Ergo AN OPEN ACCESS JOURNAL OF PHILOSOPHY

Algorithmic Fairness Criteria as Evidence

WILL FLEISHER
Georgetown University

Statistical fairness criteria are widely used for diagnosing and ameliorating algorithmic bias. However, these fairness criteria are controversial as their use raises several difficult questions. I argue that the major problems for statistical algorithmic fairness criteria stem from an incorrect understanding of their nature. These criteria are primarily used for two purposes: First, evaluating AI systems for bias, and, second, constraining machine learning optimization problems in order to ameliorate such bias. The first purpose typically involves treating each criterion as a necessary condition for fairness. The second use involves treating criteria as sufficient conditions for fairness. Since the criteria are used for both roles, some researchers have treated them as both necessary and sufficient conditions, i.e., as definitions of algorithmic fairness. However, serious problems have been raised for the use of these fairness criteria. Under ordinary circumstances, it is impossible to satisfy multiple criteria at the same time. Moreover, there are counterexamples to both the sufficiency and necessity for fairness of each criterion. I argue that we should instead understand fairness criteria as merely providing evidence of fairness. In other words, satisfaction (or violation) of these criteria should be understood as potential evidence of fairness (or bias). Whether a criterion counts as evidence in a particular case will depend on stakeholders' background knowledge and the specific features of the system's task. This evidence account of fairness conditions provides guidance for recognizing both the appropriate uses and the limitations of fairness criteria.

1. Introduction

Algorithmic systems, especially those trained using machine learning techniques, are being deployed for a rapidly expanding number of important decisions affecting people's lives. These include hiring, school admissions, policing, and

Contact: Will Fleisher <will.fleisher@georgetown.edu>

https://doi.org/10.3998/ergo.9869

pre-trial detention decisions. Unfortunately, these algorithmic systems often display pernicious biases against marginalized groups (Angwin et al. 2016; Benjamin 2019; Buolamwini & Gebru 2018; Fazelpour & Danks 2021; O'Neil 2016). This bias results in systems that make erroneous predictions about people, or that lead to unjust or incorrect decisions. Moreover, because these systems are often trained using machine learning (ML), they can become biased or unfair without their developers intending to be discriminatory or unfair. The problem of algorithmic bias will only become more pressing as new developments in AI allow for the creation of AI systems with a wider range of capabilities (Gabriel et al. 2024).

In response to this problem of algorithmic bias, researchers interested in promoting fair machine learning have proposed a wide variety of *fairness criteria* (sometimes also called metrics or conditions) in order to diagnose and ameliorate such bias (Barocas et al. 2023; Corbett-Davies et al. 2023; Mehrabi et al. 2021; Verma & Rubin 2018). Most of these criteria consist in statistical parity conditions on the output of algorithmic systems. In other words, they concern whether the results of an algorithmic system are the same (in some sense) for people from different social groups. For instance, some criteria require that an AI system makes errors at the same frequency for members of different groups.

Statistical fairness criteria are primarily used for two purposes. First, they are used to evaluate the output of algorithmic systems for fairness and bias, even in cases where the internal operation of the system is opaque. For evaluation purposes, the criteria are treated as necessary conditions for fairness. In other words, when practitioners use a criterion for evaluation, they assume that *if the system is operating fairly, then the criterion will be satisfied* (Eva 2022; Hedden 2021; Long 2021). The second purpose is to ameliorate bias. This involves using a criterion as a constraint on a machine learning optimization problem (Barocas et al. 2023; Eliassi-Rad 2020; Ongun et al. 2019). For this purpose, the criteria are treated as sufficient conditions for fairness, i.e., practitioners assume that *if the criterion is satisfied, then the system is fair*.

Statistical fairness criteria are often called *definitions* of fairness (Hutchinson & Mitchell 2019; Mehrabi et al. 2021; Narayanan 2018; Verma & Rubin 2018). This makes sense, given the two purposes just mentioned. A definition is a set of necessary and sufficient conditions, so if a criterion really was a definition, it could serve both purposes.

However, these ways of understanding the fairness criteria—as necessary conditions, sufficient conditions, or definitions—have led to difficulties. For one thing, a wide variety of quite different statistical fairness criteria have been proposed (Dwork et al. 2012; Narayanan 2018; Verma & Rubin 2018). Unfortunately, these various criteria cannot serve to characterize a unified

conception of fairness, as the criteria are mathematically impossible to satisfy at the same time in ordinary circumstances (Chouldechova 2017; Corbett-Davies et al. 2023; Eliassi-Rad & Fitelson 2021; Kleinberg et al. 2016). Moreover, there are counterexamples for treating each of the proposed criteria as either necessary or sufficient for fairness.

I will argue that we should understand statistical fairness criteria as providing evidence of fairness, and their violation as evidence of unfairness. I will call this the *evidence account* of algorithmic fairness criteria. On this account, whether satisfying or violating a particular criterion counts as relevant evidence about the fairness of an AI system will depend on the particular circumstances where the system is deployed and the background knowledge of developers, domain experts, and stakeholders. Understanding these criteria as evidence of fairness avoids the problems posed for the other conceptions of fairness criteria by incompatibility and counterexamples.

Below, I will first provide some additional background about the fairness criteria in question. Then, in §3, I will introduce some of the problems that arise from treating these criteria as conditions of fairness. These include a wide variety of counterexamples from the literature, but I will also offer some new cases. In §4, I present the evidence account and argue that it helps to solve these problems. Then, in §5, I will defend the evidence account from a potential competitor, which I call contextualism.

2. Fairness Criteria

Research concerning fair machine learning has primarily been concerned with evaluating fairness and ameliorating bias in *classification* systems: algorithmic systems trained to accurately classify an individual based on their features into one or more categories. Classification systems are in widespread use as aids (or replacements) for human decision-making. I will focus on classification systems as a running example.

An important motivating case for the Fair ML literature—one which has also proven divisive—is the COMPAS risk-scoring system, specifically as it was used by Broward County, Florida (Angwin et al. 2016). COMPAS is a widely used system meant to help judges make pre-trial detention and sentencing decisions. It assigns a risk score of between 1 and 10 to a criminal defendant. This score is meant to represent the risk of that defendant committing another crime (i.e., their risk of recidivism). Precisely how COMPAS's internal model works is not completely clear, as its code is proprietary. However, we do know that the system takes as input up to 137 features about individual defendants, drawn both from answers to questions provided by the defendant, along with criminal records

(Angwin et al. 2016).[1] The COMPAS score is provided to judges as advice for making pre-trial detention and sentencing decisions.

In order to evaluate COMPAS for bias, ProPublica's researchers compared its predictions regarding defendants in Broward County with subsequent arrest records (Angwin et al. 2016). The results, they suggested, showed evidence of bias against Black defendants. Specifically, they found that the overall rates of errors made about white and Black defendants were similar, but that the system was prone to make different kinds of errors for Black defendants than for white defendants. In particular, Black defendants who were not rearrested were much more likely to be falsely flagged as high risk—where this means having a risk score above 5. In other words, Black defendants received higher rates of *false positives*. ProPublica's appeal to false positives inspired a specific statistical criterion of fairness: *false positive rate parity* between social groups.

In response, the company that created COMPAS, known at the time as Northpointe (later called Equivant), argued that ProPublica was mistaken to focus on FPR parity as a relevant criterion to detect bias (Flores et al. 2016). They, along with some independent academic researchers (Corbett-Davies et al. 2023), argued that a better fairness criterion is *calibration by group*. Calibration by group—or *group calibration* for short—is a requirement that applies to risk assessment systems that assign scores. It concerns the percentage of people assigned a certain score who actually have the property the score is meant to track (i.e., who are in the *positive class*). Group calibration requires that this percentage is the same across relevant groups.

FPR parity and calibration by group are both statistical fairness criteria that have been defended as necessary or sufficient conditions for algorithmic fairness. A wide variety of similar statistical fairness criteria—called, collectively, *group fairness* criteria—have been suggested. Hedden (2021) identifies eleven such criteria. Verma and Rubin (2018) consider twenty distinct criteria. Narayanan (2018) identifies twenty-one. Like FPR parity, many of these criteria concern ratios defined using values from the four quadrants of an error matrix (Verma & Rubin 2018). That is, they are defined by appeal to the number of predictions the system makes that are true positives, false positives, true negatives, and false negatives.[2]

For ease of discussion, I will focus on FPR parity and group calibration, as they are among the most popular proposed criteria, have close relations to some

1. Though note that it has been reverse-engineered. It is comparable in accuracy to a system using only a few features concerning age and criminal history (Angelino et al. 2018; Dressel & Farid 2018).

2. Individual and counterfactual fairness criteria provide distinct paradigms for fair machine learning (Dwork et al. 2012; Kearns & Roth 2019; Kearns et al. 2017; Loi et al. 2023). Here, I focus on so-called "group fairness" criteria for ease of discussion. These other fairness criteria suffer from issues of their own (Fleisher 2021; Loi et al. 2023). Moreover, the evidence account can straightforwardly accommodate individual and counterfactual fairness.

other influential criteria, and because they helpfully illustrate problems that apply more generally. I will define them in commonly used terms (Hedden 2021; Verma & Rubin 2018). For binary classification problems, such as predicting a person to be high-risk or low-risk, we can use the term *positive class* to refer to those individuals who actually have the property the classifier is attempting to predict (e.g., defendants who will be rearrested). That an individual is in the positive class is often represented symbolically as $Y = 1$. *Negative class* refers to those individuals who do not have this property (e.g., defendants who will not be rearrested). This is represented as $Y = 0$. A false positive occurs when an individual is given a positive classification while actually being in the negative class—e.g., when a defendant is rated as high-risk but is not rearrested. The classification made by a system is typically represented with $\hat{Y}$, where $\hat{Y} = 1$ means the system predicts the individual is in the positive class, while $\hat{Y} = 0$ suggests a negative prediction.

FPR Parity requires equal false positive rate between social groups. False positive rate is the ratio of false positives (FP) to the number of individuals in the negative class, which is equal to the false positives + true negatives (TN), i.e., $FPR = \frac{FP}{FP+TN}$.

FPR parity is a member of a broader group of criteria, sometimes called *separation* criteria (Barocas et al. 2023: 56–57). Separation criteria all concern a certain kind of conditional probability: the probability that a classifier will assign an individual to the positive (or negative) class, given that they are in the positive (or negative) class. Separation criteria require that some conditional probability of this sort is independent of group membership.

The second exemplar fairness criterion is calibration by group. In this context, *calibration* (simpliciter) is a condition on risk scores that are interpreted as probabilities (Barocas et al. 2023: 61). It requires that, for individuals who are assigned a score *s*, the percentage of those individuals who are in the positive class is also *s*. For COMPAS, this would mean that we interpret a score of 7 as meaning (roughly) that the probability the defendant will be rearrested is 0.7. Then, the COMPAS score would be calibrated if 70% of those assigned a score of 7 are in fact rearrested. Calibration by group was originally devised as a requirement that a score be equally well-calibrated (*simpliciter*) for members of each important social group. Again for COMPAS, that would mean that both Black and white individuals assigned a score of 7 are rearrested 70% of the time.[3]

Calibration by group is taken to be an important requirement because it aims to ensure that the score "means the same thing", or has the same evidential import, when it is applied to members of different groups (Corbett-Davies et al. 2023;

3. Note that satisfying calibration (simpliciter) does not entail the satisfaction of calibration by group.

Hedden 2021; Verma & Rubin 2018). We can expect that all individuals assigned the same score have the same chance of being in the positive class, regardless of their social group membership. For instance, if COMPAS is calibrated by (racial) group, then any person it assigns a score of 7 is 70% likely to be rearrested, regardless of what racial group they belong to.

However, there are two issues with this way of defining calibration by group. First, it only applies to scores that can be interpreted as probabilities, and second, it implausibly treats calibration *simpliciter* as a fairness requirement. An alternative, weaker version of the principle avoids both of these issues[4]:

Calibration by group (weak) requires that for each score that a system assigns, individuals assigned a score of s ($S = s$) have the same probability of actually being in the positive class ($Y = 1$), independent of whether they are a member of a certain social group G, i.e.,

$$Pr(Y = 1 \mid S = s, G = 1) = Pr(Y = 1 \mid S = s, G = 0).$$

According to this weaker version of the criterion, a system is calibrated by group when, for each possible score, the percentage of individuals assigned that score who are actually in the positive class is the same for each relevant social group (Barocas et al. 2023; Verma & Rubin 2018). Applied to COMPAS, this means that for every score between 1 and 10, Black defendants and white defendants assigned that score must actually be rearrested with (approximately) the same frequency. So, the percentage of white defendants assigned 7 who are rearrested must equal the percentage of Black defendants assigned 7 who are rearrested. What makes this version of the criterion weak is that the percentage in question need not be 70%. Going forward, I will only be concerned with the weak version of the criterion.

Calibration by group is a plausible candidate for a fairness criterion, as it aims to ensure a score carries the same information for members of different groups. If a system like COMPAS violates calibration, it may result in judges over-estimating the risk posed by members of one group compared to another. Allowing failures of group calibration for COMPAS would mean allowing a score of 7 to indicate a different risk of rearrest for white and Black defendants. This could predictably lead to mistakes when the scores are used by judges to make bail decisions. Moreover, scores that violate group calibration may incentivize different (and potentially discriminatory) treatment for members of different groups (Corbett-Davies et al. 2023). For instance, we can imagine a 1–10 scoring

4. This version is discussed by Chouldechova (2017: 3), Hedden (2021: 214), and Eva (2022: 47), among others. Eva also provides further arguments and counterexamples against the strong version of the principle.

system that scores college applicants based on whether they are likely to graduate. Suppose this system is not calibrated by gender group: men who receive a 7 are 70% likely to graduate, while women who receive a 7 are only 60% likely to graduate. (A score could have this feature, even if women are more likely to graduate than men, in general, and even if the score is otherwise reasonably accurate). This would motivate admissions committees to prefer men over women who have the same score.[5]

Calibration by group is a criterion that inspires strong disagreement among fair ML researchers. It is strongly favored by some researchers (e.g., Flores et al. 2016; Hedden 2021; Long 2021). However, it is also very commonly satisfied by unconstrained machine learning (Barocas et al. 2023: 19). In other words, it is often achieved simply by aiming for standard accuracy measures. For this reason, group-calibrated systems tend to reflect underlying disparities in a data set. Hence, some wonder if it is much use in diagnosing bias or unfairness (19).

Group calibration is central to another class of related criteria, sometimes called *sufficiency* criteria (Barocas et al. 2023). These criteria are all related to another conditional probability: the probability that an individual belongs in the positive (or negative) class, given that the classifier assigns them to the positive (or negative) class. Sufficiency criteria require that this probability is independent of group membership.

Because FPR parity and calibration by group are representative members of separation criteria and sufficiency criteria, respectively, they serve as useful examples for discussion.[6] With these examples on the table, we can turn to discussing the uses of fairness criteria, and their relation to bias and fairness. Recall that the two primary ways these criteria are used is for evaluation of AI systems, and for constraining machine learning in the development of such systems.

ProPublica's evaluation of COMPAS (Angwin et al. 2016) illustrates the evaluative use of fairness criteria. This use is supported by the assumption that fairness criteria serve as necessary conditions on fairness. Angwin et al. suggest that COMPAS violates FPR parity, and is therefore unfair. This inference makes sense if FPR parity is a necessary condition for fairness. For this reason, some other researchers—particularly philosophers— have taken fairness criteria to be necessary conditions (Eva 2022; Hedden 2021; Long 2021).

In addition to their evaluative role, fairness criteria are also used in attempts to ameliorate bias. This involves using the criteria as constraints for a machine learning optimization problem (Barocas et al. 2023; Eliassi-Rad 2020; Hardt et al. 2016; Ongun et al. 2019; Saleiro et al. 2018; 2020). To see what

5. Thanks to an anonymous referee for suggesting this point.

6. For detailed overview and discussion of separation and sufficiency and criteria in each category, see (Barocas et al. 2023: ch. 3; Verma & Rubin 2018).

this means, it helps to have a rough understanding of how machine learning works.[7] For illustration, I will continue to focus on predictive classification tasks, and the kind of supervised learning typically used to develop programs for this task.

Roughly, then, machine learning is a set of methods for using data to develop a computer system that is effective at some task. Classification tasks often involve developing a program for predicting the behavior of some target phenomena. This kind of classification program computes a function that is meant to represent whatever actual, real-world relationship holds between the features in the data set and the target phenomena in question. For instance, a program that is meant to predict whether a person will default on a loan, based on data about their income and age, will compute a function from salary and age to probability of default. The function implemented by such a program is called a "model" because it is meant to represent the real-world relationship between income, age, and default. A model that is useful for ML will have parameters that can be adjusted to affect what the system does. A model can be as simple as a 2-dimensional linear equation, familiar from grade-school math, where the parameters are the slope (m) and y-intercept (b). Alternatively, a model can be a complex deep neural network whose parameters are the weights of the connections between artificial neurons. Machine learning involves solving an optimization problem: given a model, the objective is to learn parameter values that minimize a chosen loss function when the model is applied to its training data. The loss function represents the ML system's error: how badly the system does at its task when applied to the training data. For a classification system trained with supervised learning, loss minimization is achieved by repeatedly asking the system to classify individuals in its training data, then using the loss function to calculate how badly the system did, and then slowly adjusting the parameters so that the loss goes down over time until parameters are discovered that have the lowest loss (i.e., the lowest degree of error).

Fairness criteria can serve as constraints on this optimization problem. The constrained ML optimization problem is to find the parameters that minimize loss while also satisfying the chosen fairness criteria. This use is justified by the assumption that fairness criteria are sufficient conditions for fairness: If the criteria are satisfied, then the resulting classifier is fair. This sufficient-condition understanding of the criteria makes sense of their use as constraints for ML.[8]

7. For a better overview, see (Barocas et al. 2023) or (Russell & Norvig 2020).

8. Beigang (2023) offers an explicit endorsement of treating fairness conditions as sufficient conditions for avoiding unfair discrimination.

Fairness criteria are often discussed as if they are definitions *of* fairness (Chouldechova & Roth 2018; Corbett-Davies et al. 2023; Hutchinson & Mitchell 2019; Mehrabi et al. 2021; Verma & Rubin 2018). On this view, each criterion is meant as a precise, mathematical formulation of the definition of the concept of fairness, at least as that concept applies to algorithmic systems. This definitional understanding vindicates both the evaluative and ameliorative uses of fairness criteria: if a criterion is a definition of fairness, then it is both a necessary and sufficient condition for fairness. Sometimes, these fairness definitions are taken to be competing accounts of the same underlying notion or property. In other cases, fair ML researchers are committed to a more pluralist, contextualist view about the criteria. That is, they take different fairness criteria to capture different types of fairness—and different senses of "fairness"—that are applicable in different contexts. I will consider a sophisticated version of this contextualist view in the final section, before arguing that the evidence account offers a better theory of fairness criteria.

3. Problems for Fairness Criteria

There are two primary problems for treating fairness criteria as either necessary or sufficient conditions (or as definitions) for fairness. First, there are incompatibility results showing that, given certain plausible assumptions, some sets of fairness criteria cannot be mutually satisfied in ordinary circumstances. In particular, separation and sufficiency requirements cannot be mutually satisfied under ordinary conditions. So, this incompatibility holds for group calibration and FPR parity, our two running examples of criteria from these families. Second, there are compelling counterexamples to the necessity and sufficiency of each criterion as a condition for fairness.

3.1. Incompatibility results

There is an obvious response to the previously mentioned dispute between defenders of COMPAS and ProPublica regarding FPR parity and group calibration. Specifically, since both criteria seem like plausible requirements of fairness, the obvious solution is to require risk assessment tools like COMPAS to satisfy both constraints. However, it turns out that this is mathematically impossible whenever the base rates (or prevalence) of the property at issue are different for the two groups. That is, since the base rate of rearrest is higher for Black defendants than white defendants in Broward County, it is impossible for a criminal risk

assessment system deployed there to satisfy both group calibration and FPR parity at the same time (Chouldechova 2017; Corbett-Davies et al. 2023; Eliassi-Rad & Fitelson 2021; Kleinberg et al. 2016).

These impossibility results obtain because the two criteria in question are determined by appeal to some of the same statistical frequencies in the output of a system (Barocas et al. 2023; Long 2021). Among other connections, they are both sensitive to the frequencies of false positives. However, group calibration is concerned with the ratio of false positives compared to the system's predictions ($\hat{Y}$). So, whether group calibration is satisfied depends, in part, on the ratio of false positives to the overall number of defendants predicted to be positive. In contrast, FPR parity tracks the number of false positives as a share of actual outcomes (Y). FPR just is the ratio of the number of false positives to the number of individuals in the negative class.

When the base rates between social groups are different, using a group-calibrated classifier will result in a different number of false positives for the two groups. This creates a violation of false positive rate parity. Meanwhile, taking a group-calibrated classifier and changing it so that it satisfies FPR parity requires changing the number of false positives received by at least one of the groups. This will result in violation of calibration by group. Again, this is because the two criteria are partially determined by the same quantity: the number of false positives. However, the two criteria compare this number of false positives with different values that are partially determined by the base rates. These results generalize to the other proposed criteria because they are all similarly defined in terms of (a) the error/accuracy rates displayed in the classifier's confusion matrix and (b) the numbers determined by the base rates.[9]

These incompatibility results raise a problem for thinking of fairness criteria as necessary conditions. Assuming fairness is possible, the fairness criteria cannot all be necessary conditions for fairness as they cannot be satisfied at the same time. For the same reason, they cannot serve as a definition of a single concept or property of fairness, i.e., as a set of necessary and jointly sufficient conditions for a classifier to be fair.[10]

---

9. For a more detailed discussion of the incompatibility results and helpful clarification, see (Barocas et al. 2023; Eliassi-Rad 2020; Hellman 2020; Long 2021).

10. Beigang (2023) offers modified versions of separation and sufficiency conditions that are more compatible with one another. There isn't enough space to consider this proposal in detail here. Instead, I will just note that the "matching" procedure Beigang uses is similar to individual fairness measures, and so will suffer from some of the difficulties associated with those criteria (See Fleisher 2021). Moreover, the matching procedure depends on certain idealizing assumptions that often won't be met in real-life cases. Finally, Beigang's modified criteria are still vulnerable to the counterexamples discussed below (§3.2.2). Despite these issues, the evidence account could vindicate the use of Beigang's matching criteria in a variety of circumstances.

3.2. Counterexamples

The second main problem for understanding fairness criteria as either necessary or sufficient conditions for fairness is that there are also compelling counterexamples to each of the criteria. In fact, there are a wide variety of easily-produced counterexamples. This variety and ease of production will be important later in the discussion of contextualism (§5). I will provide a non-exhaustive taxonomy of counterexample types.

3.2.1. Necessity Counterexamples

First, there are a variety of counterexamples to treating each criterion as a necessary condition. These are cases where the use of an algorithmic system seems *entirely fair*, and yet the fairness criteria in question are violated.

Randomization Hedden (2021) provides a compelling necessity counterexample that simply involves coin flips and random group assignment. We can call this a *randomization* counterexample. It provides a counterexample to most group fairness criteria, including FPR parity, with the notable exception of calibration by group.

> *People, Coins, and Rooms* "Suppose that there are a bunch of coins of varying biases. Each individual in the population is randomly assigned a coin. Then those individuals are randomly assigned to one of two rooms, A and B. Our aim is to predict, for each person, whether that person's coin will land heads or tails . . . Luckily, each coin comes labeled . . . with a real number in the interval [0, 1] indicating its bias, or its objective chance of landing heads. Here is a perfectly fair and unbiased predictive algorithm: For each person, take their coin and read its label. If it says 'x' assign that person a risk score of x. And if $x > 0.5$, make the binary prediction that they are a heads person (positive), while if $x < 0.5$, make the binary prediction that they are a tails person (negative)." (2021: 219)

There is no unfairness in this setup. However, this setup may simultaneously violate all the proposed statistical fairness criteria *except* group calibration. Hedden suggests an example with the following distribution: Room A has 12 people with coins labeled ".75" and eight people with coins labeled "0.125". Room B has ten people with "0.6"-labeled coins and ten people with "0.4"-labeled coins.

This example violates FPR Parity between people in the two rooms: Assuming we assign risk scores and predictions as above, and that the bias labels are accurate, then the false positive rates for rooms A and B are 3/10 and 4/10, respectively.[11]

Hedden's randomization case is a counterexample to the necessity of most proposed fairness criteria at issue, but by design calibration by group cannot be violated in such a setup. Hedden suggests that this is support for thinking that group calibration is the only necessary statistical criterion. However, there are counterexamples to group calibration as a necessary condition for fairness, too.

Gerrymandering — Non-Necessity Eva offers counterexamples to calibration by group that depend on identifying groups and sub-groups using cross-cutting categories (2022: 249–250). Eva's gerrymandering cases require categories of two types: first, sensitive or protected categories that are socially important but are not permissible to use for making the decision (e.g., race, gender, etc.); and second, categories that are relevant to the decision but are not socially sensitive (e.g., test scores, credit scores). These categories are combined to create four subgroups. For instance, he suggests dividing applicants for auto insurance into four groups using distinctions between young and old drivers, and between high and low credit score drivers. A risk scoring classifier can fail to be group-calibrated between these four groups, while also intuitively failing to count as biased against a protected category. That is, a score that is not group-calibrated with respect to the four groups can be intuitively fair and unbiased with respect to how it treats the two sensitive social groups. In his case, a non-group-calibrated score still means the

11. For discussion of how this violates the other criteria see Hedden's original treatment. For further discussion, see also Eva (2022). One might worry that this case does not involve machine learning, important decisions, or socially salient groups. Hence, it might not seem like a relevant counterexample. Viganò et al. (2022), for instance, argue that this case does not provide a counterexample to separation criteria such as FPR parity. This is because the predictive algorithm involved makes its judgments based on the coins' labels which are in turn based directly on the coins' objective chances. They do not involve making inferences about one individual based on data from other individuals. This, Vigano et al. suggest, is a reason to think it is not a counterexample to FPR parity (or other separation requirements). Fairness criteria should only apply, they think, to prediction systems that involve making inferences about one person based on other people's data, as is typically the case for algorithmic systems trained with ML. However, I think it is relatively simple to produce a case that is structurally identical to Hedden's case, but where the labels are determined by statistical inference based on other people's behavior, and where the prediction system concerns important social goods. We can simply imagine that the coins in question are produced in particular factory runs, and those runs all have the same bias in their objective chances. Then, we label the coins we distribute based on the past behavior of coins from the same run. Then, we can imagine that the system is being used to distribute important but scarce resources that society has determined should not be distributed according to merit, but by lottery.

same thing when applied to members of the socially important categories of old and young drivers.

Eva subsequently presents his own novel fairness criterion that he calls *Base Rate Tracking*: "The difference between the average risk scores assigned to the relevant groups should be equal to the difference between the (expected) base rates of those groups" (2022: 18). The motivation for this criterion is to track whether, and to what degree, a risk score worsens (or improves) things for a particular group from the starting point of the base rates for that group.

Brute Force The following counterexample undermines both calibration by group and base rate tracking as necessary conditions for fairness:

> *Lead Abatement* A municipal government is determining which houses require lead abatement. Black homeowners are more likely to have homes with lead paint. The moral cost of false negatives is enormously high as children may be poisoned. Meanwhile, the cost of lead abatement is comparatively low. The city deploys an AI system that is designed to reflect these costs. The system assigns a maximum score of 10 to every house, i.e., it recommends that *every* house requires lead abatement. This score violates group calibration: a score of 10 means something different for white homeowners and Black homeowners. Specifically, a Black homeowner with a score of 10 is more likely to have lead problems. Moreover, the difference in the average score (of 0) between the two groups is far lower than the (nonzero) difference in base rates. Hence, base rate tracking is violated. But the system is perfectly fair: no one has a reasonable complaint of being treated unfairly, given the course of action chosen and the relative costs involved.

This example shows that neither group calibration nor base rate tracking are necessary conditions for algorithmic fairness. Examples like these belong to a large category of what I call *brute force* counterexamples. They work by forcing values to some extreme for every individual being classified—so that the parity conditions in question are obviously violated—but in a way that does not violate our intuitive notion of fairness because the outcomes benefit all participants in unobjectionable ways.[12]

12. (Grant 2023: 101) offers a related type of counterexample aimed at base rate tracking. In his case, a credit scoring algorithm overestimates risk of default for Black applicants, but only for applicants who are so far below the threshold of risk that they all still receive loans. Helpfully, Grant's counterexample makes a complementary move. It relies on making the failure of base rate tracking so small that the difference makes no impact on people's welfare.

Valence Reversal — Non-necessity There are a variety of counterexamples that involve reversing the valence of an algorithmic system's decisions. To create counterexamples in this way, we look at the cases used to motivate a criterion, where the ML classification system is intuitively unfair, and where the proposed fairness criterion is violated. Then, we reverse the valence of how the system's classifications are interpreted and used.

Castro (2022) offers a valence reversal case based on a COMPAS-style risk-score classifier.[13]

> *Violent Offense* You are deciding which defendants to give free anger management counseling to. There are a limited number of counselors, and your task is to increase public safety by giving counseling vouchers to defendants who are at high risk of committing violent offenses while out on bail. Male defendants are much more likely to commit violent crimes while out on release than female defendants, and your data reflect this. You construct an accurate—but not perfectly accurate—and group-calibrated system for identifying individuals who are at high risk of committing violent offenses while out on bail. You give "high risk" individuals vouchers. (2022: 175)

In this case, FPR parity is violated (much as in the original COMPAS case), yet the case seems perfectly fair. There are very good reasons, in this case, to care more about reaching more men (and protecting more women), rather than worrying about men having high false positive rates for being offered vouchers.

Another example of a kind of valence reversal case involves a violation of group calibration due to affirmative action.[14] Here, we consider cases where sensitive categories, the most common being race, are used as a means for making classification decisions. In the initial, unfair case, we imagine that a college admissions system is designed to classify applicants based on their likelihood of graduating, but is also designed to discriminate against Black people. For this system, a Black applicant with a score of 7 is much more likely to graduate than a white applicant with a score of 7. The system therefore violates group

13. Compare this also to discussion in Rodolfa et al. (2020) and Loi et al. (2024) concerning the distribution of advantages vs. disadvantages.

14. Whether affirmative action is fair or just is, of course, not without controversy. However, I am convinced it can be a fair means for promoting diversity and for redressing past injustice. I will continue with that assumption here. For arguments that affirmative action can be fair or just in general, see (Fullinwider 2018). For discussion of affirmative action in Fair ML, see Barocas et al. (2023). Castro et al. (2023) offer similar cases that turn on affirmative action-style interventions (though they don't use the term).

calibration and is also unfair. To construct the valence reversal counterexample, we instead imagine a college admissions system designed to engage in affirmative action. Here, Black applicants with a score of 7 are *less* likely to graduate than white applicants with the same score. So, once again, calibration by group is violated. However, this system is designed to promote diversity and redress historical injustice. Thus, the system is intuitively fair, despite the violation of group calibration.[15]

3.2.2. Sufficiency Counterexamples

There are also clear counterexamples to each criterion's sufficiency for fairness.

Red-lining Eva (2022: 14) offers a counterexample for the sufficiency of group calibration that appeals to red-lining: the policy of discriminating against Black applicants for mortgage loans using Black neighborhoods as a proxy for identifying Black applicants. He imagines a case where a bank's risk assessment system assigns scores meant to track an applicant's likelihood of loan default. The score is based purely on zip code, which is a reliable enough proxy for the likelihood of default. In the imagined case, Black and white applicants within a zip code have similar rates of default, however, Black applicants live predominantly in poorer zip codes. The result is a score that is group-calibrated for white and Black applicants while being clearly and intentionally designed to discriminate against Black applicants.[16]

Leveling Down Long (2021: §4.3) and Corbett-Davies et al. (2023) both offer cases that serve as sufficiency counterexamples to FPR parity. In these cases, a risk score is made less accurate (and so no longer group-calibrated) at evaluating a protected social group in order to cause the score to satisfy FPR parity despite differences in base rates. An individual from the protected group facing such a system would face a greater risk of error, and can thus reasonably complain they had been treated unfairly. The upshot is that forcing FPR parity satisfaction is insufficient (and may be counterproductive) for ensuring fairness. It can constitute a kind of leveling-down: FPR parity is achieved by making some new people

15. We can also construct similar, intuitively fair affirmative action cases that violate base-rate tracking. To do so, we construct a similar college application case, but ensure that the score assigned by the admissions system for Black applicants indicates that they will graduate at a rate that outstrips the base rate for graduation among Black students generally.

16. Eva's own proposed criterion of base-rate tracking is not meant as a sufficient condition, suggesting that Eva recognizes that it faces similar sufficiency counterexamples.

worse off while failing to help the people who were being treated unfairly in the first place.[17]

Gerrymandering — Insufficiency Dwork et al. (2012) and Kearns et al. (2018) offer gerrymandering counterexamples to the sufficiency of simple demographic parity requirements. However, the gerrymandering examples can also provide sufficiency counterexamples for separation conditions such as FPR parity. Gerrymandering cases work by ensuring that the condition in question is satisfied, while only engaging in accurate classification for some of the groups in question.

> *Randomized Group* Imagine a classifier for college admissions that predicts whether applicants will graduate. Suppose, for simplicity, that there are only two racial groups in the applicant pool: 80% are white and 20% are Black. The classifier is designed to satisfy demographic parity: it recommends admitting the same number of white and Black students. The classifier is highly accurate for white applicants. However, it selects Black applicants entirely at random. Then the demographic parity condition will be satisfied, but intuitively the system is biased or unfair. We can construct a similar case for separation conditions like FPR parity. Suppose that the system again selects Black applicants randomly, but the system gets lucky and selects Black students who happen to graduate at roughly the same rates as white students.

Valence Reversal — Insufficiency Valence reversal cases can also be used to demonstrate insufficiency. Castro offers another case for this purpose, based on the *violent offense* case above, aimed at showing the insufficiency of FPR parity for fairness. In this variant of the case, he imagines that the system is constructed so it has "lower standards for identifying women as high risk than it does for men so that the false positive rates among men and women will be in parity" (2022: 177). In other words, it is designed to treat women differently than men in order to ensure FPR parity. However, this is intuitively unfair: it places a "burden on certain women to make up for the bad behavior of men" (177).

In sum, there are a wide variety of types of counterexamples that have been offered for both the necessity and sufficiency of all proposed fairness criteria. Hence, the proposed fairness criteria cannot serve as necessary or sufficient conditions for fairness. In addition, the case types identified here serve as recipes

17. For discussion of Leveling down objections to egalitarian theories of justice, see (Arneson 2015; Parfit 2002). Similar leveling-down cases have been discussed by Holm (2023a), though Holm attempts to respond to objections based on these cases.

for developing counterexamples to each criterion in a wide variety of contexts (which will be important for evaluating contextualism in §5.)

4. The Evidence Account of Fairness Criteria

The evidence account offers a different explanation for why statistical fairness criteria like FPR parity and group calibration have seemed important for determining whether an algorithmic decision-making process is fair. It suggests that all of these proposed criteria simply offer evidence regarding whether an algorithmic decision was fair or biased. That is, the satisfaction (or violation) of each proposed criterion can serve as a reason to believe that an algorithm—or a broader social system of which the algorithm is a part—is fair (or unfair).

4.1. The Account

The Evidence Account The satisfaction or violation of a genuine algorithmic fairness criterion (potentially) provides defeasible evidence regarding whether an algorithmic system is fair. The violation of a genuine algorithmic fairness criterion provides defeasible evidence that the system is unfair. The satisfaction of such a criterion provides defeasible evidence that the system is fair.

According to this account, fairness criteria offer a way to obtain evidence concerning a system's fairness. When applied to a particular task, they provide ways of testing a system for fairness, i.e., ways of detecting fairness or unfairness. Whether satisfaction or violation of a fairness condition actually counts as evidence in a particular case—and how strong that evidence is—depends on the context an algorithmic system is operating in and the background knowledge of those who are evaluating it.

The evidence account only claims that the satisfaction or violation of fairness criteria provides defeasible evidence for the fairness of an AI system. This evidence is *prima facie*, meaning the violation of a criterion like FPR parity may provide what appears to be evidence of unfairness, but further investigation or reflection undercuts the apparent evidential connection. Moreover, the evidence provided is also only *pro tanto*: it could be overridden by stronger counter-vailing evidence.[18] To see the importance of these things, consider a case like one of the affirmative action cases from §3.2.1. Suppose it is first discovered that a system violates group calibration. However, upon further investigation, it

18. For a quick explanation of the terms *prima facie* and *pro tanto* see (Stanton-Ife 2022: fn. 2).

turns out that this is because the system is designed to promote restorative justice by providing benefits to a historically marginalized group. This further evidence defeats any support for the idea that group calibration violation in this particular case counts as unfair. The violation was merely *prima facie* evidence in this case.

The key move of the evidence account is to re-frame the discussion as concerning an epistemic relation, rather than a metaphysical one. Instead of considering what the right definition or conception of fairness is, and which formal criteria best capture it, we can instead consider what satisfying (or violating) a criterion tells us in particular circumstances given our background normative commitments and domain knowledge. This evidential relationship between fairness and the criteria is a weaker relation than we might have hoped for. But the account does vindicate the usefulness of such criteria in many circumstances (see §4.2). Recognizing the nature of this relation also helps to identify some limitations of using fairness criteria (see §4.3).

Although an evidential relation is weaker than a necessary or sufficient one, the evidence provided by the violation (or satisfaction) of a criterion may be strong, when considered *as evidence*. Violation of a fairness criterion can serve as very strong evidence of bias in the right context. Even when the evidence relation is a weak one, however, this can still be useful for the purposes of auditing or evaluating a system for fairness. For instance, it can be the starting point into an investigation, when our expectations about a system's output are violated.

The evidence account is neutral between various theories of evidence, as it requires only a few weak and widely accepted assumptions regarding the nature of evidence.[19] First, what counts as evidence, and how strong the evidence is, depends on the background knowledge of the subjects involved (Longino 1990).[20] Second, evidence is defeasible: it can be undermined or overridden by further evidence (Koons 2022). Third, the degree of evidential support that a proposition E provides a hypothesis H depends on how probable E is, given H.[21] This third assumption will help in offering guidance on which particular fairness criterion is relevant in a particular circumstance, given the incompatibility results discussed above.

19. For background on various philosophical views of evidence that support these assumptions, see (Kelly 2016).

20. Note that this is true even for highly objectivist views about evidential favoring relations, e.g., Williamson's evidential probabilities (2002), objective Bayesianism (Lin 2022). There might be an objective fact about how much evidence P provides for Q, given background knowledge set K. But what counts as K depends on the context, at least insofar as it depends on who the subjects in question are and what they already know.

21. This last assumption is a weaker corollary of the "law of likelihood" (Sober 2008).

4.2. Advantages of the Account

The evidence account offers several advantages. First, the counterexamples raised in §3 pose no difficulty for understanding fairness criteria as providing evidence of fairness. In general, evidential relations are possible even in the absence of sufficiency or necessity relations. A proposition *E* need not be a necessary condition for the truth *H* in order for *E* to be evidence for *H*. For instance, in a criminal investigation, that a suspect's fingerprints are on the murder weapon (*E*) is evidence that they are the murderer (*H*). However, the presence of the fingerprints is not a necessary condition for the suspect to have committed the crime; they could have worn gloves. The fingerprints are also not a sufficient condition for the suspect having committed the crime; someone else could have stolen a weapon owned by the suspect in order to frame them. The same point applies to fairness criteria. That a system violates FPR parity can be evidence of the system's unfairness, even though the violation of the criterion is neither necessary nor sufficient for unfairness. Thus, the fact that the cases from §3 are counterexamples to the necessity or sufficiency of the criteria for fairness does not necessarily undermine the criteria's usefulness as evidence.

A second advantage of the evidence account is its compatibility with extant philosophical theories of fairness. Moreover, it is compatible with treating such theories as providing a univocal sense of "fairness." In other words, the account is consistent with thinking there is a unified core conception of what it means for a social system to be fair. For instance, it fits well with a contractualist notion of fairness, where a person's treatment is fair only if no one could reasonably reject the principle permitting such treatment (Scanlon 2000). Alternatively, the view is compatible with a theory of fairness that treats it as essentially involving equality of opportunity (Barocas et al. 2023; Heidari et al. 2019; Holm 2023a; Loi et al. 2024), or one that sees fairness as requiring a proportional satisfaction of claims (Holm 2023b; Lippert-Rasmussen 2022).[22]

The evidence account is compatible with a univocal notion of fairness because it can explain why different fairness criteria are applicable in different contexts, even if there is only one kind of fairness. For instance, what is required by principles no one could reasonably reject will depend on context. Rawls suggests that in a just, well-ordered society, a substantive equality of opportunity principle is strictly more important than the prioritarian difference principle (Rawls 1971).

22. Note that the evidence account is also compatible with pluralist accounts that suggest there is more than one sense of "fairness", or accounts that make important distinctions between fairness and justice.

However, in circumstances of existing injustice, the prioritarian principle is instead of paramount importance (215–217, 54–55).[23]

Crucially, what would serve as evidence that those principles have been satisfied will *also* depend on context: even in two cases where a non-rejectable principle requires the same thing—e.g., substantive equality of opportunity—the evidence that this principle is satisfied (or unsatisfied) might be different in the two cases. In one case, satisfying group calibration may be evidence of fairness, because respecting substantive equality of opportunity may require ensuring that information is equally distributed among all loan applicants. In another case, the satisfaction of FPR parity will provide evidence that the same equality of opportunity principle is satisfied, because it requires that risk of harm is distributed equally across two groups, amongst those who have the same target feature—e.g., Black and white criminal defendants. Thus, the very same theory of *what fairness is* can result in different criteria providing evidence in different contexts.[24] The evidence account is in a strong position to explain this because evidence is generally context-sensitive. As noted above, whether a proposition serves as evidence for a subject depends on their background knowledge and assumptions.

According to the evidence account, the procedure for effectively using a criterion should then look like this: we consider (a) what moral reasons or principles we should be sensitive to in a context, along with (b) what we know about the social and other causal features of the context. Given (a) and (b), we can ask what the outputs of a fair (or unfair) classification would look like in this situation. If we know enough about the situation, we should have some expectations about what the results should look like. If those expectations are satisfied (or violated) this gives us some evidence regarding the system. This could suggest that we don't adequately understand (b), the descriptive facts about the context. But it can also suggest things about (a), that is, whether the moral reasons in question are being adequately respected.

An example will help to illustrate the second advantage and to show how the evidence account can offer guidance for using fairness criteria. Suppose we expect that, for a particular task in a particular context, a fair classifier will produce a group-calibrated output. This might occur when a classifier is used by individuals for their own decision-making (Loi et al. 2024). For instance, we might design a credit scoring system that helps consumers judge their own ability to pay back a loan. We would expect that a fair system of this sort will offer a score that means the same thing—that provides the same quality of evidence—for both white and Black users. If we discover that such a classifier's score is not group-calibrated,

23. See (Taylor 2009: 485) for discussion.

24. Castro et al. (2023) and Loi et al. (2024) make similar points.

this gives us evidence that the classifier is unfair: its information content is unfairly biased when used by members of certain groups. Of course, further investigation might defeat this evidence. But it is at least a *prima facie* problem worth investigating.

A third advantage of the evidence account is that the incompatibility of different fairness criteria poses no difficulty for it. One way this incompatibility is explained is by the defeasibility of evidence. There are two relevant kinds of defeat: undercutting and rebutting.

Determinate/determinable relationships illustrate how undercutting defeat helps make sense of incompatible evidence. If I learn that a can of paint is not crimson, that is some (weak) evidence that the paint is not red. Crimson is one way of being red, a way that is now ruled out, so the probability that the can is red is (slightly) lower. However, if I subsequently learn that the paint is scarlet, this is (maximally) strong evidence that the paint is red. Moreover, this second piece of evidence undercuts the evidential force of the first piece: once I know the paint is scarlet, the fact that it fails to be crimson is irrelevant to whether I should think it is red.

This general point about evidence and undercutting defeat also applies to evaluating the evidential import of violating or satisfying fairness criteria. Consider again the example of the credit-scoring algorithm designed to be used by loan applicants. We expect that a fair classifier of this kind (in this context) will be group-calibrated. Furthermore, we know there are uneven base rates of loan repayment between groups. Given our knowledge of the impossibility results discussed above, we should expect that a fair version of such a classifier will *violate* FPR parity. In this case, FPR parity violation will *not* provide evidence of unfairness.

The evidence account can also accommodate incompatibility by suggesting that in some cases violation (satisfaction) of a criterion can serve as a rebutting defeater for the satisfaction (violation) of another criterion. A rebutting defeater R is a piece of evidence that provides strong enough evidential support for a proposition P that it overrides the evidential force of another piece of evidence Q which is evidence *against* P. In rebutting defeat, R doesn't provide a reason to doubt Q or to doubt that Q is evidence for $\neg P$. Instead, R just provides such strong support for P that it outweighs the support Q provides against P. That Lisa's coat is on the rack is some (*pro tanto*) evidence she is in the office. However, if Lisa's trustworthy colleague Tom tells me Lisa is out of the office today, this latter piece of evidence is much stronger evidence that she is out. The testimony provides a rebutting defeater.

According to the evidence account, violation and satisfaction of fairness criteria can have this same structure. A system might satisfy calibration by group, which provides some evidence it is fair. However, we might then find out that

it also violates FPR parity. This violation of FPR parity might be much stronger evidence, one which serves as a rebutting defeater for the evidence provided by the satisfaction of group calibration. This can be illustrated by appeal to the COMPAS case.

FPR parity seems like a very important indicator in the context of pre-trial detention risk assessment. This is because the task concerns who is forced to bear the costs of being unnecessarily imprisoned. That FPR parity is violated by COMPAS indicates that a greater percentage of Black defendants are held on bond despite not posing a high risk of committing additional crimes. If the criminal justice system—including algorithmic risk assessment for pre-trial detention—were functioning in a fair manner, then we would expect (rough) FPR parity for Black and white defendants. Hence, the violation of FPR parity is evidence of unfairness. At the same time, COMPAS satisfies calibration by group. Group calibration is an important indicator in many cases because it tracks whether a risk score provides the same information when applied to each group. However, the justice considerations that are here being tracked by FPR parity are significantly more important. Hence, one might judge that, even though satisfying group calibration is some evidence of fairness, it is here overridden.

In addition to appealing to defeat, the evidence account can also explain the relevance of incompatible fairness criteria in another way: by suggesting that one of the background conditions that leads to incompatibility is itself unfair. Here, too, COMPAS provides an example. We know that when there are differing base rates for the target property, if a risk score is group calibrated, then the score cannot also satisfy FPR parity. So, discovering that a group-calibrated score like COMPAS cannot be made to satisfy FPR parity is evidence that there are differing base rates. For COMPAS, this means a different prevalence of rearrest among white and Black defendants. However, this itself is evidence of past (and potentially ongoing) injustice. Black people are not inherently more criminal than white people. The difference in base rates here is obviously the result of unfair, oppressive social structures. These might include differences in economic opportunities or bias in arrest rates (probably both). An algorithmic system that perpetuates such biases, as reflected in the higher FPRs for Black defendants, will compound the injustice that has been visited on Black communities (Hellman 2023). In a just society, we would expect that both calibration and FPR parity would be satisfied by a fair risk score.

In sum, in this section I have argued there are three advantages provided by the evidence account. First, it is unthreatened by the counterexamples. Second, it is compatible with extant philosophical theories of fairness, including univocal ones. Third, it explains how mutually incompatible criteria can all count as legitimate fairness conditions.

4.3. Evidence, Interventions, and News Management

Another advantage of the evidence account is that it identifies certain limitations on the use of algorithmic fairness criteria. One such limitation concerns the kinds of interventions the criteria can be used for. Even in a context where the violation of a particular fairness criterion does provide evidence of unfairness, not just any way of intervening to get the algorithm to satisfy that criterion will promote fairness. This is a general feature of evidence: Changing your evidence does not necessarily change what the evidence concerns. One has to avoid illicit *news management*.[25] Sticking your head in the sand changes what evidence you will receive about approaching danger, but not in a way that will help you avoid the danger.[26]

Similarly, changing an algorithm so that it satisfies FPR parity can hide or even compound injustice, rather than ameliorate it. Such a change can give you evidence that an AI system is fair, without changing what would be required to make using the system fair. As mentioned above, Long (2021) offers several leveling down cases where forcing satisfaction of FPR will lead to outcomes that seem no better—and may seem worse—even to members of the group such an intervention is meant to help. For instance, one way to change COMPAS so that it satisfies FPR parity would be by making it assign higher risk scores to all white defendants, thereby increasing the number of their false positives. However, this does not improve things for the Black defendants: they still face the same high false positive rates. This does not appear to improve the fairness of the situation. Instead, it constitutes managing the news one has about COMPAS—*viz.* by engineering the news that the system satisfies FPR parity. This does not change what made the system unfair in the first place, which concerns the high burden the criminal justice system places on Black defendants, and the disparate base rates of rearrest that are the result of a long history of injustices, such as disproportionate enforcement targeted at the Black community in the US.[27]

Long (2021) takes such examples to provide reasons to doubt the usefulness of FPR parity as a fairness criterion. However, I think the actual upshot is

25. I am taking the notion of news management from Lewis, who uses it as an objection to evidential decision-theory (1981: 5). Fazelpour and Lipton make a related point in the context of arguing for the importance of non-ideal theory in machine learning, suggesting that ideal theory is not action-guiding (2020: 61).

26. Terry Pratchett, in *Men at Arms*, offered a colorful illustration of news management. He describes a fictional practice of "retrophrenology", that involves hitting people over the head with a hammer to mold their skull into shapes that indicate greater intelligence. Even granting the false causal model assumed by phrenology, this is a bad idea.

27. Gonen and Goldberg (2019) offer a related case, where debiasing methods for word embeddings can make it appear that an embedding no longer displays gender-based bias, even though the bias remains.

that intervening to change our evidence about whether something is wrong is not always the right way to respond to such evidence. One must ensure the intervention will change the evidence in the right kind of way: by changing the underlying causes of unfairness, rather than by simple news management.

Telling the difference between effective intervention and mere news management will depend on the case. For instance, it is plausible that some violations of group calibration in a risk score will count as evidence of bias, and that this bias actually can be ameliorated by forcing the system to be calibrated. This will occur in cases where it seems the algorithm is biased by its training data, and where the chief harm of this bias involves new harms caused by the algorithmic system, rather than compounding historical injustices. In other cases, forcing a system to satisfy group calibration will lead to worse outcomes, as in several cases discussed above. This will occur, for example, whenever group calibration is being violated because the system is designed to implement affirmative action in response to historical injustice.

The evidence account thus provides a better understanding of the limitations of fairness criteria, particularly in their use as constraints on ML optimization.

4.4. Related work

To further clarify the account, I will contrast the evidence account with some related ideas.

One thing that sets the evidence account apart is that it emphasizes the *value* of using the criteria to obtain evidence. Some philosophers and other fair ML researchers have mentioned, typically in passing, the possibility that violations of some fairness criteria can serve as "weak evidence" of unfairness. However, this is often mentioned as a contrast with a researcher's preferred criterion, e.g., in purported contrast with group calibration (Corbett-Davies et al. 2023; Eva 2022; Hedden 2021; Long 2021). Alternatively, the idea is occasionally mentioned as a concession after pointing out various problems with using the criteria, e.g., in (Fazelpour & Lipton 2020; Fazelpour et al. 2022; Herington & Glymour 2019; Vredenburgh 2024).

Hellman (2020) offers a view somewhat similar to the evidence account for some kinds of fairness criteria. However, Hellman's account differs in what she takes various criteria to be evidence *for*. She argues that calibration by group is evidence for believing an individual is or is not in the positive class. Meanwhile, she thinks other criteria (including false positive rate parity) offer evidence regarding what we ought to do. The evidence account, in contrast, proposes that all the proposed fairness criteria should be understood as evidence regarding the fairness of an AI system.

Other researchers have mentioned the idea of fairness criteria as evidence in a more favorable light, without endorsing the claim that fairness criteria primarily serve as evidence. Jacobs and Wallach (2021) are concerned with fairness criteria understood as measurement modeling constructs. On their account, each criterion can be assessed using the social science tools of construct validity and construct reliability. I take this work to be complementary: measurements are one kind of evidence. Friedler et al. (2021) provide another related discussion. Their framework treats different fairness requirements as reflecting different assumptions about the relationship between our knowledge of the world (what they call the *observed space*) and the underlying features of the world (what they call the *construct space*).[28] Castro (2022) briefly suggests that fairness criteria can serve as heuristics for detecting unfairness, though I interpret his other work to endorse contextualism.

Some proponents of group calibration (or other related criteria) have admitted that violations of FPR parity by risk assessment systems do indicate past injustice in the sociotechnical system in which the algorithm is embedded (Corbett-Davies et al. 2023; Eva 2022; Loi & Heitz 2022; Long 2021; Simoiu et al. 2017). However, these researchers often wish to distinguish this sociotechnical injustice from the "intrinsic" or "inherent" bias of an algorithm itself, which they suggest is operationalized by their preferred criterion.

I think this is a mistake. Fairness criteria are not used for detecting bias in some morally neutral sense. Whether any system is unfair or discriminatory—i.e., biased in a morally pernicious way—will depend on the social system it is embedded in. This is helpfully demonstrated by the valence reversal cases discussed above (§3). Moreover, we should be primarily concerned with the impact of an algorithmic system on justice more broadly (Hellman 2023). Despite these points, the evidence account is compatible with making various distinctions among kinds of fairness—e.g., between intrinsic unfairness and injustice (Eva 2022), or between "prediction-fairness" and "decision-fairness" (Beigang 2022)—for those who prefer to do so.

5. Contextualism

Contextualism is an alternative to the evidence account for explaining the value of algorithmic fairness criteria. It aims to salvage the idea that algorithmic fairness

28. Note, however, that Friedler et al. argue that individual fairness and group fairness metrics differ in their assumptions about the relation between these two spaces, and are thereby incompatible. The evidence account does not make the commitments necessary to force that conclusion, so is compatible with deploying both group and individual fairness criteria in different contexts.

conditions are necessary conditions for fairness by suggesting that each condition is only operative in certain contexts.[29] There isn't space here to offer a complete argument against this view, nor to offer a full accounting of what I take to be the advantages the evidence account has over it. Instead, I will discuss two significant problems for contextualism that are not shared by the evidence account. This will provide preliminary reasons to prefer the evidence account, though a full evaluation of the two views' comparative advantages will require further research. The first worry concerns contextualism's commitment to—and motivation for—pluralism about fairness. The second concerns the view's continued vulnerability to counterexample.

According to contextualism, different fairness criteria accurately operationalize or capture different notions of fairness. These different notions of fairness are operative in different contexts. The criteria serve as conditions on achieving fairness that correspond to—and operationalize—these different notions. The motivation for contextualism is quite similar to the motivation of the evidence account: to vindicate the intuitively valuable use of fairness criteria, despite the incompatibility results and despite the counterexamples discussed above. However, contextualists are still committed to understanding fairness criteria as conditions—typically as necessary conditions.[30]

According to contextualism, in a particular context, FPR parity is a necessary condition for fairness, while in another context group calibration is a necessary condition. The fact that the two cannot be satisfied at the same time is not a problem as they operationalize distinct notions of fairness that are operative in distinct contexts. It is a familiar point that different moral reasons or principles are applicable in different circumstances. Moreover, distinct moral reasons or principles can conflict with one another. Contextualists claim that the fairness principles applicable in cases like COMPAS—i.e., of recidivism risk scoring for pre-trial detention decisions—are distinct from the fairness principles applicable in other cases—e.g., college admissions. In some cases, multiple kinds of fairness are intuitively important, but cannot be mutually satisfied, so we must pick one or the other. Thus, contextualism purports to explain away problems of incompatibility.

29. This view has been defended explicitly by (Castro et al. 2023; Loi et al. 2024; Castro & Loi 2022; and Loi & Heitz 2022), among others. See also (Heidari et al. 2019). It also is a plausible interpretation of how many other researchers see fair ML criteria relating to fairness and bias. For examples of this, see e.g., (Verma & Rubin 2018; Dwork et al. 2012; Rodolfa et al. 2020; Saleiro et al. 2018; Saleiro et al. 2020; and Barocas et al. 2023).

30. Note that contextualism goes beyond *contextual sensitivity*: the claim that we should be sensitive to context in selecting which fairness criterion to deploy. Contextualism claims that fairness has different necessary conditions in different contexts and that fairness criteria must serve to operationalize these conditions. The evidence account is also committed to contextual sensitivity but offers a distinct explanation for it. Moreover, this explanation is derived from the uncontroversial fact that evidence itself is context-sensitive.

At first glance, contextualists also have a ready response to the issue of the counterexamples discussed above: The counterexamples are cases where the criterion in question is not the relevant one for the context. In fact, the explicit proponents of contextualism typically make use of counterexamples to argue for their position (Castro 2022; Loi & Heitz 2022; Loi et al. 2024). The counterexamples motivate contextualism because they show that a simple condition-based view doesn't work (see §3).

Contextualism represents a serious competitor to the evidence account. However, I would suggest that the commitment to a condition-based understanding of the fairness criteria is still a mistake, one that leads the view into the two difficulties at issue.

The first problem for contextualism is that it requires pluralism about the nature of fairness as a way of vindicating the use of fairness criteria in light of the counterexamples. Not only does this mean the view is incompatible with extant univocal views of fairness, it also involves what might seem like the wrong kind of motivation for this pluralism. It is motivated by a desire to vindicate the technical fairness criteria, rather than by appeal to independent moral argument.

Essentially, contextualism requires pluralism about fairness because it is attempting to accommodate the practice of using statistical fairness criteria. The motivation for pluralism here is the failure of the fairness criteria to be compatible, and their failure to serve as general conditions on fairness in light of the counterexamples. However, this seems like an ad hoc maneuver to protect the fairness criteria. We are led to heavyweight, normative and metaethical claims about fairness in response to the fact that certain simple statistical criteria fail to be universally necessary (or sufficient) conditions on fairness. The worry, in sum, is that we are being moved to make significant changes in our view about fairness for weak reasons.

I'm tempted to think that there is a univocal sense of "fair" as it is used in ordinary language that philosophers have been attempting to capture in their theorizing.[31] However, even if there are genuine distinctions between kinds of fairness—or between justice and fairness, or direct and indirect discrimination—these more traditional kinds of pluralism are motivated by considerations other than saving the statistical fairness criteria. They track distinctions that are more obviously present in ordinary discourse and in ethical and political theory. Moreover, the more limited traditional pluralisms may not provide enough distinctions to vindicate the use of the wide variety of fairness criteria, nor to explain away the wide variety of counterexamples to them. For instance, it is unclear that the distinction between direct and indirect discrimination—or a purported distinction between fairness and justice—would

31. For instance in (Rawls 1971; 2001; Scanlon 2000).

explain why calibration by group fails to be a condition of fairness in the context of the *Gerrymandering — Non-Necessity* case (§3.2.1), while base-rate-tracking is such a requirement. This last point is not a knock-down objection, but contextualists who wish to justify their view by appeal to existing pluralisms or distinctions owe us additional argument showing this can be accomplished.[32]

A second, and more serious, worry for contextualism stems from the counterexample types from §3. As noted, the counterexamples initially seem to provide motivation for the contextualist view. But the counterexamples also come in a wide variety of recognizable categories. They are also easy to produce new versions of. This means that we have recipes for building counterexamples in a wide range of different contexts for any criterion conceived of as a necessary or sufficient condition.[33]

The recipes work like this: First, we select a case that contextualists use to motivate their view. Then, we hold the general context of the case fixed, but change details of the situation. That is, we select an algorithmic system that is used for the very same task as in the original example. Then, we modify the example by changing details concerning things like the results of the algorithmic system or the background motivations of those deploying it. Suggestions for just how to change these details can be derived from the different types of counterexamples cataloged in §3—e.g., randomization, brute force, gerrymandering, etc. The trick is to ensure that the background context and situation remain recognizably the same, but that the intuitive verdict about whether the situation is fair changes.

These recipes allow us to generate counterexamples in almost any context. Moreover, they show that sometimes a criterion cannot serve as a necessary (or sufficient) condition for fairness in a context, *even though it is the right criterion to use in that context*. Together, these points give us reason to doubt that the criteria should be conceived of as necessary or sufficient conditions, even within specified contexts.

Loi et al. (2024) offer one of the most plausible versions of contextualism. Their discussion provides several useful illustrating cases to which we can apply the counterexample recipes. According to Loi et al., the underlying principle of fairness that all fairness criteria are meant to capture is a luck egalitarian-style principle they call Fair Equality of Chances (FEC).[34] FEC claims roughly

32. Castro et al. (2023) provide what I take to be the best extant attempt at this argument.

33. Here I am inspired by Zagzebski's recipe for creating Gettier counterexamples (1994).

34. This commitment to FEC as an underlying, univocal notion of fairness helps this version of contextualism avoid some of the motivation problems raised above. However, this appeal to a univocal underlying account of fairness means the view has less flexibility in avoiding counterexamples than more pluralist views. Grant (2023) offers a related account that interprets Equalized Odds, a separation-based criterion, in a way that makes it closely akin to luck egalitarian equality of opportunity principles.

that "a procedure is unfair if it allocates advantages and disadvantages to the predictable advantage or disadvantage of certain groups, across classes of equally deserving* individuals" (2024: 26–27). Their formal explication of this view is clever and nuanced, but it roughly amounts to the idea that people should have equal chances of receiving advantages based on what they deserve, regardless of their membership in socially salient but morally arbitrary social groups (i.e., race, gender, disability status). Desert* here is a term meant to pick out whatever morally justifies different treatment of people, be it need, effort, or some other consideration.

Loi et al.'s account is contextualist because which fairness criteria corresponds to FEC—and thus serves as a necessary condition on fairness—depends on context. The context determines which criterion operationalizes FEC. It does so by specifying what counts as the advantages (or disadvantages) being distributed by the decision, what features determine desert*, and which groups are relevant. They argue that in some cases a separation-based criterion—one which entails FPR parity—is the right operationalization of FEC, and so a necessary condition on fairness. The same is true for sufficiency (and so group calibration) in other cases.

We can apply the counterexample recipes to the cases that Loi et al use to motivate their account. For instance, their primary case for motivating separation-based criteria (including FPR parity) concerns distributions of cash assistance to help children stay in school (2024: 32). They imagine a simple decision procedure that uses a binary classification system. The classifier is designed to predict whether each student needs assistance, i.e., whether a student is likely to drop out of school without additional help. They assume, for simplicity, that the value of the cash assistance for each student who receives it is the same. In such a scenario, they argue, separation is a necessary condition on fairness. There should be equal false positives for, e.g., white and Black students. This seems plausible: otherwise, the burdens of mistakenly missing out on the cash assistance will be borne by an already marginalized group. Other things being equal, we should expect that a fair decision procedure will satisfy separation in such a case.[35]

Despite the plausibility of separation-based criteria in this case, we can still generate counterexamples for separation in cases *within the very same context*. These are not counterexamples to the value of using FPR parity as a criterion. Rather, they are counterexamples to thinking it serves as a necessary condition for fairness in this case.

Consider another case that is exactly like the one above, except that the school is committed to using affirmative action-type interventions to help students. That is, it wants the cash assistance program to be targeted especially at helping

35. This judgment is controversial, as Loi et al. recognize (2024: 33). The evidence account is not committed to this particular judgment.

Black students, as a way of promoting diversity and redressing historical wrongs. Everything about the surrounding context is the same except for this explicit aim. In this case, separation will be violated. By giving out a disproportionate amount of cash assistance to Black students, the false positive rate for Black students (i.e., the rate of students who receive assistance but still drop out) will naturally rise. Yet this procedure seems entirely fair.

Note that, without foreknowledge that the school is aiming at affirmative action, learning that there is a violation of FPR parity will still serve as evidence of unfairness. This evidence may be overridden—i.e., face rebutting defeat—as in the case I just offered. But that doesn't mean it is the wrong criterion to appeal to. In most cases like this one, learning there is a violation of FPR parity will be a useful indicator of unfairness. So, the evidence account can accommodate both the fact that this criterion is the intuitive one to appeal to, and the fact that there are cases where it can be violated despite the situation being entirely fair. Contextualism, however, cannot make the same claim.

Moreover, there are other counterexamples to the contextualized criterion that we can appeal to here. The counterexample types discussed in §3 provide recipes for building counterexamples out of cases like the one Loi et al imagine. One could easily build a brute force counterexample that simply awards cash assistance to all the students. Or one could build an Eva-style gerrymandering case, where there is disparity between sub-groups despite there being no disparity between Black and white students. We can similarly appeal to our taxonomy of counterexample types to develop cases that show the non-necessity of other criteria in other cases.[36]

The problem with contextualism is that it still treats the criteria as providing conditions on fairness. This suggests a modification that produces another potential competitor to the evidence account. I will call this the contributory view.[37] The basic idea is that we should treat satisfaction of a criterion as contributing to fairness and violation as contributing to unfairness. In other words, satisfaction of a condition is a *pro tanto*, fairness-based reason for using an AI system, while violation is a similar kind of reason against using the system. Which criteria are relevant for providing such reasons will depend on context.

36. One response the contextualist could make to this objection is to claim that the cases I imagine are in a different context than the original cases. After all, I have changed the background circumstances. However, note that the cases I imagined above keep the very features that Loi et al. take to be relevant to determining a context: what the advantages being distributed are, what the justifying features are, and what the relevant sensitive social groups are. More importantly, I think this is an unhelpful response for the contextualist since it makes the position seem even more ad hoc: it now requires even more fine-grained distinctions between contexts in order to accommodate intuitions about cases. Meanwhile, the evidence account has no difficulty explaining what is going on in these cases, in a way that is independently motivated.

37. Thanks to an anonymous reviewer for pointing out the relevance of this possible view.

The contributory view is like the evidence view in that it appeals to *pro tanto* (and *prima facie*) reasons to avoid the problems associated with a condition-based view of the criteria. The views are also similar in that context sensitivity is baked into the background normative views being appealed to.[38]

However, I think the evidence view does a better job of respecting the nature of the fairness criteria. The criteria concern simple facts about distributions of outcomes. They seem poor candidates for explaining the nature of what is wrong with biased algorithmic systems. The very same distribution of outcomes can result from different kinds of processes. The moral acceptability of the distribution depends on background features of the case such as historical oppression, culpable negligence in addressing oppression, and even intentional discrimination. The relevant moral reasons concern discrimination, equal respect, and democratic legitimacy. Insofar as mere statistical frequencies are relevant to these things, it is in providing an indication of whether they are satisfied. This is the root of the problem for contextualism, too.

These considerations again cannot provide the full story about how the contextualist and contributory views should be evaluated. However, I think they give good preliminary reasons to prefer the evidence account for further pursuit and research.

6. Conclusion

The evidence account vindicates the use of fairness criteria. It makes sense of how these criteria can be useful for recognizing unfairness, despite the fact that they are incompatible in ordinary circumstances and face a slew of counterexamples. The account also helps make sense of the limitations of these criteria. It suggests that there is a straightforward way of using the criteria to evaluate or audit algorithmic decision-making systems. However, it also suggests that the use of such criteria for constraining ML optimization is more fraught. This is because such usage can amount to mere news management, rather than an effective intervention for promoting fairness and justice. Furthermore, the evidence account helps make sense of why context matters to determining which criteria to appeal to. It does this

38. Indeed, if a prominent view about the nature of reasons is correct, then the evidence account and the contributory account are essentially identical. According to the *reasons as evidence* account (Kearns & Star 2009), a proposition *P* is a reason to Φ iff *P* is evidence that one ought to Φ. However, many find this view unsatisfactory, as they think practical/moral reasons should explain why an action ought to be done (Brunero 2018). If that is right, the evidence and contributory accounts are distinct. In fact, this worry is the same as the one I raise below for the contributory account. So, if the worry isn't a problem for the contributory view, this is a reason to take the reasons as evidence view to be correct, and thus the contributory view simply is the evidence view.

without the difficulties with counterexamples and motivation that contextualism suffers from.

The argument here is not meant to suggest that fairness criteria provide everything needed to ensure that algorithmic decision-making is fair. That is very far from the case, as other fair ML researchers have convincingly argued (Fazelpour & Danks 2021; Fazelpour et al. 2022; Mayson 2019; Zimmermann & Lee-Stronach 2021). Efforts to achieve fairness and justice will require much more significant changes to both our technology and the social systems it is embedded in. The evidence account just provides a way of making sense of how (and when) statistical fairness criteria can contribute to that effort, rather than distracting from or impeding it.

Acknowledgments

For helpful comments, I would like to thank Scott Alfeld, John Basl, Peter van Elswyk, Branden Fitelson, Megan Feeney, Tim LaRock, David Liu, Lisa Miracchi, and Zohair Shafi. I would also like to thank audiences at Northeastern University, the University of Pennsylvania, the Milan Workshop on Fairness and Machine Learning in Healthcare, and the 2023 Pacific Division Meeting of the APA. Brett Karlan and John Sullins provided particularly helpful feedback as commentators at the Pacific APA. Special thanks to Tina Eliassi-Rad for repeated readings and much helpful advice.

References

Angelino, Elaine, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin (2018). Learning Certifiably Optimal Rule Lists for Categorical Data. *Journal of Machine Learning Research*, *18*(234), 1–78.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner (2016). Machine Bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Arneson, Richard (2015). Equality of Opportunity. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2015 edition. https://plato.stanford.edu/archives/sum2015/entries/equal-opportunity/.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan (2023). *Fairness and Machine Learning: Limitations and Opportunities.* fairmlbook.org. http://www.fairmlbook.org.

Beigang, Fabian (2022). On the Advantages of Distinguishing Between Predictive and Allocative Fairness in Algorithmic Decision-Making. *Minds and Machines*, *32*(4), 655–682.

Beigang, Fabian (2023). Reconciling Algorithmic Fairness Criteria. *Philosophy and Public Affairs*, *51*(2), 166–190.

Benjamin, Ruha (2019). *Race After Technology: Abolitionist Tools for the New Jim Code.* Wiley.
Brunero, John (2018). Reasons, Evidence, and Explanations. In Daniel Star (Ed.), *The Oxford Handbook of Reasons and Normativity* (321–341). Oxford University Press.
Buolamwini, Joy and Timnit Gebru (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* (77–91).
Castro, Clinton (2022). Just Machines. *Public Affairs Quarterly*, *36*(2), 163–183.
Castro, Clinton, David O'Brien, and Ben Schwan (2023). Egalitarian Machine Learning. *Res Publica*, *29*(2), 237–264.
Castro, Clinton and Michele Loi (2022). The Fair Chances in Algorithmic Fairness: A Response to Holm. *Res Publica*, 1–7.
Chouldechova, Alexandra (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, *5*(2), 153–163.
Chouldechova, Alexandra and Aaron Roth (2018). The Frontiers of Fairness in Machine Learning. *arXiv preprint arXiv:1810.08810*. https://arxiv.org/abs/1810.08810.
Corbett-Davies, Sam, Johann D Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel (2023). The Measure and Mismeasure of Fairness. *Journal of Machine Learning Research*, *24*(312), 1–117.
Dressel, Julia and Hany Farid (2018). The Accuracy, Fairness, and Limits of Predicting Recidivism. *Science Advances*, *4*(1), eaao5580.
Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel (2012). Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (214–226).
Eliassi-Rad, Tina (2020). Just Machine Learning. Colloquium, Santa Fe Institute.
Eliassi-Rad, Tina and Branden Fitelson (2021). Exploring Impossibility Theorems for Algorithmic Fairness with PrSAT. Technical report, RADLAB, Northeastern University, Boston, MA. http://fitelson.org/exploring_impossibility.pdf.
Eva, Benjamin (2022). Algorithmic Fairness and Base Rate Tracking. *Philosophy and Public Affairs*, *50*(2), 239–266.
Fazelpour, Sina and David Danks (2021). Algorithmic Bias: Senses, Sources, Solutions. *Philosophy Compass*, *16*(8), 1–16.
Fazelpour, Sina and Zachary C. Lipton (2020). Algorithmic Fairness from a Non-Ideal Perspective. (57–63).
Fazelpour, Sina, Zachary C. Lipton, and David Danks (2022). Algorithmic Fairness and the Situated Dynamics of Justice. *Canadian Journal of Philosophy*, *52*(1), 44–60.
Fleisher, Will (2021). What's Fair About Individual Fairness? In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (480–490).
Flores, Anthony W., Kristin Bechtel, and Christopher T. Lowenkamp (2016). False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks. *Federal Probation*, *80*, 38–46.
Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian (2021). The (Im) Possibility of Fairness: Different Value Systems Require Different Mechanisms for Fair Decision Making. *Communications of the ACM*, *64*(4), 136–143.
Fullinwider, Robert (2018). Affirmative Action. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2018 edition. https://plato.stanford.edu/archives/sum2018/entries/affirmative-action/.

Gabriel, Iason, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez,, et al. (2024). The Ethics of Advanced AI Assistants. *arXiv preprint arXiv:2404.16244*. https://arxiv.org/abs/2404.16244.

Gonen, Hila and Yoav Goldberg (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings but Do Not Remove Them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (609–614).

Grant, David Gray (2023). Equalized Odds Is a Requirement of Algorithmic Fairness. *Synthese*, *201*(3), 1–25.

Hardt, Moritz, Eric Price, and Nati Srebro (2016). Equality of Opportunity in Supervised Learning. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, *29*, 3315–3323.

Hedden, Brian (2021). On Statistical Criteria of Algorithmic Fairness. *Philosophy and Public Affairs*, *49*(2).

Heidari, Hoda, Michele Loi, Krishna P. Gummadi, and Andreas Krause (2019). A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* (181–190).

Hellman, Deborah (2020). Measuring Algorithmic Fairness. *Virginia Law Review*, *106*, 811–866.

Hellman, Deborah (2023). Big Data and Compounding Injustice. *Journal of Moral Philosophy*, *21*(1–2), 62–83.

Herington, Jonathan and Bruce Glymour (2019). Measuring the Biases That Matter: The Ethical and Causal Foundations for Measures of Fairness in Algorithms. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, *2019*, 269–278.

Holm, Sune (2023a). Egalitarianism and Algorithmic Fairness. *Philosophy and Technology*, *36*(1), 1–18.

Holm, Sune (2023b). The Fairness in Algorithmic Fairness. *Res Publica*, *29*(2), 265–281.

Hutchinson, Ben and Margaret Mitchell (2019). 50 Years of Test (un) Fairness: Lessons for Machine Learning. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* (49–58).

Jacobs, Abigail Z. and Hanna Wallach (2021). Measurement and Fairness. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* (375–385).

Kearns, Michael and Aaron Roth (2019). *The Ethical Algorithm.* Oxford University Press.

Kearns, Michael, Aaron Roth, and Zhiwei Steven Wu (2017). Meritocratic Fairness for Cross-Population Selection. In *Proceedings of the 34th International Conference on Machine Learning*, *PMLR 70* (1828–1836).

Kearns, Michael, Seth Neel, Aaron Roth, and Zhiwei Steven Wu (2018). Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning*, *PMLR 80* (2564–2572).

Kearns, Stephen and Daniel Star (2009). Reasons as Evidence. *Oxford Studies in Metaethics*, *4*, 215–42.

Kelly, Thomas (2016). Evidence. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2016 edition. https://plato.stanford.edu/archives/win2016/entries/evidence/.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv preprint arXiv:1609.05807*. https://arxiv.org/abs/1609.05807.

Koons, Robert (2022). Defeasible Reasoning. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2022 edition. https://plato.stanford.edu/archives/sum2022/entries/reasoning-defeasible/.

Lewis, David (1981). Causal Decision Theory. *Australasian Journal of Philosophy*, *59*(1), 5–30.

Lin, Hanti (2022). Bayesian Epistemology. In Edward N. Zalta and Uri Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition. https://plato.stanford.edu/archives/fall2022/entries/epistemology-bayesian/.

Lippert-Rasmussen, Kasper (2022). Using (Un)Fair Algorithms in an Unjust World. *Res Publica*, *29*(2), 283–302.

Loi, Michele, Anders Herlitz, and Hoda Heidari (2024). Fair Equality of Chances for Prediction-Based Decisions. *Economics and Philosophy*, *40*(3), 557–580.

Loi, Michele and Christoph Heitz (2022). Is Calibration a Fairness Requirement? An Argument From the Point of View of Moral Philosophy and Decision Theory. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (2026–2034).

Loi, Michele, Francesco Nappo, and Eleonora Viganò (2023). How I Would have been Differently Treated. Discrimination Through the Lens of Counterfactual Fairness. *Res Publica*, *29*(2), 185–211.

Long, Robert (2021). Fairness in Machine Learning: Against False Positive Rate Equality as a Measure of Fairness. *Journal of Moral Philosophy*, *19*(1), 49–78.

Longino, Helen E. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry.* Princeton university press.

Mayson, Sandra G. (2019). Bias in, Bias Out. *The Yale Law Journal*, 2218–2300.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), 1–35.

Narayanan, Arvind (2018). Translation Tutorial: 21 Fairness Definitions and their Politics. In *The ACM Conference on Fairness, Accountability, and Transparency*. https://www.youtube.com/watch?v=jIXIuYdnyyk.

O'Neil, Cathy (2016). *Weapons of Math Destruction.* Crown.

Ongun, Talha, Timothy Sakharaov, Simona Boboila, Alina Oprea, and Tina Eliassi-Rad (2019). On Designing Machine Learning Models for Malicious Network Traffic Classification. *arXiv preprint: arXiv:1907.04846*. https://arxiv.org/abs/1907.04846.

Parfit, Derek (2002). Equality or Priority? In Matthew Clayton and Andrew Williams (Eds.), *The Ideal of Equality* (81–125). New York: Palgrave Macmillan.

Pratchett, Terry (2013). *Men at Arms.* Random House.

Rawls, John (1971). *A Theory of Justice.* Harvard University Press.

Rawls, John (2001). *Justice as Fairness: A Restatement.* Harvard University Press.

Rodolfa, Kit T., Erika Salomon, Lauren Haynes, Iván Higuera Mendieta, Jamie Larson, and Rayid Ghani (2020). Case Study: Predictive Fairness to Reduce Misdemeanor Recidivism through Social Service Interventions. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* (142–153).

Russell, Stuart J. and Peter Norvig (2020). *Artificial Intelligence: A Modern Approach.* Pearson Education, 4th edition.

Saleiro, Pedro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani (2018). Aequitas: A Bias and Fairness Audit Toolkit. *arXiv preprint arXiv:1811.05577*. https://arxiv.org/abs/1811.05577.

Saleiro, Pedro, Kit T. Rodolfa, and Rayid Ghani (2020). Dealing with Bias and Fairness in Data Science Systems: A Practical Hands-on Tutorial. In *SIGKDD 2020*.

Scanlon, Thomas (2000). *What We Owe to Each Other.* Belknap.

Simoiu, Camelia, Sam Corbett-Davies, and Sharad Goel (2017). The Problem of Infra-marginality in Outcome Tests for Discrimination. *The Annals of Applied Statistics*, *11*(3), 1193–1216.

Sober, Elliott (2008). *Evidence and Evolution: The Logic Behind the Science.* Cambridge University Press.

Stanton-Ife, John (2022). The Limits of Law. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2022 edition. https://plato.stanford.edu/archives/spr2022/entries/law-limits/.

Taylor, Robert S. (2009). Rawlsian Affirmative Action. *Ethics*, *119*(3), 476–506.

Verma, Sahil and Julia Rubin (2018). Fairness Definitions Explained. In *Proceedings of the International Workshop on Software Fairness* (1–7).

Viganò, Eleonora, Corinna Hertweck, Christoph Heitz, and Michele Loi (2022). People are Not Coins: Morally Distinct Types of Predictions Necessitate Different Fairness Constraints. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* (2293–2301).

Vredenburgh, Kate (2024). Fairness. In Justin B. Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew M. Young and Baobao Zhang (Eds.), *The Oxford Handbook of AI Governance*. Oxford University Press.

Williamson, Timothy (2002). *Knowledge and Its Limits.* Oxford University Press.

Zagzebski, Linda (1994). The Inescapability of Gettier Problems. *Philosophical Quarterly*, *44*(174), 65–73.

Zimmermann, Annette and Chad Lee-Stronach (2021). Proceed with Caution. *Canadian Journal of Philosophy*, (1), 6–25.