Does the End Justify the Means?

On the Moral Justification of Fairness-Aware Machine Learning

HILDE WEERTS, Eindhoven University of Technology, The Netherlands
LAMBÈR ROYAKKERS, Eindhoven University of Technology, The Netherlands
MYKOLA PECHENIZKIY, Eindhoven University of Technology, The Netherlands

Despite an abundance of fairness-aware machine learning (fair-ml) algorithms, the moral justification of how these algorithms enforce fairness metrics is largely unexplored. The goal of this paper is to elicit the moral implications of a fair-ml algorithm. To this end, we first consider the moral justification of the fairness metrics for which the algorithm optimizes. We present an extension of previous work to arrive at three propositions that can justify the fairness metrics. Different from previous work, our extension highlights that the consequences of predicted outcomes are important for judging fairness. We draw from the extended framework and empirical ethics to identify moral implications of the fair-ml algorithm. We focus on the two optimization strategies inherent to the algorithm: group-specific decision thresholds and randomized decision thresholds. We argue that the justification of the algorithm can differ depending on one's assumptions about the (social) context in which the algorithm is applied - even if the associated fairness metric is the same. Finally, we sketch paths for future work towards a more complete evaluation of fair-ml algorithms, beyond their direct optimization objectives.

1 INTRODUCTION

Despite an abundance of fairness-aware machine learning (fair-ml) algorithms, there is still little guidance on the suitability of different approaches in practice. In most studies, fair-ml algorithms are evaluated in terms of their direct optimization objectives: the predictive performance and fairness of the resulting machine learning model, quantified as a set of metrics. A growing body of work considers the circumstances and moral frameworks under which we can justify the use of these fairness metrics [7, 8, 14, 18, 20, 27]. In contrast, the moral justification of *how* fair-ml algorithms optimize for fairness metrics remains largely unexplored. Importantly, there are various ways in which the optimization problem can be approached with varying side effects and final outcomes. The aim of this paper is to consider a new research question: under which circumstances is the way in which fair-ml algorithms optimize for fairness metrics morally justifiable?

Fairness-aware machine learning algorithms can be roughly subdivided into three categories. *Pre-processing* algorithms adjust training data directly to mitigate downstream model unfairness [13, 23, 37]. *Constrained learning* approaches incorporate fairness constraints into the machine learning process, either by directly incorporating a constraint in the loss function [23, 36, 38] or by learning an ensemble of predictors [1]. *Post-processing* techniques make adjustments to existing machine learning models to satisfy fairness constraints, either by adjusting the parameters of a trained model directly [24] or by post-processing the predictions of the model [17].

In this paper, we will analyze the moral implications of the post-processing algorithm proposed by Hardt et al. [17]. We have chosen this algorithm due to its popularity as a bench-marking algorithm (e.g., in [1]) as well as its simplicity. This allows us to analyze the algorithm as a decision-making policy even in absence of a specific data set.

The main contributions of this work are as follows. First, we present three propositions that can morally justify the use of three popular fairness metrics for which the post-processing technique can optimize. To this end, we extend the framework introduced by Hertweck et al. [20], which distinguishes between different causes of (observed) inequality. We argue that the original framework is insufficient to judge fairness, because it does not consider the consequences of predicted outcomes. Second, to the best of our knowledge, we are the first to present a moral justification of a fair-ml

algorithm. In particular, we demonstrate that the defensibility of the post-processing algorithm can differ depending on one's assumptions about the (social) context in which the algorithm is applied - even if the associated fairness metric is the same.

The remainder of this paper is structured as follows. In Section 2, we introduce three running examples that will be used throughout the paper. In Section 3, we set the stage for our discussion of the post-processing algorithm by analyzing the justification of the fairness metrics for which it optimizes. We first present the three fairness metrics. Next, we introduce Hertweck et al.'s framework [20] and propose an extension that considers the utility of predicted outcomes. We use the extended framework to arrive at three propositions that can justify the use of fairness metrics. In Section 4, we discuss technical details of the post-processing algorithm proposed by Hardt et al. [17]. With all relevant components introduced, we turn to a discussion on the moral implications of the two main optimization strategies inherent to the post-processing algorithm: (1) group-specific decision thresholds (Section 5), and (2) randomized decision thresholds (Section 6). Finally, we sketch several paths for future work (Section 7) and conclude our findings (Section 8).

2 RUNNING EXAMPLES

Our arguments will be illustrated using three running examples. Albeit highly stylized, we believe these examples highlight key differences across scenarios that are relevant from both ethical and practical perspectives.

2.1 Resume Selection

In *resume selection*, we consider a machine learning model for selecting job applicants for interviews based on their resumes. Due to resource constraints, only a limited number of interviews can take place. Consequently, the problem is formulated as a ranking problem where the top-ranking candidates are selected for the interview round. We define a candidate to belong to the *positive* class when they have the qualities to be a high-performing employee. The model is trained to predict appraisal scores of current employees, based on their resume at the time of application. We further assume that all candidates are sufficiently qualified such that they would benefit from getting the interview.

2.2 Lending

The *lending* scenario considers a machine learning model that predicts whether an applicant will default on the loan within the first year, based on the characteristics of the requested loan and prospective borrower. A *positive* prediction corresponds to the applicant receiving the loan and a *negative* prediction to a rejection. The predictions will be used to automatically accept or reject loan applications. We further assume that there is no direct resource constraint, but that the lender does want to balance the false positives and false negatives to manage the overall risk. Finally, we assume that it is generally beneficial to get a loan if an applicant is able to pay off the loan in time (*true positive*), but harmful for an applicant who will default (*false positive*).

2.3 Disease Detection

In the *disease detection* scenario, it is assumed that a machine learning model is developed to predict whether somebody is highly likely to get a deadly disease. We use *positives* to refer to patients who have the disease and *negatives* to patients who do not. The model is trained using historical patient data and is used to determine whether a patient will receive preventive treatment. We assume that the treatment has severe side effects. That is, similar to our *lending* scenario, the patient benefit depends on the accuracy of the prediction.

3 A FRAMEWORK FOR THE JUSTIFICATION OF FAIRNESS METRICS

To set the stage for our discussion of the post-processing algorithm proposed by Hardt et al. [17], we will first consider the moral justification of the three group fairness metrics for which it can optimize. We start by formally defining the three metrics (Section 3.1). Then, we will introduce the framework proposed by Hertweck et al. [20] (Section 3.2), which distinguishes between different causes of (observed) inequalities. In Section 3.3, we propose an extension of the framework that allows us to also consider the consequences of a particular predicted outcome. We use the extended framework to come to three propositions that can justify the fairness metrics (Section 3.4). Finally, we discuss the limitations of the framework (Section 3.5).

3.1 Three Group Fairness Metrics

The majority of fair-ml algorithms, including the algorithm proposed by Hardt et al. [17], optimize for group fairness. Group fairness is a notion of fairness that requires group statistics over the model's predictions to be equal across (sub)groups, defined by one or more sensitive characteristics. Classic examples of sensitive characteristics are race and gender, but depending on the context other traits may be relevant. We consider three metrics: *demographic parity*, *equalized odds*, and *equal opportunity* [17]. We use the following notation: X denotes features, Y denotes the target variable, $\widehat{Y}$ denotes the model's predictions, and A the (set of) sensitive features. In accordance with our running examples, we limit the definitions to the binary classification scenario.

3.1.1 Demographic parity. Demographic parity requires that the *selection rate* is equal across sensitive groups. For a binary classifier, demographic parity is met if the following holds:

$$\Pr(\widehat{Y} = 1 \mid A = a) = \Pr(\widehat{Y} = 1 \mid A = a'). \tag{1}$$

for all $a, a' \in A$. For example, in the *lending* scenario, demographic parity is met if the proportion of applicants who receive a loan is equal across groups. Note that demographic parity does not depend on Y, which implies that when base rates differ between groups (i.e., $Pr(Y = 1 \mid A = 0) \neq Pr(Y = 1 \mid A = 1)$, this metric rules out a perfect predictor.

3.1.2 Equalized odds. Equalized odds enforces that the model's predictive performance is equal for all sensitive groups. In particular, it requires the *false positive rate* and *true positive rate* to be the same for each sensitive group. For a binary classifier, equalized odds is satisfied if the following holds:

$$\Pr(\widehat{Y} = y \mid A = a, Y = y) = \Pr(\widehat{Y} = y \mid A = a', Y = y) \tag{2}$$

for all $a, a' \in A$ and $y \in \{0, 1\}$. In the *disease detection* scenario, the false positive rate constitutes the proportion of people who will *not* get the disease (negatives) for which the model falsely predicts that they will get the disease (false positives). The true positive rate, on the other hand, amounts to the proportion of people that will get the disease (positives) that are correctly predicted to get the disease (true positives). As opposed to demographic parity, equalized odds does take into account the ground-truth variable Y, which implies that the observed outcomes are relevant for judging fairness.

3.1.3 Equal Opportunity. Equal opportunity is a weaker version of equalized odds that only requires the *true positive rate* to be equal across groups:

$$\Pr(\widehat{Y} = 1 \mid A = a, Y = 1) = \Pr(\widehat{Y} = 1 \mid A = a', Y = 1) \tag{3}$$

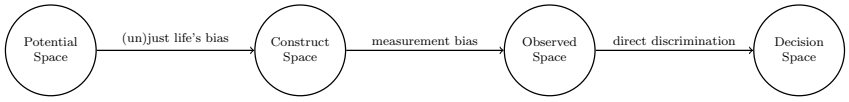
for all $a, a' \in A$. For example, in the *resume selection* scenario, equal opportunity would require the proportion of applicants who would be good employees (positives) that are selected for the interview (true positives) to be equal across groups, but would set no restrictions on the proportion of poor candidates that continue to the interviewing round. Consequently, equal opportunity assumes that only the positive outcome is important for judging fairness.

3.2 The Original Framework

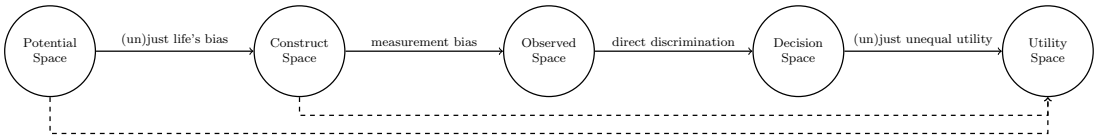
To arrive at a justification of the fairness metrics introduced in the previous section, we will build upon the framework proposed by Hertweck et al. [20], which is an extension of Friedler et al. [14]. A key insight of the framework is that the moral justification of demographic parity (Equation 1) depends on the *cause* of observed inequalities. Importantly, these causes cannot be deduced from the data alone and require one to make assumptions about the social context. Friedler et al. [14] refer to such assumptions as *worldviews*. To distinguish between different worldviews, the framework makes a distinction between four *spaces*.

- *Decision Space* (DS): represents the predictions of the machine learning model ($\widehat{Y}$).
- *Construct Space* (CS): consists of an individual's characteristics we would *like* to base decisions on.
- *Observed Space* (OS): consists of the measurements of the characteristics in the *Construct Space* that we *actually* base decisions on, i.e., the features present in a data set (X).
- *Potential Space* (PS): represents an individual's innate potential to develop the characteristics in the *Construct Space*.

For example, in *resume selection*, the *Decision Space* would consist of employee quality, a *Construct Space* attribute could be knowledge about the job, the *Observed Space* the number of years in a similar job, and the *Potential Space* a person's innate potential to become knowledgeable. Biases and direct discrimination can distort transformation from one space to another space (Figure 1(a)), three of which will be discussed in more detail below.



(a) The original framework by Hertweck et al. [20].



(b) Our extension. The utility space is a function of the *Decision Space* and may also depend on the *Potential Space* or *Construct Space*.

Fig. 1. The relationship between the different spaces and distortion mechanisms. Each mechanism can cause inequality between groups going from one space to the next space.

3.2.1 Direct Discrimination. According to Friedler et al. [14], direct discrimination refers to distortions between the *Observed Space* and the *Decision Space*. The intuition is that if two groups are, on average, equal in the *Observed Space*, inequalities in the *Decision Space* are due to differences in treatment by the predictive model.

3.2.2 Measurement Bias. Distortions between the *Construct Space* and *Observed Space* are a form of *measurement bias*: a systematic error in measurement. For example, a person's number of years of experience is a proxy for their actual knowledge (i.e., a measurement) and could over- or underestimate actual knowledge (i.e., the construct the proxy is supposed to measure). If measurement bias is associated with sensitive group membership, this can result in inequalities in the *Observed Space*. For example, social bias in past hiring decisions may have made it generally more difficult for women to get working experience in certain fields, compared to equally knowledgeable men.

3.2.3 Life's Bias. Distortions between the *Potential Space* and *Construct Space* are due to *life's bias*: inequalities in circumstances, such as income of a person's parents, that influence whether an individual's potential materializes [20]. For example, in *resume selection*, a lack of access to high-quality education could deprive a talented student of getting excellent grades. In *lending*, women might be "more likely to be single parents with unpredictable outgoings due to their dependants." [7]. Finally, in *disease detection*, some demographic groups may be at higher risk for developing the disease due to confounding factors such as poorer nutrition.

In the original framework, the PS, CS, and OS are restricted to the characteristics we would like to base decisions on, whereas the DS considers the model's predictions. In this work, we explicitly include (predicted) *outcomes* in the PS, CS, and OS. The reason for this is that a machine learning model is trained using features (X) *as well as* a target variable Y. As a result, measurement bias in the target variable can result in unfairness. For example, in *resume selection*, predicting appraisal scores is a proxy for predicting employee quality. However, appraisal scores do not fully capture the substantive nature of employee quality, threatening construct validity [34]. If historical appraisal scores are generated through biased evaluation practices, using the scores as a proxy for quality will constitute measurement bias. Borrowing the example of Binns [7], measurement bias could occur in our *lending* scenario if clerks are more lenient with repayment deadlines for men compared to women, resulting in more 'late' payments for women. In *disease detection*, the disease may have been historically under-diagnosed in some demographic groups, causing measurement bias in the observed prevalence of the disease. We refer to [22] for a more detailed overview of the relationship between measurement and fairness.

3.3 Extension: The Utility Space

The framework described by Hertweck et al. [20] is insufficient to judge the moral justification of demographic parity. In this section, we propose an extension to the framework that alleviates its main shortcoming.

The main intuition of Hertweck et al.'s framework is that the *causes* of observed inequalities, i.e., measurement bias and life's bias, are relevant for judging the morality of a fairness metric. Consequently, the final space in the original framework is the *Decision Space* (Figure 1(a)). Judging a fairness metric solely based on the distribution of predictions may be appropriate when predictions directly correspond to generally beneficial outcomes. For example, in the *resume selection* scenario, we assume that all candidates can benefit from a positive prediction, irrespective of their true class. However, as illustrated in [20], causes alone are insufficient when a predicted outcome cannot be considered universally beneficial. For example, in the *disease detection* scenario, clearly a positive prediction is only beneficial if it is a *true* positive - a false positive would unnecessarily expose the patient to severe side effects. Here, it seems impossible to judge the fairness of a particular distribution of predictions $\widehat{Y}$, without taking into account the ground truth class Y.

3.3.1 Utility Space. In order to consider the consequences of a predicted outcome, we propose the addition of a fifth space, the *Utility Space* (US). This space represents the *utility*, i.e., net benefit or harm, of a particular prediction for the

individual who is affected by the prediction (Figure 1(b)). In this work, we remain agnostic to an exact definition of utility, as it may differ between contexts and philosophical frameworks. In some scenarios, such as our *resume selection* example, the decision space directly affects the distribution of resources and we could directly consider resources as our measurement of utility. In *disease detection*, we may be more interested in the distribution of physical well-being. In other scenarios, we may consider the distribution of wealth, social welfare, or capabilities. More generally, one could define utility based on different fairness-related harms [10]: (1) allocation harms, which consider the distribution of resources and opportunities, and (2) quality-of-service harms, which consider the distribution of predictive performance. The question of how we should value (predicted) outcomes mirrors the 'equality of what?' debate in egalitarianism [6, 21].

A key component of the *Utility Space* is that, depending on how utility is defined, it may be a function of both the *Decision Space* and the *Potential Space* or *Construct Space* (Figure 1(b)). For example, in *disease detection*, a utility of physical well-being depends on whether somebody will actually get the disease (*Construct Space*) and the predictions (*Decision Space*). Here, negative utility can be interpreted as a quality-of-service harm. In *lending*, we may want to account for *life's bias* and consider utility as a function of the *Decision Space* and *Potential Space*. In other cases, such as *resume selection*, we may define utility solely based on whether somebody gets the interview (*Decision Space*). In this example, we consider negative utility as an allocation harm - although it remains difficult to determine utility without considering the entire recruitment pipeline.

Importantly, the *Utility Space* does not (necessarily) constitute a utilitarian approach to fairness, where the goal is to maximize the sum of individual utilities. Instead, the space serves to highlight that disparate predicted outcomes do not necessarily reflect disparate harms or benefits. How we evaluate a particular distribution of utilities depends on one's philosophical framework. Additionally, we acknowledge that the concept of 'utility' is controversial, as quantifying utility is notorious for the inherently political nature of assigning a (numerical) value to an outcome [8, 21, 28]. This issue is at least partially alleviated by our focus on *individual* utility for the person under classification, as opposed to more general utility of a particular outcome for society. For ease of presentation, we will make some general assumptions about the net benefit (or harm) of a particular outcome on a group level. We do not consider variance in utility due to variability of people and their preferences beyond group level differences. We believe this a reasonable simplification for the present paper, but future work may consider more fine-grained assumptions about utility in specific contexts.

3.3.2 Unequal Utility. On a group level, distortions between the *Decision Space* and *Utility Space* may arise if a particular outcome has, on average, a different utility across groups. Similar to life's bias, this can be due to inequalities in circumstances. In *resume selection*, defining utility in terms of whether an applicant gets an interview will not lead to distortions, as predictions directly correspond to utility. Instead, if we define utility as the opportunity to acquire wealth, a job interview plausibly provides a higher marginal utility for groups with higher unemployment rates. Notably, unequal utility in this case can exist even if we have accounted for life's bias in our hiring process. In *lending*, being approved a financial loan may be more beneficial for groups with, on average, fewer financial means. Similarly, in *disease detection*, a false positive will generally constitute negative utility, but the consequences to physical well-being may be even worse for groups who, on average, are in a poorer physical condition.

3.4 A Justification of Fairness Metrics

In this section, we will use the extended framework to set out three propositions that justify the use of demographic parity, equalized odds, and equal opportunity.

3.4.1 Worldviews. The framework by Hertweck et al. [20] allows us to distinguish between different *worldviews*: assumptions about the extent to which the distortions have occurred in a particular social context.

- *We're All Equal in the Construct Space* (WEA-CS): in the *Construct Space*, all groups look essentially the same and differences in the *Observed Space* are due to *measurement bias*.
- *We're All Equal in the Potential Space* (WEA-PS): in the *Potential Space*, all groups look essentially the same and differences in the *Construct Space* are due to *unjust life's bias*.
- *What You See Is What You Get* (WYSIWYG): there is no unjust difference between the *Potential Space* and the *Construct Space* nor between the *Construct Space* and *Observed Space*.

So how does this relate to fairness metrics? An important distinction between demographic parity on the one hand and equalized odds and equal opportunity on the other hand, is that demographic parity does not take into account ground-truth variable Y (Equation 1), whereas equalized odds (Equation 2) and equal opportunity (Equation 3) do condition on Y. A direct consequence of these different valuations of predicted outcomes is that demographic parity and equalized odds are mathematically incompatible [9, 26]: we can generally not achieve both demographic parity and equalized odds at the same time. Now, the main proposition in Hertweck et al. [20] is that if Y (which lies in the *Observed Space*) is either an unreliable measurement (WAE-CS) or affected by unjust life's bias (WAE-PS), the use of demographic parity can be justified.

In the remainder of this section, we will explain and extend this justification using the *Utility Space*. In particular, we will argue that the justification depends on how utility is defined: on which spaces does utility depend?

3.4.2 Measurement Bias and We're All Equal in the Construct Space. We first consider the justification of demographic parity under measurement bias. It seems intuitively unfair to make decisions based on incorrect measurements. If we assume that utility depends on the correctness of a measurement, this intuition is quite robust under different ethical theories. We formalize this assumption as *US(CS): utility is a function of the Decision Space and Construct Space*.

A Kantian decision-maker might argue that "making decisions based on the most valid data that can be obtained" is right, as they would be willing to make it a universal law. From a utilitarian perspective, if US(CS), valid measurements result in higher utility compared to invalid measurements. So as long as the cost of obtaining better measurements does not exceed this increase in utility, a utilitarian may argue that there is a moral obligation to correct for measurement bias. Finally, Hertweck et al. [20] appeal to the question of responsibility: individuals cannot be held responsible for incorrect measurements decision-makers have about them, so they should not have to bear the harmful consequences of incorrect measurements. Although not stated explicitly by Hertwerck et al., this argument implies that incorrect predictions compared to the *Construct Space* are harmful, i.e., US(CS).

From each of these arguments, it follows that if US(CS), we should not make decisions based on Y in the presence of measurement bias. Consequently, any fairness metric that conditions on Y is not justifiable, which is the case for equalized odds and equal opportunity. If we further assume that WAE-CS, i.e., the base rate in the *Construct Space* is equal across groups, we can argue that all groups are equally deserving of $\widehat{Y}$, justifying the use of demographic parity. In conclusion, we arrive at the following proposition:

Proposition 3.1. `IF WEA-CS AND US(CS) AND measurement bias, THEN demographic parity`

3.4.3 Life's Bias and We're All Equal in the Potential Space. We now turn to the justification of demographic parity under life's bias. The extent to which we should account for this type of bias is more controversial. Under which circumstances can life's bias be considered unjust? Evoked by Binns [6], Hertweck et al. [20] pose that the principles of

egalitarianism may provide guidance. Egalitarianism is a school of thought in political philosophy about distributive justice: the just allocation of rewards and costs. In the remainder of this section, we will summarize the arguments of Hertweck et al. [20] and present an extension grounded in the *Utility Space*.

An important concept within egalitarianism is equality of opportunity: the idea that (1) social positions should be open to all applications who possess the relevant attributes, and (2) all applicants must be assessed only on relevant attributes [2]. We can distinguish between several interpretations of equality of opportunity. A formal interpretation requires all people to formally get the opportunity to apply for a position. Additionally, direct discrimination based on arbitrary factors such as race or gender is prohibited. Substantive theories go one step further and pose that everyone should also get a substantive opportunity to *become* qualified. Two prominent substantive interpretations are Rawlsian equality of opportunity and luck egalitarianism. According to John Rawls' theory of justice [30], everyone with similar innate talent and ambition should have the same prospects for success, irrespective of their socio-economic background. Additionally, Rawls only considers social-economic inequalities acceptable if these inequalities benefit the most disadvantaged members of society. Luck egalitarianism [12], on the other hand, poses that inequalities are only just if they are the result of informed, free choice, as opposed to brute luck [6]. For example, taking a gamble with known risks is an informed choice, but being born into a rich household can be considered luck. Varieties of luck egalitarianism differ in how they distinguish between 'luck' and 'choice'.

Hertweck et al. [20] use the luck egalitarian distinction between 'luck' and 'choice' to distinguish between just and unjust life bias. If a distortion between the *Potential Space* and *Construct Space* is due to personal choice, this can be considered *just* life bias. If the distortion is due to circumstances (i.e., luck) this can be considered *unjust* life's bias. For example, in *lending*, we might assume that one's ability to pay off a loan is solely due to informed choices they have made in their life and conclude that inequality in the *Construct Space* is just (though one may be skeptical whether this is a well-founded assumption). Hertweck et al. [20] argue that if WAE-PS and differences in the *Construct Space* are due to unjust life's bias, demographic parity can be justified. Indeed, through an egalitarian lens, demographic parity assumes that individuals cannot be held accountable for their ground-truth class Y [18] and are equally deserving of $\widehat{Y}$. Equalized odds, on the other hand, assumes that all individuals with the same ground-truth class *can* be held equally accountable for their class [18] and, conditioned on $Y = 1$, all groups are equally deserving of $\widehat{Y} = 1$.

What the original framework fails to capture is that $\widehat{Y}$ does not always equal utility. In particular, if utility is a function of the *Construct Space*, accounting for unjust life's bias will *decrease* utility for disadvantaged groups. For example, in *disease detection*, the relevant outcome is whether somebody actually gets the disease (*Construct Space*) not their potential to get the disease (*Potential Space*). As such, we argue that demographic parity is only appropriate under WAE-PS and unjust life's bias, if utility is a function of the *Potential Space*. We specify the additional assumption *US(PS): utility is a function of the Decision Space and the Potential Space.* and arrive at the following proposition:

Proposition 3.2. `IF WEA-PS AND US(PS) AND unjust life bias, THEN demographic parity`

3.4.4 Unequal Utility and What You See is What You Get. We now turn to the final worldview: *What You See Is What You Get* (WYSIWYG). This worldview assumes that both measurement bias and unjust life's bias are absent. For the reasons outlined in Sections 3.4.2 and 3.4.3, optimizing for demographic parity under WYSIWYG is arguably not justifiable when US(CS): in absence of the biases, we can hold individuals with the same ground-truth class Y equally accountable for their class. Under WYSIWYG, optimizing for a metric that conditions on Y, such as equalized odds or equal opportunity, seems more appropriate.

Again, we pose that the *Utility Space* is relevant for a moral justification. Both equalized odds and equal opportunity assume that the negative utility of errors is equal across groups. However, in case of unequal utility, this may not be true. Similar to unjust life's bias, we can distinguish between unequal utility due to informed choice (*just* unequal utility) and unequal utility for which individuals should not be held accountable (*unjust* unequal utility). For example, the consequences of a false negative in *lending* are worse for an individual that generally has a harder time getting a loan due to unconscious gender bias, which is related to luck rather than choice. On the other hand, if for some reason an individual purposefully withholds payment, unequal utility could be considered just. We specify the additional assumption *We're All Equal in the Utility Space (WAE-US): there is no unjust unequal utility*. In conclusion, we propose the following rule:

Proposition 3.3. `IF WYSIWYG AND US(CS) AND WAE-US, THEN (equalized odds OR equal opportunity)`

Which of these two metrics is appropriate depends on how different types of errors are valued, which we can interpret as different sets of utilities. Equalized odds implicitly assumes that all types of errors are equally important, whereas equal opportunity only considers false negatives relevant for fairness. Consequently, optimizing for equal opportunity could result in arbitrarily large inequalities in terms of the false positive rate. In cases where a false negative constitutes a missed opportunity but false positives a lucky advantage, as in *resume selection*, equal opportunity might be defended by the non-maleficence principle, which obliges one to not inflict harm on others [28]. We leave a more fine-grained moral analysis of error-based group fairness metrics for future work.

3.5 Limitations

In the previous section, we have identified a set of conditions under which the three group fairness metrics introduced in Section 3.1 are justifiable, which will be the basis for our analysis of the post-processing algorithm in Section 5 and Section 6. However, the set is incomplete, due to both an incomplete coverage of all possible assumptions as well as fundamental limitations of group fairness metrics.

3.5.1 WAE-PS or WAE-CS does not hold exactly. It is unclear which fairness metric is appropriate in scenarios where either WAE-PS or WAE-CS does not hold exactly, yet unjust life's bias and measurement bias are present. In such cases, enforcing demographic parity seems unreasonably strict, yet error-based metrics are not justifiable either. Instead, more fine-grained assumptions about the underlying base rates, in either the *Potential Space* or *Construct Space*, are required. Modelling the extent of unjust life's bias is complicated by a central objection to luck egalitarian principles: factors labelled as 'luck' and 'choice' cannot always be reasonably separated [6, 28]. Similarly, it can be challenging to accurately model measurement bias.

3.5.2 Other valuations of outcomes. Equalized odds and equal opportunity are just two ways to value false positives and false negatives. In some cases, we may prefer to assign more fine-grained weights to different types of errors. If we assume unjust unequal utility, these weights may even be group-specific.

3.5.3 Disregard absolute utility. A more fundamental limitation of group fairness metrics is that it is assumed that the absence or presence of inequality is what matters for fairness. As such, group fairness metrics rely on a *relative* interpretation of utility, but do not set any constraints on *absolute* utility. As a result, fair-ml approaches that optimize for stricter group fairness criteria do not necessarily improve the absolute outcomes for the least well-off groups and may even decrease utility for all groups in order to achieve parity. For example, Hu and Chen [21] show that an SVM

classifier that enforces a smaller difference in selection rates can result in a lower number of overall resources available to the most disadvantaged group.

We can interpret the emphasis on inequality as a deontological notion of justice. Deontological ethical theories judge actions based on a set of predefined rules. In the case of fairness metrics, a deontologist may argue that inequality is wrong in and of itself. Consequentialist ethical theories, on the other hand, hold that the consequences of an action are what matters. We can view an emphasis on absolute utility as a more consequentialist perspective - albeit a limited one. A full consequentialist interpretation would consider the full range of consequences of predictions on individuals and society, including the consequences of inequality [8, 28].

4 GROUP-SPECIFIC RANDOMIZED DECISION THRESHOLDS

We will now detail the post-processing step introduced by Hardt et al. [17]. The goal of the algorithm is to identify a decision threshold[1] (t) of the model's predicted scores (R) such that the post-processed predictions ($\widetilde{Y}$) adhere to a given fairness constraint while retaining as much predictive performance as possible. By moving the decision threshold up or down, we can tweak the trade-off between (false) positives and (false) negatives. The post-processing technique uses two specific strategies to satisfy fairness constraints: group-specific thresholds and randomized thresholds.

4.1 Group-Specific Thresholds

If base rates differ across groups, it will often not be possible to identify one unique threshold t such that a fairness metric hold across all groups. In that case, Hardt et al. [17] propose to choose a separate threshold t_a for all sensitive groups $a \in A$. For example, in order to satisfy demographic parity (Equation 1), we could decrease the decision threshold for a group with a low selection rate, such that more instances are classified as positive. Similarly, lowering the decision threshold for a group will increase the true positive rate for that group (though at the cost of false positives), allowing us to achieve equal opportunity (Equation 3).

4.2 Randomized Thresholds for Equalized Odds

Unfortunately, group-specific thresholds are not always sufficient to achieve equalized odds. Recall that the main requirement of equalized odds is to ensure that the false positive rate and true positive rate is equal across groups (Equation 2). The trade-off between false positives and false negatives is often analyzed using a Receiver Operating Characteristic (ROC) curve, which sets out a classifier's false positive rate (*fpr*) against its true positive rate (*tpr*) over varying decision thresholds. Considering equalized odds, group-specific thresholds still limit us to the (*fpr*, *tpr*) combinations that lie on the intersection of group-specific ROC curves. In some cases, the group-specific ROC curves may not intersect or represent a poor trade-off between false positives and false negatives. To further increase the solution space, Hardt et al. [17] allow the decision thresholds to be randomized. That is, the decision threshold T_a is a randomized mixture of two decision thresholds $\underline{t}_a$ and $\overline{t}_a$ (Figure 2).

Randomization allows us to achieve *any* combination of (*fpr*, *tpr*) that lies within the convex hull of the ROC curve (Figure 3(a)). In cases where group-specific ROC curves do not intersect apart from trivial end points, the predictive performance of the model for the best-off group is artificially lowered through randomization until the performance is

[1]Many machine learning classification algorithms do not directly output a class, but a probability or score R, which indicates the confidence of the model that an instance belongs to a certain class. The decision threshold t is the cut-off value of the model's predicted values at which you classify an instance as belonging to that class, i.e., if $R > t$, $\widetilde{Y} = 1$.

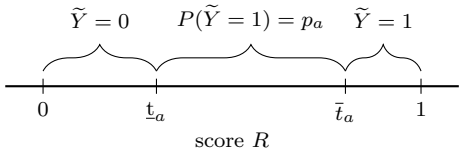
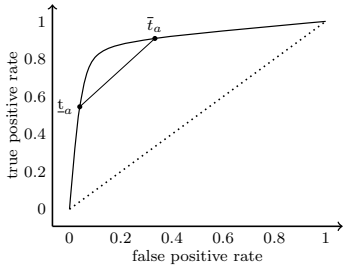
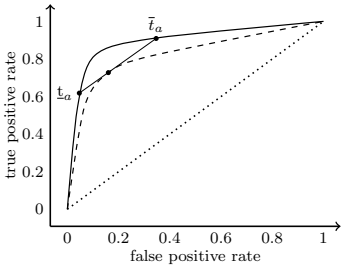


Fig. 2. The randomized decision threshold for a probabilistic classifier. The randomized threshold T_a is equal to $\underline{t}_a$ with probability p_a and equal to $\overline{t}_a$ with probability $1 - p_a$.

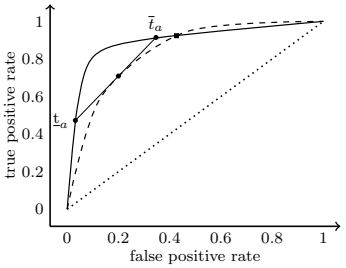
equal to that of the worst-off group (Figure 3(b)). In cases where the ROC curves do intersect, but at a sub-optimal point, which group is affected depends on the specific trade-off that is considered (Figure 3(c)).



(a) Randomization of the decision threshold between values $\underline{t}_a$ and $\overline{t}_a$ allows us to achieve any (fpr, tpr) combination on the line segment between the corresponding points on the ROC curve. The exact coordinate is determined by p_a, where higher values of p_a are closer to $\overline{t}_a$ and vice versa.



(b) The group-specific ROC curves of group A (solid) and group B (dashed) do not intersect, except at trivial end-points. We can achieve equalized odds by randomizing the decision threshold of group A such that it coincides with the ROC curve of group B.



(c) The ROC curves of group A (solid) and group B (dashed) only intersect at a point with a high false positive rate (square). If the cost of false positives is high, this point may not be optimal. Randomization of the thresholds of group A allows for a different trade-off.

Fig. 3. ROC curves of randomized decision thresholds.

5 JUSTIFICATION OF GROUP-SPECIFIC THRESHOLDS

We will now turn to our analysis of the moral justification of group-specific thresholds. We first list several general objections against this strategy, after which we consider the justification under measurement bias and unjust life's bias.

Arguments against group-specific thresholds. A common objection against group-specific thresholds is that it implies that each group is held to a different standard: $\widehat{Y}$ has a different meaning depending on sensitive group membership [15]. Consequently, group-specific thresholds can be interpreted as a form of direct discrimination, in the sense described by Friedler et al. [14]: it introduces a group-level distortion between the *Observed Space* (Y) and the *Decision Space* ($\widehat{Y}$). Moreover, legal scholars have argued that employing a different threshold for each group is likely to be illegal in several legal systems, as it may constitute direct discrimination (in the legal sense) in European Union law and disparate impact in United States labour law [19].

Similar objections can be made from a moral perspective. For example, we could interpret a prohibition against direct discrimination under Kant's categorical imperative by considering it as an *a priori* rule that must universally hold. Taking an egalitarian perspective, group-specific thresholds seem to violate formal equality of opportunity (see Section 3.4.3): we base the decision on sensitive attributes irrelevant to the prediction problem. But what if holding

everyone to the same standards results in an unequal distribution of utility? Then, equality of opportunity is at odds with equality of outcome [16], which requires an equal distribution irrespective of the relevant criteria.

This reveals a tension between *procedural justice*, which considers the means by which we achieve a particular goal, and *distributive justice*, which considers the fairness of the distribution of outcomes that follows from the process. In the remainder of this section, we will explore this tension in the light of measurement bias and unjust life's bias. Following propositions 3.1 and 3.2, we limit this discussion to demographic parity.

5.1 Measurement Bias

Do the objections against group-specific thresholds hold in the presence of measurement bias? As argued in Section 3.4.2, individuals should not be held accountable for invalid measurements. Continuing the egalitarian argument, under measurement bias, it is harder for a disadvantaged group to get a certain score, due to no fault of their own [7]. In fact, under measurement bias, using a single threshold in the *Observed Space* implies the use of a *different* threshold for each group in the *Construct Space*. In this light, even under formal equality of opportunity, there seems to be a moral obligation to hold people to a 'different' standard in the *Observed Space*. We conclude that under measurement bias and WAE-CS, a claimed tension between procedural and distributive justice is false, because it erroneously judges individuals against the *Observed Space* as opposed to the *Construct Space*.

The question remains whether group-specific thresholds are *sufficient* under measurement bias: if we cannot guarantee validity of data, should machine learning be used at all? Clearly, historical policy was biased and accounting for this through group-specific thresholds is likely an improvement to the status quo. However, even an improvement may be a cold comfort if the underlying mechanisms that produced the bias are not resolved. As such, group-specific thresholds only seem justifiable if it is not possible to avoid measurement bias in the first place.

5.2 Unjust Life's Bias

As we have argued in Section 3.4.3, accounting for unjust life's bias is required if we take a substantive egalitarian perspective. Compared to other fair-ml techniques, which often intervene earlier in the machine learning training process, group-specific thresholds are a rather crude form of direct discrimination that disregard within-group differences. This may be justifiable in cases where, due to historical discrimination, sensitive group membership is *directly* related to a particular outcome. Following a Rawlsian interpretation of equality of opportunity, inequalities are accepted if they are based on nature, talent, and preference, but not socio-economic background [30]. A luck egalitarian, on the other hand, likely prefers a more fine-grained attribute to distinguish between 'luck' and 'choice'. Rather than adjusting decision thresholds post-hoc, we should base decisions only on relevant characteristics that are not biased by unjust life's bias in the first place.

6 JUSTIFICATION OF RANDOMIZED THRESHOLDS

'Flipping a coin' is a tried and tested approach for choosing between two competing alternatives or settling disputes. But is randomization justifiable as a fair-ml strategy? In this section, we will explore the justification of group-specific *randomization*, which can be applied to optimize for equalized odds (Figure 3(a)). We will focus on group-specific randomization, but some arguments apply to randomization and group-specific thresholds more broadly.

Arguments for randomization. Drawing from empirical ethics, we can identify general conditions under which randomization is considered ethically acceptable [25]. First, randomization is often recommended when it is important to ensure an unbiased procedure, such as in lotteries or drafts for military service. Here, randomization is not as much a means towards justice, but a means towards *preventing* injustice. Since the post-processing step applies *group-specific* randomization, which is biased by definition, this condition is not relevant for the present paper. A second circumstance in which randomization can be justified is at indecision. If it is impossible for a decision maker to decide which of the alternatives is better, randomization between alternatives seems the most sound procedure. In machine learning, this may be the case if the model's predictions are uncertain. We will explore this condition further in Section 6.1.

Arguments against randomization. A major objection against randomization revolves around the belief that a good decision must be justified by explicit reasons [25]. We can interpret this through Aristotle's concept of justice as consistency: "a decision maker should be able to produce a single, predictable, and correct judgement in each case" [7]. As the gravity of the decision increases, such as in a case of life and death, a coin toss becomes less acceptable [4, 5]. In machine learning, randomization of predictions implies that two very similar or even identical instances can receive a different prediction, violating consistency. Here, we can see a parallel with the notion of *individual fairness*, which poses that similar individuals should be treated similarly [11] and is often contrasted with group fairness. A related objection is that randomization makes it difficult to hold decision-makers accountable for the outcome [25]. These objections bring into question the proportionality of randomization as a means to an end.

In the remainder of this section, we will first explore the justification of randomization in light of indecision. Second, following Section 3.5.3 and Section 3.4.4, we will discuss randomized thresholds in the light of absolute and unjust unequal utility.

6.1 Indecision

Violations of equalized odds typically arise when the data distribution differs substantially across sensitive groups. Disparities are exacerbated if the amount of samples or predictability of the outcome based on the selected features differs per group [3]. This reveals a moral obligation to ensure that the collected data is equally informative for all groups, i.e., that it allows for equally accurate predictions. Importantly, the post-processing step cannot improve the overall predictive performance - it can only allow us to choose a different trade-off between false positives and false negatives, based on the existing predictions. Randomized thresholds even ensure that the performance for all groups is decreased until it is equal to the worst-off group. Therefore, Hardt et al. [17] stress that their post-processing step should only be applied if investing in better features and more data is not an option. This may be the case for outcomes that are inherently difficult to predict. For example, human resources outcomes such as in *resume selection* have an inherently stochastic nature [34].

The uncertainty of the predictions of the model can be interpreted as a state of indecision and could potentially justify randomization. Hardt et al. [17] implicitly incorporate this assumption by applying randomization only if the predicted score is uncertain, i.e., if $\underline{t}_a < R \leq \overline{t}_a$ (Figure 2). The predicted score of the classifier is likely the best estimate we have about whether an individual is a 'border case' or not, although the decision boundary of the classifier can be very different from the 'ground-truth' decision boundary. Moreover, the procedure does not put any restrictions on the width of the interval $[\underline{t}_a, \overline{t}_a]$. At the extreme, we may find that any person in a group subject to randomization

receives a random prediction, disregarding their personal characteristics. This calls into question the proportionality of the approach.

However, even if the amount of randomization seems proportional, an important difference between 'flipping a coin' and randomized thresholds is that it only applies randomization to some groups, but not to others. As such, we argue that indecision by itself is likely insufficient to justify this fair-ml strategy.

6.2 Absolute and Unjust Unequal Utility

In Section 3.5, we have seen that equalized odds disregards absolute utility. This fundamental limitation is mirrored in the post-processing algorithm. First of all, group-specific thresholds may increase the model's misclassification cost for some groups compared to the "optimal" cost that could have been achieved for that group. Randomization further exacerbates this effect, as it artificially decreases performance for all groups until it is equal to that of the worst-off group.

Again, we can come to a moral justification by considering utility. First of all, utility of individuals in a disadvantaged group may depend on the predicted outcomes of another group. For example, the resource constraint in *resume selection* implies that a false negative in one group may provide an opportunity for a member of a historically disadvantaged group (even if it is a false positive). Moreover, in the presence of unjust unequal utility, misclassification of a member of a disadvantaged group may be more harmful compared to misclassification of a member of an advantaged group.

Under such circumstances, we can take a contractualist perspective to justify randomization. Simply put, contractualism is a moral reasoning approach that judges an action based on whether a person could reasonably reject the action [31]. In *resume selection*, one may argue that a member of group A would be willing to forfeit a (true) positive prediction to compensate for historical injustice of individuals in group B. In *disease detection*, we cannot make such an argument: disregarding equality for the sake of equality, decreasing the predictive performance for group A does not benefit individuals in group B. Indeed, in absence of resource constraints, artificially decreasing the performance of one group seems unjustifiable from a contractualist perspective.

7 TOWARDS A MORAL JUSTIFICATION OF FAIRNESS-AWARE MACHINE LEARNING

Fairness-aware machine learning techniques are by no means the only part of the puzzle towards fairer machine learning systems. In many cases, they should arguably not be a part of the solution at all. Several scholars have highlighted how efforts such as more thoughtful measurement [22], expanding our view beyond the algorithmic frame [15, 32], and meaningful participatory design [33] can provide critical paths towards fairer machine learning. That being said, when fair-ml algorithms *are* adopted, we believe it is crucial to carefully consider the moral implications of doing so. With this work, we have taken a first step towards the construction of a comprehensive reasoning framework for assessing the moral justification of fairness metrics and fair-ml algorithms. In the long term, we envision a set of guidelines that supports practitioners in considering and communicating the moral implications of particular strategies in various contexts. We expect this work to inspire new fair-ml algorithms tailored towards more nuanced notions of fairness. In particular, we envision the following lines of future work.

An important research direction is to expand the present work to other common optimization strategies applied in fair-ml as well as the implications of learning without explicit fairness constraints [29]. Some of our arguments can be directly applied to other fair-ml approaches that rely on randomization or group-specific thresholds. Other algorithms, particularly pre-processing and constrained learning algorithms, may be more challenging to understand as a decision-making policy and will require an approach more tightly connected to a specific social context.

In the present paper, we have considered the moral justification of fair-ml through several highly stylized examples, with a heavy focus on technical metrics and algorithms. A limitation of this narrow lens is that it only allows us to consider the direct short-term impact of the machine learning model in isolation. In reality, there may exist downstream or long-term effects on individuals and communities [32]. An important future research direction is to consider the moral justification of fair-ml algorithms within a specific real-world context through case studies. In particular, we envision a comparison of (fairness-aware) classification to existing (non-algorithmic) policies.

Finally, an important question for any ethical issue is who gets the power to decide. In machine learning development, decisions conceptualized as 'technical' are usually left at the discretion of the machine learning practitioner [35]. Perhaps unsurprisingly, we have seen that even a 'technical' choice such as which fair-ml algorithm to use can have ethical implications. As such, we urge researchers to explore ways to incorporate additional stakeholder perspectives in decision-making regarding fair-ml algorithms.

8 CONCLUSIONS

In this work, we have raised the important question of when the use of fairness-aware machine learning algorithms is morally justifiable and proposed one way to answer it. In particular, we considered the moral justification of two strategies used in fair-ml techniques: group-specific decision thresholds and randomized thresholds. Our first contribution is an extension of the framework proposed by Hertweck et al. [20], which allows us to consider the utility of predicted outcomes and introduce three propositions that can justify the use of fairness metric. Our second contribution is a moral analysis of the post-processing algorithm proposed by Hardt et al. [17]. While group-specific thresholds for enforcing demographic parity seem defensible under measurement bias, the thresholds may be too course under a luck egalitarian interpretation of unjust life's bias. Additionally, under resource constraints, the use of randomized thresholds for equalized odds may be defensible from a contractualist perspective. However, we can question the proportionality of the approach in absence of a state of indecision. Our arguments show that the justification of the post-processing algorithm varies across use cases - even if the associated fairness metric is the same. We hope our work inspires other scholars to pursue a more holistic evaluation of fair-ml strategies, beyond their direct optimization objectives.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, Stockholm, Sweden, 60–69. https://proceedings.mlr.press/v80/agarwal18a.html

[2] Richard Arneson. 2018. Four conceptions of equal opportunity. *The Economic Journal* 128, 612 (2018), F152–F173.

[3] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *California Law Review* 104 (2016), 671.

[4] Jonathan Baron and Mark Spranca. 1997. Protected values. *Organizational behavior and human decision processes* 70, 1 (1997), 1–16.

[5] Jane Beattie, Jonathan Baron, John C Hershey, and Mark D Spranca. 1994. Psychological determinants of decision attitude. *Journal of Behavioral Decision Making* 7, 2 (1994), 129–144.

[6] Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 149–159. https://proceedings.mlr.press/v81/binns18a.html

[7] Reuben Binns. 2020. On the Apparent Conflict between Individual and Group Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 514–524. https://doi.org/10.1145/3351095.3372864

[8] Dallas Card and Noah A. Smith. 2020. On Consequentialism and Fairness. *Frontiers in Artificial Intelligence* 3 (May 2020). https://doi.org/10.3389/frai.2020.00034

[9] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[10] Kate Crawford. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker.*

[11] Cynthia Dwork, 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.

[12] Ronald Dworkin. 1981. What is equality? Part 1: Equality of welfare. *Philosophy & public affairs* (1981), 185–246.

[13] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 259–268. https://doi.org/10.1145/2783258.2783311

[14] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im) possibility of fairness. *arXiv:1609.07236* (2016). https://arxiv.org/abs/1609.07236

[15] Ben Green. 2021. Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness. *arXiv:2107.04642* (2021). https://arxiv.org/abs/2107.04642

[16] Jerald Greenberg. 1987. A taxonomy of organizational justice theories. *Academy of Management review* 12, 1 (1987), 9–22.

[17] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016), 3315–3323.

[18] Hoda Heidari, Michele Loi, Krishna P. Gummadi, and Andreas Krause. 2019. A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 181–190. https://doi.org/10.1145/3287560.3287584

[19] Deborah Hellman. 2020. Measuring algorithmic fairness. *Virginia Law Review* 106 (2020), 811.

[20] Corinna Hertweck, Christoph Heitz, and Michele Loi. 2021. On the Moral Justification of Statistical Parity. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 747–757. https://doi.org/10.1145/3442188.3445936

[21] Lily Hu and Yiling Chen. 2020. Fair Classification and Social Welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 535–545. https://doi.org/10.1145/3351095.3372857

[22] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 375–385. https://doi.org/10.1145/3442188.3445901

[23] Faisal Kamiran and Toon Calders. 2011. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (Dec. 2011), 1–33. https://doi.org/10.1007/s10115-011-0463-8

[24] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2010. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*. IEEE, 869–874.

[25] Gideon Keren and Karl H Teigen. 2010. Decisions by coin toss: Inappropriate but fair. *Judgment and Decision Making* 5, 2 (2010), 83.

[26] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv:1609.05807* (2016). https://arxiv.org/abs/1609.05807

[27] Derek Leben. 2020. Normative principles for evaluating fairness in machine learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 86–92.

[28] Derek Leben. 2020. Normative Principles for Evaluating Fairness in Machine Learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) *(AIES '20)*. Association for Computing Machinery, New York, NY, USA, 86–92. https://doi.org/10.1145/3375627.3375808

[29] Lydia T Liu, Max Simchowitz, and Moritz Hardt. 2019. The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning*. PMLR, 4051–4060.

[30] John Rawls. 1971. *A theory of justice*. Harvard University Press.

[31] Thomas Scanlon. 2000. *What we owe to each other*. Belknap Press.

[32] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[33] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2020. Participation is not a design fix for machine learning. *arXiv:2007.02423* (2020). https://arxiv.org/abs/2007.02423

[34] Prasanna Tambe, Peter Cappelli, and Valery Yakubovich. 2019. Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review* 61, 4 (2019), 15–42.

[35] Pak-Hang Wong. 2020. Democratizing algorithmic fairness. *Philosophy & Technology* 33, 2 (2020), 225–244.

[36] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 54)*, Aarti Singh and Jerry Zhu (Eds.). PMLR, 962–970. https://proceedings.mlr.press/v54/zafar17a.html

[37] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 28)*, Sanjoy Dasgupta and David McAllester (Eds.). PMLR, Atlanta, Georgia, USA, 325–333. https://proceedings.mlr.press/v28/zemel13.html

[38] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) *(AIES '18)*. Association for Computing Machinery, New York, NY, USA,

335–340. https://doi.org/10.1145/3278721.3278779