A Review of the Different Notions of Algorithmic Fairness

Bobby (Zelin) Lv[1], Shengwen Yang[2], and Rajan Dalal[3]

[1]UW-Madison, zlv7@wisc.edu
[2]UW-Madison, syang382@wisc.edu
[3]UW-Madison, rdalal@wisc.edu

Abstract

Algorithmic fairness is an emerging field in Computer Science that aims for "fairer" probabilistic classifications across different groups and protected features. Many formal definitions and notions of fairness have been proposed over the years, each with their own trade-offs and applications. In this paper, we review and compare the different definitions of fairness policies, inter-compatibilities of fairness conditions, and plausible use cases. This review will show that there is no single right over-arching notion of fairness. Rather, different notions offer varying compromises in different applications.

1 Introduction

Machine learning models are increasingly being used in assessment frameworks across a wide variety of fields such as law enforcement, medicine, banking, advertisement, and others. These statistical frameworks affect people from various demographics and a desirable, often necessary, aspect of these models is their *fairness*. In recent years, multiple definitions and notions of fairness have been proposed, based on backgrounds ranging from the statistical features of predictions to the inherent worldviews based on mapping spaces.

The most basic form of fairness is possibly *anti-classification*, where the algorithm does not consider any protected features like gender, race or general group membership. However, even this strict form of fairness can cause *disparate impact* - disadvantageous effects - for some groups. Following this are various definitions of metric parity. In this family of fairness policies, the algorithm tries to equate certain metrics across all groups, such as AUC of the ROC curve or the error rate. The incompatibilities of these definitions has been a focus of some recent research [KMR17; Cho17]. Various other definitions are also discussed.

Another way of looking at fairness is recognising how algorithms model the real world. To this end, concepts of *feature spaces* have been put forth recently [FSV16]. The discussion of these world-view type frameworks also leads into the discussion of the inherent impossibilities of implementing multiple fairness criteria.

'Fair' models are used today in criminal justice[Ang+16], medicine[Gar97], city planning[Shr17], and a wide range of other fields. Implementing fairness in these fields is important not just for a bureaucratic check mark but also for the development of affected communities as a whole. The 'how to' of fair modelling is an important debate and indiscriminate applications of mathematical fairness conditions may not be a necessary or sufficient condition for achieving non-technical notions of fairness. It will always be important to analyse the disparate impacts of a model on the various groups being touched upon, and even, say, a simple forgoing of the high predictive value protected features is not enough to achieve 'fairness.'

2 Background

To understand the impacts of fairness it is necessary to looks at how *unfairness* can be defined.

2.1 Treatment and Impact

The Criminology / Law paper [BS16] lists two notions of *unfairness* that are important:

Disparate Treatment: From the paper:

> Disparate treatment comprises two different strains of discrimination: (1) formal disparate treatment of similarly situated people and (2) intent to discriminate. . . . Disparate treatment recognizes liability for both explicit formal classification and intentional discrimination. . . . In the world of data mining, . . . [e]ven if membership in a protected class were specified as an input, the eventual model that emerges could see it as the least significant feature. In that case, there would be no discriminatory effect, but there would be a disparate treatment violation, because considering membership in a protected class as a potential proxy is a legal classificatory harm in itself.

In other words, violation of anti-classification can possibly, legally speaking, be seen as a sign of discrimination.

Disparate Impact: From the paper:

> Disparate impact refers to policies or practices that are facially neutral but have a disproportionately adverse impact on protected classes. Disparate impact is not concerned with the intent or motive for a policy; where it applies, the doctrine first asks whether there is a disparate impact on members of a protected class, then whether there is some business justification for that impact, and finally, whether there were less discriminatory means of achieving the same result.

This metric of unfairness is a useful tool when looking at the limitations of certain fairness policies.

2.2 Confusion Matrix

Some of the primary metrics to equate classification on are derived from a confusion matrix, like one below. It catalogues the prediction-vs-truth values of the predictor model. In the following sections we will describe some definitions with respect to this matrix.

True Value	Positive Prediction	Negative Prediction
Positive	a : True Positive	b : False Negative
Negative	c : False Positive	d : True Negative

2.3 Related Work

There is a large line of work on fairness in Computer Science. In order to remove discrimination and preserve fairness, one approach that has been used is to change the datasets and remove the information that causes 'unfair' results [KC09; HD13]. Another similar approach is to learn intermediate representations of data that preserve enough information and fairness [Zem+13]. Previous work also considers whether issues regarding fairness could arise in the learning process [Jos+16].

In [Héb+17], Hebert-Johnson et al. state that training data might inadvertently or maliciously introduce biases that are not borne out in the data. In [Blu+18], Blum et al. showed that fairness using an extension of classical on-line learning is unachievable for the prevalent notion of "equalized odds" that requires equal false negative rates and equal false positive rates across groups (also see Section 3.8).

A great survey of sources of unfairness in model learning and the disadvantages of using bandits for learning has been discussed in [CR18]. [Zaf+15] introduces a mechanism to design fair classifiers in machine learning by adapting decision boundaries to notions of fairness. [Kus+17] talks about fairness in the context of causality.

3 Definitions

3.1 Anti-classification

[CG18] describe anti-classification as the process of ignoring all protected or discriminatory features. Formally, anti-classification requires that:

$$d(x) = d(x') \;\; \forall \; x, x' \text{ such that } x_u = x'_u \tag{1}$$

Where $d(x)$ is the final prediction (or decision) based on input x, and x_u is the projection of x onto its unprotected features.

3.2 Classification Parity

Classification Parity, as described in [CG18], refers to a broad family of fairness conditions where some measure of classification is equated across groups. This measure is usually derived from the confusion matrix in Section 2.2. [Ber+17] lists a few possible classification parity class conditions:

1. *Overall accuracy equality* : The predictor satisfies this parity when the overall accuracy for each class is equated. From the confusion matrix, this means that the value $(a+d)/(a+b+c+d)$ is equated across all groups.

2. *Statistical parity* : The predictor satisfies this when the marginal distribution of predictions are equated. That is, when $(a+c)/(a+b+c+d)$ and $(b+d)/(a+b+c+d)$ are the same across all groups.

3. *Conditional procedure accuracy equality* or *Error rate balance*: The predictor satisfies this when the CPA (conditional procedure accuracy) is the same across groups. That is, $a/(a+b)$ and $d/(c+d)$ are equated across groups. Note that this the same as saying that $b/(a+b)$ and $c/(c+d)$ be equated across the groups, or that the error rates be equated.

4. *Conditional use accuracy equality* : The predictor satisfies then when, across all groups, the conditional use accuracy, or $a/(a+c)$ and $d/(b+d)$, are equated.

5. *Treatment equality* : Treatment equality looks at the ratio of false negatives and false positives. The predictor satisfies this when b/c is equated across all groups.

In addition to these, [CG18] also consider equating AUC under a ROC curve as a type of classification parity.

Another common condition that falls in this grouping is *demographic parity*. In this parity, we equate the proportion of positive decisions. That is, $(a+c)/(a+b+c+d)$ is equated across all groups. This is somewhat similar to statistical parity.

3.3 Calibration

Under calibration, we ensure that the outcomes of the predictor are independent of all protected attributes when conditioned on risk score. In other words, risk scores should be consistent across all groups, and the same risk score should 'mean the same thing' for all members of all groups. Formally, given a risk score $s(x)$,

$$Pr(Y = 1|s(X), X_p) = Pr(Y = 1|s(X)) \tag{2}$$

In other words, an algorithm is *well-calibrated* if the algorithm predicts a group as having a positive instance with probability z, then approximately a z fraction of this group should indeed be positive instances, applicable to all groups.

[Cho17] considers a framework that is well-calibrated as being *free from predictive bias*.

3.4 Total Fairness

[Ber+17] describe a predictor as satisfying *total fairness* when it achieves, simultaneously, all of overall accuracy equality, statistical parity, conditional procedure accuracy equality, conditional use accuracy equality, and treatment equality.

3.5 Thresholding

Most of the notions of fairness above can be written in terms of a threshold function. [Cho17] describes some of these threshold variants. For example, *predictive parity* is defined as being demographic parity over a threshold. Given $s(X)$ as a score function, s_{th} as a threshold, and g_0, g_1 as two groups,

$$Pr(Y = 1|s(X) > s_{th}, X \in g_0) = Pr(Y = 1|s(X) > s_{th}, X \in g_1) \ \forall \ g_0, g_1 \tag{3}$$

Under this system, *statistical parity* is redefined in terms of the threshold as well,

$$Pr(s(X) > s_{th}|X \in g_0) = Pr(s(X) > s_{th}|X \in g_1) \ \forall \ g_0, g_1 \tag{4}$$

3.6 Balance For Classes

Balance for classes is balancing for average scores for both positive and negative instances. From [KMR17], *balance for the positive class*, for instance, is achieved when the average score assigned to people from one group who belong to the positive class is the same as the average score assigned to people from every other group who also belong to the positive class. A similar requirement applies to *balance for the negative class*.

Note that this is different from statistical parity. In statistical parity, we look at the average estimate across all members of the groups, and not at the average scores for particular class members.

3.7 Individual and Group Fairness

[Dwo+12] considers the notion of Group and Individual Fairness.

Group Fairness: Also called statistical parity - a description of fairness that looks at global rates across groups.

Individual Fairness: This is the notion that any two individuals who are similar with respect to a particular task should be classified similarly. 'Similarity' in this notion is described using a distance metric in a task space. Two such distance metrics are:

1. *Lipschitz Distance*: In this, we create first a map M of a set of individuals, V to a set of outcomes in the algorithm, A. Specifically, $X \in V$ has outcome $M(X)$. Then, using the above definition of individual fairness, this mapping satisfies the $(D, d) - Lipschitz$ property, where D and d are distance metrics, if, $\forall X_0, X_1 \in V$:

$$D(M(X_0), M(X_1)) \leq d(X_0, X_1) \tag{5}$$

Creating a loss function L over this mapping converts the search function into an optimisation problem : find an appropriate mapping M that minimises L.

Two such D-metrics are D_{tv} and D_∞.

D_{tv}: This *statistical distance* or *total variation norm* is denoted by (S and T being two different distributions):

$$D_{tv}(S, T) = \frac{1}{2} \sum_{a \in A} |S(a) - T(a)| \tag{6}$$

The idea behind D_{tv} is that similar individuals have d close to 0 while dissimilar individuals have d closer to 1.

D_∞: Also called the *relative* l_∞ *metric*, the assumption here is that similar individuals have $d \ll 1$ while dissimilar individuals have $d \gg 1$.

$$D_\infty(S, T) = \sup_{a \in A} log\left(max \left(\frac{S(a)}{T(a)}, \frac{T(a)}{S(a)} \right) \right) \tag{7}$$

2. *Earthmover Distance*: Defining 'bias' between two distributions S and T from a particular Lipschitz-metric as:

$$bias_{D,d}(S, T) = max(\mu_S(0) - \mu_T(0)) \tag{8}$$

and redefining *statistical parity* in terms of bias ϵ:

$$D_{tv}(\mu_S, \mu_T) \leq \epsilon \tag{9}$$

Then the *Earthmover distance* d_{EM} relates D_{tv} to a bias as:

$d_{EM}(S, T) = min \sum_{X_0, X_1 \in V} h(X_0, X_1) d(X_0, X_1)$

where $\sum_{X_1 \in V} h(X_0, X_1) = \sum_{X_1 \in V} h(X_1, X_0) + S(X_0) - T(X_0)$

and $h(X_0, X_1) \geq 0$

Then for a distance d,

$$bias_{D_{tv},d}(S, T) \leq d_{EM}(S, T) \tag{10}$$

If $d(X_0, X_1) \leq 1 \; \forall \; X_0, X_1$, we have:

$$bias_{D_{tv},d}(S,T) = d_{EM}(S,T) \tag{11}$$

Thus, Earthmover distance constraints the Lipschitz distance to within the bounds of statistical parity.

3.8 Odds and Opportunity

In [HPS16] we encounter the notions of:

Equalized odds: The predictor satisfies this when, with respect to protected features A and the true value Y', if the prediction Y and A are independent conditional on Y'. That is, when

$$Pr(Y = 1|A = 0, Y' = y) = Pr(Y = 1|A = 1, Y' = y) \forall y \in 0, 1 \tag{12}$$

In other words, a predictor has equalized odds when we equate both the True Positive and False Positive Rates across all groups. Note than achieving equalized odds also achieves demographic parity.

Equalized opportunity: The (binary) predictor satisfies this when, with respect to protected features A and the true value Y', if the prediction Y and A are independent conditional on Y', when $Y' = 1$. That is, when

$$Pr(Y = 1|A = 0, Y' = 1) = Pr(Y = 1|A = 1, Y' = 1) \forall y \in 0, 1 \tag{13}$$

This assumes that $Y' = 1$ is an 'advantageous' outcome. In other words, a predictor has equalized opportunity when we equate the True Positive Rate across all groups.

3.9 World Views

A fundamentally different notion of fairness can encode the way models represent the world. [FSV16] defines world-views using notions of *feature spaces*. We have:

1. *Construct Space (CS)*:
   The construct space is a metric space $CS = (P, d_P)$ consisting of individuals and a distance between them. It is assumed that the distance d_P correctly captures closeness with respect to the task. For example, for college admissions, personal qualities, such as "self-control", "growth mind-set", or "grit" would belong to the abstract construct space.

2. *Observed Space (OS)*:
   The observed space (with respect to a task T) is a metric space $OS = (\hat{P}, \hat{d})$. [FSV16] assumes an observation process $g : P \rightarrow \hat{P}$ that generates an entity $\hat{p} = g(p)$ from a person p $\in$ the CS. For example, in the admissions context above, "grit" other such qualities are not directly observable. The "amount of grit" is a feature in the CS: it is unobservable but appears to have some unknown influence on the desired predicted outcome. The inferred metric from the CS, through imperfect proxy features, such as a "survey-based grit score," would lie in the OS.

3. *Decision space (DS)*:
   A decision space is a metric space $DS = (O, d_O)$, where O is a space of outcomes and d_O is a metric defined on O. A task T can be viewed as the process of finding a map from P or $\hat{P}$ to O. In the example context above, the decision space could consist of the information that makes up the final admissions decision. The decision space might be the binary yes/no decisions, or might be the predicted "potential" of an applicant or their predicted "performance" in college, similar to a risk score.

3.9.1 Fairness w.r.t Feature Spaces

A mapping $f : CS \to DS$ is said to be 'fair' if objects that are close in CS are also close in DS. For two thresholds ε, ε', f is defined as (ε, ε')-fair if for any $x, y \in P$,

$$d_P(x, y) \leq \varepsilon \Rightarrow d_O(f(x), f(y)) \leq \varepsilon' \tag{14}$$

3.9.2 Distances within Spaces

Within a metric space, there are various ways of defining the distance used above. Unlike in Section 3.7 these are measured not in the context of a bias, but with what is called a *coupling measure*:

Coupling Measure: Let X, Y be sets with associated probability measures μ_x, μ_y. A probability measure v over $X \times Y$ is a coupling measure if $v(X, .)$ (the projection of v on X) equals μ_x, and similarly for $v(., Y)$ and μ_y. The space of all such coupling measures is denoted by $\mathcal{U}(X, Y)$.

Using this notion we can now calculate the distance between points within a metric space. One such distance is:

Wasserstein Distance (WD): If (X, d) is a metric space and Y, Y' are two subsets of X, μ is a probability measure defined on X, which in turn induces probability measures μ_Y and $\mu_{y'}$ on Y and Y' respectively. The Wasserstein Distance between Y, Y' is given by:

$$W_d(Y, Y') = \min_{v \in \mathcal{U}(Y, Y')} \int d(y, y') \cdot v(y, y') \tag{15}$$

To calculate distance between two metric spaces, we have:

Gromov-Wasserstein Distance (GWD): If (X, d_X) and (Y, d_Y) are two metric spaces with associated probability measures μ_X and μ_Y, the Gromovv-Wasserstein Distance between X and Y is given by:

$$GW(X, Y) = \frac{1}{2} \inf_{v \in \mathcal{U}(X,Y)} \iint |d_X(x, x') - d_Y(y, y')| d_{\mu_X} \times d_{\mu_X} d_{\mu_Y} \times d_{\mu_Y} \tag{16}$$

3.9.3 Different Worldviews

Worldviews, in [FSV16], are used to describe how we interpret the existence and importance of various construct spaces. Two important worldviews are:

What you see is what you get (WYSIWYG): Under this worldview, we assume that the CS and the OS are essentially the same. In non-technical terms, this means that, for example, if a person has a certain score on an OS proxy, like say, an IQ test, then that proxy accurately represents their metric (intelligence, here) in the CS. Formally, under WYSIWYG,

There exists a mapping $f : CS \to OS$ such that the distortion ρ_f is at most ε for some small $\varepsilon > 0$. Or equivalently, the distortion ρ between CS and OS is at most ε.

Where we describe distortion as:

If (X, d_X) and (Y, d_Y) are two metric spaces and $f : X \to Y$ is a map from X to Y, then the distortion ρ_f of f is defined as the smallest value such that for all $p, q \in X$:

$$|d_X(p, q) - d_Y(f(p), f(q))| \leq \rho_f \tag{17}$$

Structural Bias with WAE: In the *structural bias* worldview, we assume that that the OS does not accurately represent the CS. Real life factors can influence the transformations from CS to OS. For example, in college admissions, this could mean attributing some of the differences in standardised test scores to socio-economic backgrounds of the applicants. Formally,

The metric spaces CS = (X, d_X) and OS = (Y, d_Y) admit t-structural bias if the group skew $\sigma(X, Y) > t$.

Where group skew is calculated using:

The between-groups distance between $(X, d_X), (Y, d_Y)$ with measures μ_X, μ_Y is :

$$\rho_b = \frac{GW(X, Y)}{\binom{k}{2}} \tag{18}$$

The within-groups distance between X_i and Y_i, two sets in the spaces X, Y corresponding to the i^{th} group (let $\rho_i = GW(X_i, Y_i)$) is:

$$\rho_w = \frac{1}{k} \sum_{i=1}^{k} \rho_i \tag{19}$$

With this, group skew is:

If $(X, d_X), (Y, d_Y)$ are metric spaces with group partitioning X', Y' and measures μ_X, μ_Y. The group skew between X' and Y' is the quantity:

$$\sigma(X', Y') = \frac{\rho_b(X', Y')}{\rho_w(X', Y')} \tag{20}$$

One underlying assumption under structural bias is the axiom of WAE (We're all equal):

We're all equal (WAE): Under this axiom, we assume that inherently, all applicants / people / data points are equal in ability. In other words, everyone has nearly similar Construct Spaces, and the differences in the OS can be attributed to structural bias. Formally,

Let $CS = (X, d_X)$ with measure be partitioned into groups $X_1, ..., X_k$. There exists some $\varepsilon > 0$ such that for all $i, j, W_{d_X}(X_i, X_j) < \varepsilon$.

3.9.4 Fairness mechanisms

The paper describes three factors in recgonising fairness:

Richness: A mechanism $f : OS \to DS$ is rich if for each $d \in DS, f^{-1}(d) \neq \emptyset$

Individual fairness mechanism (IFM): Fix a tolerance ε. A mechanism IFM_ε is a rich mapping $f : OS \to DS$ such that $\rho_f \leq \varepsilon$.

Group fairness mechanism (GFM): Let X be partitioned into groups X_1, X_2,... as before, A rich mapping $f : OS \to DS$ is said to be a valid group fairness mechanism if all groups are treated equally. Specifically, fix ε. Then f is said to be a GFM_ε if for any $i, j, W_{d_O}(X_i, X_j) \leq \varepsilon$.

4 Trade-offs and limitations

Having looked at the various notions of fairness, we now look at the limitations and trade-offs of the definitions. Despite being desirable, simultaneous satisfaction of multiple fairness guarantees can not only be difficult, but also impossible. One important such example of a trade-off is the incompatibility of calibration and balance for the classes.

4.1 Calibration and Balance for the Classes

[KMR17] is a famous, recent paper that talks about the inherent incompatibilities of having a fairness solution that is both well-calibrated and achieves balance for both positive and negative classes.

Although these three definitions are crucial for fairness in the future predictions or classification tasks of algorithms or machine learning models - that each member of different groups should have the same results regardless of group membership, these three conditions are in general incompatible at the same time in most cases and can only be satisfied in some really constrained cases, even in the approximate versions of these conditions. The constraint cases are **perfect prediction** and **equal base rates**.

Theorem 4.1. *(**Exact version of impossibility of Fairness Condition**). Consider an instance of the problem in which there is a risk assignment satisfying fairness conditions well-calibrated, Balance for the positive class, and Balance for the negative class. Then the instance must either allow for perfect prediction or have equal base rates.*

Even when we relax the conditions to be approximate fair, we still have the following impossibility result.

Theorem 4.2. *(**Approximate version of impossibility of Fairness Condition**). There is a continuous function f, with $f(x)$ going to 0 as x goes to 0, so that the following holds. For all $\epsilon > 0$, and any instance of the problem with a risk assignment satisfying the $\epsilon -$ approximate versions of fairness conditions well-calibrated, Balance for the positive class, and Balance for the negative class, the instance must satisfy either the $f(\epsilon)-$approximate version of perfect prediction or the $f(\epsilon)-$-approximate version of equal base rates.*

4.1.1 Problem Formulation

In the proof of the impossibility result of fairness condition, we will assume that we only have two group G_1, G_2, since this results can be easily extent to the cases where we have more than 2 groups. And we will solve this problem through a model called fairness assignment, whose mathematical definition will be given below.

4.1.2 Instance

Instance includes all the information required for the two groups. An instance is given as a tuple, (G_1, G_2, σ, p). G_t is group of t and each member of each group has a feature vector from σ and a label. And each entry in $n_t \in \mathbf{R}^{|\sigma|}$ denotes the number of member in group t has feature σ. Here in this problem, we just focus on the case where the label is binary. We use $|\sigma|$ to denote the number of feature vectors. Each entry in the vector $p \in \mathbf{R}^{|\sigma|}$ in position σ denotes the probability that a member with σ belong to the positive class. And we have $P \in \mathbf{R}^{|\sigma|*|\sigma|}$ as the diagonal matrix version of p. We denote the number of members of G_t as $N_t = |G_t|$ and μ_t as the number of members of G_t belong to positive class.

4.1.3 Fairness Assignment Model

Given a model, we want to have assignment of these two groups and similar to the instance, we define the assignment as a tuple (X, v, B). B is the number of bins, here we can view it as the variable of dimension of assignment matrix X and score vector v. Since we want a non-trivial assignment, $B \geq 2$. The assignment matrix $X \in \mathbf{R}^{|\sigma|*B}$ has its entry $x_{\sigma b}$ specifies the fraction of people with feature vector σ who get mapped to bin b and score vector $v \in \mathbf{R}^B$ has its entry v_b specifies the score given to member in b. Similarly, we use $V \in \mathbf{R}^{B*B}$ to represent the diagonal matrix of v.

4.1.4 Fairness Conditions

Here we give the mathematical definitions of fairness conditions of *well-calibrated*, *Balance for the positive class*, and *Balance for the negative class*.

- Well-calibrated: ${n_t}^T PX = {n_t}^T XV$
- Balance for the negative class: $\frac{{n_1}^T XVv}{\mu_1} = \frac{{n_2}^T XVv}{\mu_2}$
- Balance for the positive class: $\frac{\mu_1 - {n_1}^T XVv}{N_1 - \mu_1} = \frac{\mu_2 - {n_2}^T XVv}{N_2 - \mu_2}$

The mathematical definitions of the cases we have mentioned above:

- Equal base rate: Two groups have the same fraction of members in the positive class. $\frac{\mu_1}{N_1} = \frac{\mu_2}{N_2}$.
- Perfect prediction: for each feature vector σ, either $p_\sigma = 0$ or $p_\sigma = 1$.

4.1.5 Main Impossibility Theorem

Theorem 4.3. *Consider an instance of the problem in which there is a risk assignment satisfying fairness conditions well-calibrated, Balance for the positive class, and Balance for the negative class. Then the instance must either allow for perfect prediction or have equal base rates.*

With the model given above, we can prove the theorem.

Proof. To have an assignment that satisfies the conditions above is equivalent to have a solution for the following system of equations:

$$\begin{cases} {n_t}^T PX = {n_t}^T XV \\ \frac{{n_1}^T XVv}{\mu_1} = \frac{{n_2}^T XVv}{\mu_2} \\ \frac{\mu_1 - {n_1}^T XVv}{N_1 - \mu_1} = \frac{\mu_2 - {n_2}^T XVv}{N_2 - \mu_2} \end{cases}$$

From these equations, we can obtain that

$$\frac{1}{N_1 - \mu_1}(\mu_1 - {n_1}^T XVv) = \frac{1}{N_2 - \mu_2}(\mu_2 - {n_2}^T XVv)$$

Then we use γ_t to represent $\frac{1}{\mu_t}{n_t}^T XVv$ and from the condition above, we must have $\gamma_1 = \gamma_2 = \gamma$. So we have

$$\frac{\mu_1/N_1}{1 - \mu_1/N_1}(1 - \gamma) = \frac{\mu_2/N_2}{1 - \mu_2/N_2}(1 - \gamma)$$

This equality implies that there are two cases that satisfies this equation:

- $1 - \gamma = 0$, which is $\gamma = 0$, the perfection preditions;
- $\frac{\mu_1/N_1}{1-\mu_1/N_1}(1 - \gamma) = \frac{\mu_2/N_2}{1-\mu_2/N_2}(1 - \gamma)$, where $\mu_1/N_1 = \mu_2/N_2$, the equal base rate.

This completes the proof of Theorem 4.1. □

Theorem 4.4. *There is a continuous function f, with $f(x)$ going to 0 as x goes to 0, so that the following holds. For all $\epsilon > 0$, and any instance of the problem with a risk assignment satisfying the $\epsilon - approximate$ versions of fairness conditions well-calibrated, Balance for the positive class, and Balance for the negative class, the instance must satisfy either the $f(\epsilon)-$approximate version of perfect prediction or the $f(\epsilon)-$-approximate version of equal base rates.*

Proof. First, we give the mathematical definitions of the approximate fairness conditions:

$$(1-\varepsilon)[n_t XV]_b \leq [n_t PX]_b \leq (1-\varepsilon)[n_t XV]_b$$
$$(1-\varepsilon)\mathsf{p}\frac{1}{N_2-\mu_2}n_t(I-P)Xv \leq \mathsf{p}\frac{1}{N_1-\mu_1}n_t(I-P)Xv \leq (1+\varepsilon)\mathsf{p}\frac{1}{N_2-\mu_2}n_t(I-P)Xv$$
$$(1-\varepsilon)\mathsf{p}\frac{1}{\mu_2}n_t PXv \leq \mathsf{p}\frac{1}{\mu_1}n_t PXv \leq (1+\varepsilon)\mathsf{p}\frac{1}{\mu_2}n_t PXv$$

For the last two conditions, we also require that these hold when μ_1 and μ_2 are interchanged.

As well as the three conditions above, we also define approximate version of equal base rate and perfect conditions respectively in terms of $f(\varepsilon)$, which is a function that goes to 0 as ε goes to 0.

- *Approximate perfect prediction.* $\gamma_1 \geq 1 - f(\varepsilon)$ and $\gamma_2 \geq 1 - f(\varepsilon)$
- *Approximately equal base rates.* $|\mu_1/N_1 - \mu_2/N_2| \leq f(\varepsilon)$

In order to save space, we omit the full proof here. □

4.2 World View Trade-offs

Looking at the worldviews, [FSV16] shows how, under the existence of structural bias in the moddel, no mechanism can guarantee fairness. Fairness (as they describe it) can only be achieved under the WYSIWYG worldview. Even then, only individual fairness can be fair. A group fairness mechanism will, on the other hadn, be unfair. If structural bias is assumed, applying an individual fairness mechanism will cause discrimination in the DS whether we assume WAE or not.

To quote the paper,

> The assumption of a worldview is vital when choosing a mechanism of fairness. Under a WYSIWYG worldview,only individual fairness mechanisms achieve fairness (and group fairness mechanisms are unfair). Under a structural bias worldview, only group fairness mechanisms achieve non-discrimination (and individual fairness mechanisms are discriminatory).

4.3 Error Rate Balance and Predictive Parity

[Cho17] looks at the trade-offs of the fairness criterion of error rate balance and predictive parity.

Consider again the confusion matrix described in Section 2.2. Consider two groups, g_0, g_1 with prevalence rates p_0 and p_1, the rates of the true outcome being 'positive.' If FPR is the False Positive Rate, FNR is the False Negative Rate, and PPV is the Positive Predictive Value (the final outcome after thresh-holding), we have,

$$FPR_i = \frac{p_i}{1-p_i} \cdot \frac{1-PPV_i}{PPV_i} \cdot (1-FNR_i) \;\forall\; i \tag{21}$$

Equating PPVs leads to achieving predictive parity, while equating FPR and FNR leads to achieving error rate balance. Equating on both of them is conditional *only* on the assumption that $p_0 = p_1$. That is,

Proposition 4.5. *Both Error Rate Balance and Predictive Parity are simultaneously achievable if and only if the underlying prevalence rates are equal among groups.*

Thus, trying to achieve predictive parity with different underlying base rates will lead to disparate impact in the final predictor due to different FNR and FPR rates across groups.

4.4 The Representation Paradox

The trade-offs mentioned above are almost all caused by the underlying differences in base rates among groups. This highlights the requirements of a representative training sample that accounts for examples from different groups. However, the mere existence of these protected attributes in the dataset violates anti-classification and is a type of disparate treatment. On the other hand, trying to implement multiple incompatible fairness criteria on differing base rates will always lead to disparate impacts. The only time when this does not happen, paradoxically, is when the base rates are equal, meaning that there is no requirement of having different representative samples. In this case, then, the protected attributes play no role as they do not distinguish the data points, as all of them lie in the same group anyway. This leads us to the following proposition:

Proposition 4.6 (The Representation Paradox)**.** *Compatible fairness criteria based on equating conditional predictor outcomes are possible only when anti-classification has no effect on the accuracy of the learning model.*

In other words,

Corollary 4.7. *It is only when representative samples are* not *required, that is, it is only in the trivial case that multiple fairness criteria can be established.*

This establishes why there is indeed an inherent trade-off in some application of fairness policies.

4.5 Limitations of Definitions

On top of the trade-offs in these notions of fairness, the definitions themselves are often inadequate in establishing a 'true' sense of fairness.

4.5.1 Limitations of anti-classification

The main drawback, looking from a critical model review standpoint, in the application of anti-classification, is the loss in accuracy of the model. Removing protected features completely from the model can remove features with important predictive values. From an ethical standpoint, [CG18] gives the example of COMPAS, a criminology software used to predict recidivism rates among prisoners. By blindly applying anti-classification, each group is essentially subjected to the average treatment for all groups. While this does not violate disparate treatment or even disparate impact, an argument can be made that it unfairly treats all members below the average recidivism

rates.

Specifically looking at gender, women tend to have statistically significantly lower rates of recidivism than men. In other words, for the same COMPAS scores, women are actually less prone to recidivism than men. To balance this, some jurisdictions, like the state of Wisconsin, now require frameworks that take gender into account. Thus, the more legal (and possibly, more ethical) option is to remove anti-classification and therefore eliminate the unfairness of those lying below the average of negative rates.

This case also highlights how enforcing anti-classification on groups with different base rates comes at the cost of calibration. The same scores for recidivism actually mean different metrics for different groups.

4.5.2 Limitations of parities

The inherent limitation of parities, as defined above, is their mutual incompatibility. Another way of looking at this is through the lens of *infra-marginality*, as [CG18] does. Popular metrics are often based on populations away from the margins (hence *infra*-marginal). However, the most impactful applications of economics and government policies often affects the people at the margins. Thus, standard metrics for parities and error rates, based on averages, can be very poor proxies for certain notions of well-being. This is caused, again, by the differing base rates of those on the margins and those away form the margins. The different risk rates mean that the same threshold has different meanings for the different groups, while different thresholds violate other notions of fairness.

Another important limitation of parities is they often consider, as a utility, the impact on the people being effected. and not as their society as a whole. By lowering the threshold for certain tasks, the burden of, say, a False Positive can be negligible to the person, but more impactful for society. For example, by maximising the utility when equalising FPRs, and setting different thresholds for different groups, people with a higher risk score in one group may possibly get a positive outcome than a person with same risk score from another group. In the case of releasing prisoners for parole, this is an undesirable outcome to the community of the released prisoner.

Moreover, even by applying parities, it is possible to subvert them and violate other general, non-technical notions of fairness.[Dwo+12] looks at some ways to 'game' statistical parity:

1. If the utility maximisation schemes differ for different groups, applying statistical parity can lead to undesirable outcomes. For example, if, in people of culture S, the most intelligent students choose subject $\hat{S}$, and in culture T, the subject of choice is $\hat{T} \neq \hat{S}$, then aiming for utility maximisation on the culture of S can lead to choosing the wrong subset in T.

2. The training data or input data can be intentionally modified to only have the lowest feature values for people from a certain group. If, for example, a company chooses to interview many 'low-quality' candidates from a certain group, they can truthfully ignore people from that group while maintaining statistical parity.

4.5.3 Limitations of calibration

Similar to 'gaming' parities, a maligned actor can truthfully satisfy calibration while still violating general notions of fairness. [CG18] describes the example of the process of redlining. In this, the fact that we do not consider the independence of features is exploited. For example, zip codes can be used as a proxy for group membership, and calibrating scores for group membership could be maintained while simultaneously factoring in the zip code feature with arbitrary weights. By changing the distributions of certain groups to be centered around a desired mean, arbitrary thresholds for maintaining the same score can be implemented.

5 Conclusion

In this paper we reviewed multiple definitions and notions of fairness. Some definitions, like *anti-classification* forgo all protected features. Other force a parity between some statistical metric. *Calibration* requires that risk scores for different groups be equally important. Yet other definitions of fairness create feature spaces and compute distance metrics within these spaces.

However, because of the nature of these definitions, there exist incompatibilities between them. For example, except in the most trivial case, it is impossible to have a well-calibrated and class balanced fairness solution. Much of this arises from the *representation paradox*, where there exist cases where the only time when multiple fairness criteria can be implemented is when those very criteria are not needed.

Ultimately, mathematical notions of fairness are just approximations of real-world non-technical notions of fairness. When implementing their own learning models, programmers must take into take what worldview they assume, what mechanism can best fit their model, and what disparate impacts, if any, are caused by their combination of criteria choices. Even after choosing a set of compatible fairness criteria, programmers and policy makers must understand the limitations of these policies, and must be acquainted with the ways in which these conditions can be subverted.

References

[Ang+16] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks". In: (2016).

[Ber+17] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. "Fairness in Criminal Justice Risk Assessments: The State of the Art". In: *Sociological Methods & Research* (Mar. 2017). DOI: 10.1177/0049124118782533.

[Blu+18] Avrim Blum, Suriya Gunasekar, Thodoris Lykouris, and Nathan Srebro. *On preserving non-discrimination when combining expert advice*. 2018. arXiv: 1810.11829 [cs.LG].

[BS16] Solon Barocas and Andrew D. Selbst. "Big Data's Disparate Impact". In: *Advances in Neural Information Processing Systems* 104 (3 2016). DOI: 10.15779/Z38BG31.

[CG18] Sam Corbett-Davies and Sharad Goel. "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning". In: (2018). arXiv: 1808.00023 [cs.CY].

[Cho17] Alexandra Chouldechova. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". In: (2017). arXiv: 1703.00056 [stat.AP].

[CR18] Alexandra Chouldechova and Aaron Roth. *The Frontiers of Fairness in Machine Learning*. 2018. arXiv: 1810.08810 [cs.LG].

[Dwo+12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. "Fairness through awareness". In: *Innovations in Theoretical Computer Science* (2012), pp. 214–226.

[FSV16] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. "On the (im)possibility of fairness". In: (2016). arXiv: 1609.07236 [cs.CY].

[Gar97] Howard N. Garb. "Race bias, social class bias, and gender bias in clinical judgment". In: *Clinical Psychology: Science and Practice* 4 (2 1997), pp. 99–120.

[HD13] Sara Hajian and Josep Domingo-Ferrer. "A methodology for direct and indirect discrimination prevention in data mining." In: *IEEE transactions on knowledge and data engineering* 25 (7 2013), pp. 1445–1459.

[Héb+17] Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. *Calibration for the (Computationally-Identifiable) Masses*. 2017. arXiv: 1711.08513 [cs.LG].

[HPS16] Moritz Hardt, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning". In: *Advances in Neural Information Processing Systems* (2016), pp. 3315–3323.

[Jos+16] Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. "Fairness in learning: Classic and contextual bandits". In: *Advances in Neural Information Processing Systems* (2016).

[KC09] Faisal Kamiran and Toon Calders. "Classifying without discriminating". In: *2nd International Conference on Computer, Control and Communication, IC4* (2009).

[KMR17] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores". In: *8th Innovations in Theoretical ComputerScience Conference, ITCS* (2017).

[Kus+17] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. "Counterfactual Fairness". In: (2017). Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, pp. 4066–4076.

[Shr17] R. Shroff. "Predictive analytics for city agencies: Lessons from children's services". In: *Big Data* 5 (3 2017), pp. 189–196.

[Zaf+15] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. *Fairness Constraints: Mechanisms for Fair Classification*. 2015. arXiv: 1507.05259 [stat.ML].

[Zem+13] Richard S Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. "Learning Fair Representations." In: *Proceedings of the 30th International Conference on Machine Learning* 28 (2013), pp. 325–333.