Avoiding Disparity Amplification under Different Worldviews

Samuel Yeom
Carnegie Mellon University
syeom@cs.cmu.edu

Michael Carl Tschantz
International Computer Science Institute
mct@icsi.berkeley.edu

ABSTRACT

We mathematically compare four competing definitions of group-level nondiscrimination: *demographic parity*, *equalized odds*, *predictive parity*, and *calibration*. Using the theoretical framework of Friedler et al., we study the properties of each definition under various *worldviews*, which are assumptions about how, if at all, the observed data is biased. We argue that different worldviews call for different definitions of fairness, and we specify the worldviews that, when combined with the desire to avoid a criterion for discrimination that we call *disparity amplification*, motivate demographic parity and equalized odds. We also argue that predictive parity and calibration are insufficient for avoiding disparity amplification because predictive parity allows an arbitrarily large inter-group disparity and calibration is not robust to post-processing. Finally, we define a worldview that is more realistic than the previously considered ones, and we introduce a new notion of fairness that corresponds to this worldview.

CCS CONCEPTS

• **Social and professional topics** → **Socio-technical systems**;
• **Mathematics of computing** → *Probability and statistics.*

KEYWORDS

fairness, worldview, disparity amplification, demographic parity, equalized odds, predictive parity, calibration

ACM Reference Format:
Samuel Yeom and Michael Carl Tschantz. 2021. Avoiding Disparity Amplification under Different Worldviews. In *Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3442188.3445892

1 INTRODUCTION

Researchers in the field of fair machine learning have proposed numerous tests for fairness, which focus on some quantitative aspect of a model that can be operationalized and checked using empirical, statistical, or program analytic methods. These tests abstract away more subtle issues that are difficult to operationalize or too contentious to decide algorithmically, such as which groups or attributes should be protected and which cases should be treated as exceptions to general rules. Our work sheds light on some of the possible assumptions behind and motivations for four common empirical tests that check for discrimination against groups.

The simplest of these tests, *demographic parity*, checks whether the model gives the favorable outcome to two given groups of people at equal rates. This test is an abstraction of the legal notion of *disparate impact*, or *indirect discrimination*, which in certain circumstances requires that some approximation of demographic parity hold. Like disparate impact, demographic parity does not depend upon the intentions of the modeler, and it can flag a model that does not directly use the protected attribute if it instead uses another attribute that is correlated with the protected one. However, demographic parity abstracts away disparate impact's exceptions for cases where there is sufficient justification for a disparity in outcomes, such as a *business necessity* [e.g., 2, 18]. By completely abstracting away such exceptions, demographic parity may lead to models so inaccurate as to become useless, such as when predicting physical strength while requiring demographic parity on gender.

This impossibility of accuracy motivates moving away from demographic parity to tests that take the ground truth into account, allowing a degree of accuracy. One such test, called *equalized odds* by Hardt et al. [19], requires equal false positive and false negative rates for each protected group. Two other commonly used tests are *predictive parity*, which requires equal predictive values for each protected group, and *calibration*, which further imposes the constraint that the model must output the correct probability [e.g., 6]. Like demographic parity, all of these tests can be seen as abstractions of disparate impact in that they too examine disparities in outcomes, not how or why they were reached. In contexts where accuracy can be considered a business necessity, these tests arguably provide a more refined abstraction of disparate impact than demographic parity does.

However, disagreement exists over which of these tests is the most appropriate, with some favoring calibration [10] and some favoring equalized odds [1, 19]. It has been argued that adopting the calibration or equalized odds test corresponds to adopting the perspective of either the person using the classification or the person being classified, respectively [1, 28]. We provide a different lens on this disagreement and study the conditions under which each test allows the amplification of pre-existing disparities.

In some cases, the "ground truth" may be tainted by past discrimination, and consulting it will help perpetuate the discrimination. In this work, we handle this issue by adopting the framework of Friedler et al. [14], who make a distinction between the observed ground truth and the *construct*, which is the attribute that is truly relevant for prediction. For example, in the context of bail decisions, the construct could be whether a defendant commits a crime while out on bail, and the observed ground truth could be whether the defendant is rearrested for a crime. Because the construct is usually unobservable, Friedler et al. introduce and analyze two assumptions, or *worldviews*, about the construct: Under the We're All Equal (WAE) worldview, there is no association between the construct and the protected attribute, and under the What You See Is What

FAccT '21, March 3–10, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8309-7/21/03.
https://doi.org/10.1145/3442188.3445892

You Get (WYSIWYG) worldview, the observations accurately reflect the construct.

By using the construct, we specify a natural criterion for discrimination. This criterion, *disparity amplification*, deals with the disparity in positive classification rates, which is a widely accepted measure of discriminatory effect in both law [12] and computer science [4, 5, 13, 21, 35, 36]. It stipulates that a disparity in the output of the model is justified by a commensurate disparity in the construct, thereby allowing accurate models even when the base rates are different for different protected groups, as equalized odds, predictive parity, and calibration do. In addition, because it uses the construct, it does not depend upon the possibly biased ground truth. Using the often unobservable construct can make testing for disparity amplification impossible; we argue that its value instead comes from its ability to organize the space of empirical tests.

In particular, one of our main contributions is our argument that the WAE and WYSIWYG worldviews, when combined with the desire to avoid disparity amplification, motivate demographic parity and equalized odds, respectively. We thus shed light on why people may disagree about which empirical test of discrimination to apply in a particular setting: Even if they agree on the need to avoid disparity amplification, they may disagree about the correct worldview to apply in that setting. We also show that, regardless of the worldview and the base rates of the observed ground truth, predictive parity does not impose any restrictions on the extent to which a model amplifies disparity. Calibration is more restrictive in this regard, but the common post-processing method of thresholding can amplify disparity to an arbitrary extent. Since equalized odds is incompatible with predictive parity or calibration [6, 8, 23], this is an argument for the use of equalized odds instead of predictive parity or calibration. Furthermore, we compare our approach to that of Zafar et al. [34] in their work on *disparate mistreatment*, or disparate misclassification rates, showing that the definition of disparity amplification can be modified to apply in their setting.

Although the WAE and WYSIWYG worldviews are useful for theoretical analysis, they are unlikely to be true in practice. To remedy this issue, we introduce a family of hybrid worldviews that is parametrized by a measure of how biased the observed data is against a protected group of people. This allows us to model many real-world situations by simply adjusting the parameter. We then create a parametrized test for discrimination that corresponds to the new family of worldviews, showing how one can apply the analysis in our paper to more realistic scenarios.

Our most fundamental contribution is introducing a framework in which to motivate empirical tests in terms of construct-based criteria of discrimination and worldviews. Disparity amplification is not the only relevant notion of discrimination, nor is it suitable in every context. Indeed, there are many other aspects of discrimination that we do not address in this paper, such as intentional discrimination [2, §II-A], individual fairness [11], proxy discrimination [9], delayed outcomes [26], and affirmative action [22]. Future work may use our approach to tease out the assumptions implicit in these tests.

We view the discussed tests and disparity amplification as diagnostics that can lead to further investigations of potentially discriminatory behavior in a model. As a result, we do not provide an algorithm for ensuring that a model does not have disparity amplification since, in our view, doing so would be treating the symptom rather than the cause. Such algorithms can eliminate one aspect of discrimination, but may in the process create a model that is obviously discriminatory from another angle. When a model does not satisfy a notion of nondiscrimination, it should be a starting point for investigation as to why. While it could be that the learning algorithm is corrupt, it could also be due to a mismatch between the construct and the observed data, or a need for better features. No one test or criterion can ensure fairness [17], and no single algorithm will be appropriate in all cases.

2 RELATED WORK

Our work is most similar in structure to that of Heidari et al. [20], who propose a unifying framework that reformulates some existing fairness definitions through the lens of equality of opportunity from political philosophy [29, 30]. They then propose a new fairness definition that is inspired by this lens. Although we also present a unifying framework, our unification is through the lens of constructs and worldviews.

Friedler et al. [14] introduced the concept of the construct in fair machine learning. Although they also use the construct in their definition of nondiscrimination, their definition uses the Gromov–Wasserstein distance and as a result is more difficult to compute and reason about. One benefit of their approach is that it enables their treatment of fairness at both the individual level and the group level. By contrast, we consider group nondiscrimination only, and this allows us to draw a parallel between the worldviews and the existing empirical tests of discrimination.

Barocas and Selbst [2] discuss in detail the potential legal issues with discrimination in machine learning. One widely consulted legal standard for detecting disparate impact is the *four-fifths rule* [12]. The four-fifths rule is a guideline that checks whether the ratio of the rates of favorable outcomes for different demographic groups is at least four-fifths. This guideline can be considered a relaxation of demographic parity, which would instead require that the ratio of the positive classification rates be exactly one.

The four-fifths rule has inspired the work of Feldman et al. [13] and Zafar et al. [35], who deal with a generalization of the four-fifths rule, called the *p% rule*, in their efforts to remove disparate impact. On the other hand, many others [4, 5, 21, 36] consider the difference, rather than the ratio, of the positive classification rates. Our discrimination criterion is a generalization of this difference-based measure, but it differs from the others in that it uses the construct rather than the observed data.

Other works in the field of fair machine learning deal with aspects of discrimination that are not well described by positive classification rates. Hardt et al. [19] characterize nondiscrimination through *equalized odds*, which requires that two measures of misclassification, false positive and false negative rates, be equal for all protected groups. *Calibration*, Chouldechova [6] points out, is widely accepted in the "educational and psychological testing and assessment literature". In another work, Friedler et al. [15] create a benchmark for empirically evaluating the consequences of imposing these and other definitions of fairness, finding that many, but not all, definitions lead to similar model behavior.

Dwork et al. [11] formally define *individual fairness* and give examples of cases where models are blatantly unfair at the individual

level even though they satisfy demographic parity. Although individual fairness is sometimes considered to be in conflict with group-based notions of fairness, Binns [3] argues otherwise, instead pointing to the difference in worldviews as the truly important factor. He then lists demographic parity and calibration as corresponding to the WAE and WYSIWYG worldviews, respectively. For the WYSIWYG worldview, he reasons that if calibration is satisfied, no applicant would receive a less favorable outcome than a less qualified applicant, assuming that the calibrated scores accurately describe the degree to which the applicant is qualified. By contrast, in this paper we prove that equalized odds, but not calibration, is an effective way to avoid disparity amplification under the WYSIWYG worldview.

As mentioned previously, discriminatory effects can be justified if there is a sufficient reason. For prediction tasks, it is natural to think of accuracy as a sufficient justification. Zafar et al. [35] handle this by solving an optimization problem to maximize fairness subject to some accuracy constraints. This reflects the idea that a classifier is justified in sacrificing fairness for accuracy. To a lesser extent, equalized odds, predictive parity, and calibration can also be thought of as motivated by the dual desires for accuracy and fairness. Our approach to justification is also motivated by these desires, but we use the construct and say that a classifier is justified in predicting the construct correctly.

3 NOTATION

In the framework introduced by Friedler et al. [14], there are three spaces that describe the target attribute of a prediction model. The *construct space* represents the value of the attribute that is truly relevant for the prediction task. This value is usually unobservable, so prediction models in a supervised learning problem are instead trained with a related measurable label, whose values reside in the *observed space*. Finally, the *prediction space* (called *decision space* by Friedler et al.) describes the output of the model. We will use Y', Y, and $\hat{Y}$ as the random variables representing values from the construct, observed, and prediction spaces, respectively. (See Figure 1.)

In addition, we will use Z to denote the protected attribute at hand, and we will assume that $Z \in \{0, 1\}$. For example, if Z is gender, the values 0 and 1 could represent male and female, respectively. Although the input features $X = (X_1, \ldots, X_n)$ are also critical for both the training and the prediction of the model, they are rarely used in this paper.

Example 1. Some jurisdictions have started to use machine learning models to predict how much risk a criminal defendant poses [25]. Judges are then allowed to consider the risk score as one of many factors when making bail or sentencing decisions [32]. Using the three-space framework of Friedler et al. [14], we can represent the risk score output by the model as $\hat{Y}$. The model would be trained with the observation Y, which in this case may be recorded data about past criminal defendants and their failures to appear in court (bail) or recidivism (sentencing). These models would also be trained with features X from the input space, such as age and criminal history.

For sentencing decisions, presumably we want to know whether the defendant will commit another crime in the future, regardless of whether the defendant will be caught committing the crime. Therefore, we argue that the recorded recidivism rate Y is merely a proxy for the actual reoffense rate Y', which is the relevant attribute for the prediction task. There is evidence that Black Americans are arrested at a higher rate than White Americans for the same crime [27], so it is reasonable to suspect that Y is a racially biased proxy for Y'.

Example 2. Universities want the students that they admit to the university to be successful in the university (Y'). Because *success* is a vague term that encompasses many factors, a model that predicts success in university would instead be trained with a more concrete measure, such as graduating within six years (Y). This model may take inputs such as a student's high-school grades and standardized test scores (X), and will output a prediction of how likely the student is to graduate within six years ($\hat{Y}$). Admissions officers can then use this prediction to guide their decision about whether to admit the student.

It is important to note that the models in the above examples do not make the final decision and that human judgments are a major part of the decision process. However, we are concerned about the fairness of the model rather than that of the entire decision process. Thus, we focus on $\hat{Y}$, the output of the model, rather than the final decision made using it.

4 PRELIMINARY DEFINITIONS

In this work, we use two notions of distance between two random variables that measure how different the random variables are. When the random variables are categorical, we use the total variation distance.

Definition 1 (Total Variation Distance). *Let Y_0 and Y_1 be categorical random variables with finite supports $\mathcal{Y}_0$ and $\mathcal{Y}_1$. Then, the* total variation distance *between Y_0 and Y_1 is*

$$d_{\text{tv}}(Y_0, Y_1) = \frac{1}{2} \sum_{y \in \mathcal{Y}_0 \cup \mathcal{Y}_1} \left| \Pr[Y_0{=}y] - \Pr[Y_1{=}y] \right|.$$

In the special case where $Y_0, Y_1 \in \{0, 1\}$, the total variation distance can also be expressed as $|\Pr[Y_0{=}1] - \Pr[Y_1{=}1]|$.

When the random variables are numerical, our notion of distance takes into account the magnitude of the difference in the numerical values. The following definition assumes that the random variables are continuous, but a similar definition is applicable when they are discrete.

Definition 2 (Earthmover Distance). *Let Y_0 and Y_1 be continuous numerical random variables with probability density functions p_0 and p_1 defined over support $\mathcal{Y}$. Furthermore, let Γ be the set of joint probability density functions $\gamma(u, v)$ such that $\int_{\mathcal{Y}} \gamma(u, v)\, dv = p_0(u)$ for all $u \in \mathcal{Y}$ and $\int_{\mathcal{Y}} \gamma(u, v)\, du = p_1(v)$ for all $v \in \mathcal{Y}$. Then, the* earthmover distance *between Y_0 and Y_1 is*

$$d_{\text{em}}(Y_0, Y_1) = \inf_{\gamma \in \Gamma} \int_{\mathcal{Y}} \int_{\mathcal{Y}} \gamma(u, v)\, d(u, v)\, du\, dv,$$

where d is a distance metric defined over $\mathcal{Y}$.

The joint probability density function γ has marginal distributions that correspond to Y_0 and Y_1. Intuitively, if we use the graphs

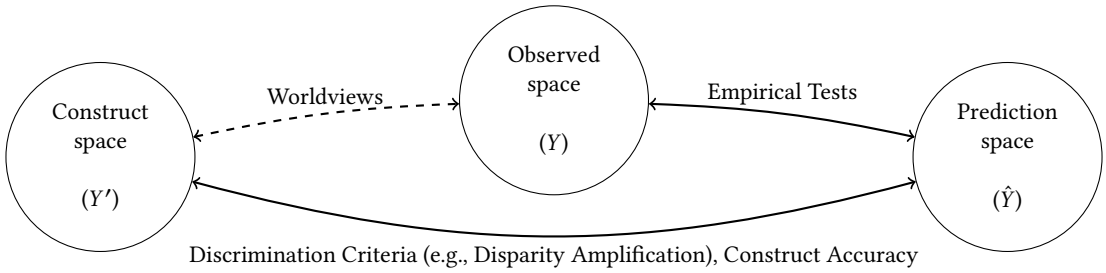

Figure 1: Three relevant spaces for prediction models. The space of input features $X = (X_1, \ldots, X_n)$ is not depicted here. The observed space and the prediction space are measurable, and the existing empirical tests (Definitions 4, 5, 6) impose constraints on the relationship between the two spaces. On the other hand, the construct space is usually unobservable, so we must assume a particular worldview (e.g., Worldview 1 or 2) about how the construct space relates to the observed space, if at all. Then, we can define disparity amplification and construct accuracy, which relate the construct space to the prediction space.

of the probability density functions p_0 and p_1 to represent mounds of sand, γ corresponds to a transportation plan that dictates how much sand to transport in order to reshape the p_0 mound into the p_1 mound. In particular, the value of $\gamma(u, v)$ is the amount of sand to be transported from u to v. The distance $d(u, v)$ can then be interpreted as the cost of transporting one unit of sand from u to v, and the earthmover distance is simply the cost of the transportation plan γ that incurs the least cost.

Now we define Lipschitz continuity.

Definition 3. *Let $f : \mathcal{Y} \to \mathbb{R}$ be a function, and let d be a distance metric defined over $\mathcal{Y}$. f is* ρ-Lipschitz continuous *if, for all $u, v \in \mathcal{Y}$,*

$$|f(u) - f(v)| \leq \rho \cdot d(u, v). \tag{1}$$

4.1 Existing Empirical Tests of Discrimination

Many fairness definitions for prediction models have been proposed previously, and here we restate four of them. Because much of the prior work does not make the distinction between the construct space and the observed space, there is some ambiguity about whether Y' or Y is the appropriate variable to use these definitions. Given that these works suggest that these definitions can be computed, we interpret them to be *empirical tests* that can help verify whether a model is fair. As a result, none of these definitions include the construct Y'. In all four definitions, the probabilities are taken over random draws of data points from the data distribution, as well as any randomness used by the model.

Definition 4 (Demographic Parity Test). *A model passes the* demographic parity test *if, for all $\hat{y}$,*

$$\Pr[\hat{Y}{=}\hat{y} \mid Z{=}0] = \Pr[\hat{Y}{=}\hat{y} \mid Z{=}1].$$

Definition 5 (Equalized Odds Test [19]). *A model passes the* equalized odds test *if, for all y and $\hat{y}$,*

$$\Pr[\hat{Y}{=}\hat{y} \mid Y{=}y, Z{=}0] = \Pr[\hat{Y}{=}\hat{y} \mid Y{=}y, Z{=}1].$$

Definition 6 (Predictive Parity Test [6]). *A model passes the* predictive parity test *if, for all y and $\hat{y}$,*

$$\Pr[Y{=}y \mid \hat{Y}{=}\hat{y}, Z{=}0] = \Pr[Y{=}y \mid \hat{Y}{=}\hat{y}, Z{=}1].$$

Unlike the above three tests, the calibration test is only defined for binary observations, i.e., $Y \in \{0, 1\}$.

Definition 7 (Calibration Test [6]). *A model with a binary Y passes the* calibration test *if, for all $\hat{y}$ in the support of $\hat{Y}$,*

$$\Pr[Y{=}1 \mid \hat{Y}{=}\hat{y}, Z{=}0] = \Pr[Y{=}1 \mid \hat{Y}{=}\hat{y}, Z{=}1] = \hat{y}.$$

4.2 Worldviews

Our intuitive notion of discrimination involves the relationship between the construct space and the prediction space. For example, consider the context of recidivism prediction described in Example 1. Suppose that one group of people is much more likely to be arrested for the same crime than another group. Then, the disparity in arrest rates can cause the recorded recidivism rate Y to be biased, and a model trained using such Y would likely learn to discriminate as a result. If in fact the two groups have equal reoffense rates Y', it would hardly be considered justified that one group tends to be given longer sentences as a result of the bias in Y.

However, because Y' is typically unobservable, in practice we do not know whether Y' is the same for both groups. Therefore, to reason about discrimination using the construct space, we must make assumptions about the construct space. Two such assumptions, or *worldviews*, have previously been introduced by Friedler et al. [14] and are described below. Our versions of these worldviews are simpler than the original because they are exact, whereas the original versions allow deviations by a parameter ϵ.

Worldview 1 (We're All Equal). *Under the* We're All Equal (WAE) *worldview, every group is identical with respect to the construct space. More formally, Y' is independent of Z, i.e., $Y' \perp Z$.*

Worldview 2 (WYSIWYG). *Under the* What You See Is What You Get (WYSIWYG) *worldview, the observed space accurately reflects the construct space. More formally, $Y' = Y$.*

5 CONSTRUCT CRITERIA

We introduce two construct criteria for models. By using the construct, these criteria must be combined with a worldview for application to a model. Unlike the more readily applied empirical tests, construct criteria depend upon the attribute truly relevant to the classification task.

Here, we consider the case where Y' and $\hat{Y}$ are categorical (but not necessarily binary), and in Section 9 we generalize the definition to numerical Y'.

5.1 Disparity Amplification

When $\hat{Y}$ is binary, the size of a model's discriminatory effect is commonly measured by the difference in positive classification rates: $|\Pr[\hat{Y}{=}1 \mid Z{=}0] - \Pr[\hat{Y}{=}1 \mid Z{=}1]|$. Output disparity generalizes this measure for the case of non-binary categorical $\hat{Y}$.

Definition 8 (Output Disparity). *Let the output $\hat{Y}$ of a model be categorical. The* output disparity *of the model is the quantity* $d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1)$.

However, not all output disparities are bad in every context. In particular, because we want the model to accurately reflect the construct, we allow an output disparity insofar as it can be explained by the inter-group disparity in Y'. This happens when

$$d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) \leq d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1). \tag{2}$$

Since a model can have issues with discrimination that are not characterized by output disparity (see below), (2) is not the conclusive definition of nondiscrimination. Thus, we use the logical negation of (2) as a criterion for one particular discrimination concern, which occurs when an output disparity is *not* explained by Y'.

Definition 9 (Disparity Amplification). *Let Y' and $\hat{Y}$ be categorical. Then, a model exhibits* disparity amplification *if*

$$d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) > d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1). \tag{3}$$

5.2 Construct Accuracy

As mentioned in Section 5.1, we want the output of the model to accurately reflect the value of Y'. However, the simple accuracy measure $\Pr[Y' = Y]$ incentivizes the model to become more accurate on the larger protected group at the expense of becoming less accurate on the smaller protected group. Therefore, we instead measure accuracy as the average of the accuracy on the two groups.

Definition 10 (Construct Accuracy). *The* construct accuracy *of a model is*

$$\tfrac{1}{2}\big(\Pr[Y'{=}\hat{Y} \mid Z{=}0] + \Pr[Y'{=}\hat{Y} \mid Z{=}1]\big). \tag{4}$$

Definition 11 (Construct Optimality). *A model is* construct optimal *if its construct accuracy is 1, i.e., its output $\hat{Y}$ and the construct Y' are always equal.*

Because the construct Y' usually cannot be observed, construct accuracy usually cannot be measured or directly optimized for. Even when it can measured, construct optimality would be rare since the quality of the features, data, or machine learning algorithm may preclude perfection. As with disparity amplification, we introduce construct accuracy not to empirically measure it, but as a theoretical tool for analyzing discrimination. In particular, note that equality holds in (2) for every construct optimal model. In other words, a construct optimal model displays the maximum amount of output disparity allowed by Definition 9. On the other hand, if the output disparity is greater than the disparity in Y', the model must be amplifying a disparity in a way that cannot be justified by the desire to achieve construct optimality.

The above definitions can be generalized to the setting where the range $\mathcal{Y}'$ of the values that Y' takes differs from the range $\hat{\mathcal{Y}}$ of $\hat{Y}$. If there exists a bijective mapping between $\mathcal{Y}'$ and $\hat{\mathcal{Y}}$, we can use the mapping to characterize when a value from $\hat{\mathcal{Y}}$ accurately reflects a value from $\mathcal{Y}'$.

5.3 Limitations

These criteria, separately or jointly, are neither necessary nor sufficient for fairness. Technical criteria allow precision but elide the context-specific and social aspects of fairness [17].

The criteria fail to be sufficient for fairness by not capturing forms of discrimination unrelated to output disparity. For example, a model could have a higher misclassification rate for one group of people [34], which goes undetected by Definition 9. (See Section 7 for discussion.) Furthermore, by examining just a model's input/output behavior, the criteria cannot catch a model produced by an unacceptable process or performing unacceptable computations internally to reach its outputs. For example, Datta et al. [9] show the impossibility of externally detecting whether a model internally reconstructs a sensitive attribute that it should not use.

We believe avoiding disparity amplification does better as a necessary condition for fairness, but limitations exist here as well. For example, when correcting historical wrongs, it may be fair to amplify certain disparities that benefit an oppressed group. Such cases also provide a counterexample to the necessity of construct accuracy. In some cases, carefully selecting a historically informed construct can avoid violating our criteria while achieving a reparative goal. However, some goals, such as achieving adequate representation for a group, cannot be expressed in terms of an individual-level construct. Nevertheless, our criteria highlight when a model's behavior is suspicious enough to warrant an explanation and can serve as a basis for selecting between empirical tests.

6 USING CRITERIA AND WORLDVIEWS TO MOTIVATE EMPIRICAL TESTS

In this section, we use our construct criteria to analyze which worldviews motivate the existing empirical tests of discrimination. If an empirical test does not guarantee the lack of disparity amplification, it may not be sufficient as an anti-discrimination measure as it effectively allows certain forms of discrimination. On the other hand, if the test disallows a construct optimal model, the test may be too strict in a way that lowers the utility of the model. Therefore, to argue that a worldview motivates an empirical test, we will prove the following two statements: (a) Every model that passes the empirical test does not have disparity amplification, and (b) every optimal model passes the empirical test.

We apply this reasoning to demographic parity (Definition 4) and equalized odds (Definition 5), showing that the WAE and WYSIWYG worldviews, respectively, motivate these empirical tests. More formally, we will prove statements (a) and (b) for every joint distribution of Y', Y, $\hat{Y}$, and Z that is consistent with the worldview. Table 1 summarizes these results.

6.1 Demographic Parity and WAE

Theorem 1. *A model that passes the demographic parity test does not have disparity amplification under Definition 9. Moreover,*

Table 1: Summary of the results in Section 6. We say that a worldview motivates an empirical test if it precludes disparity amplification (Definition 9) but does not preclude a perfectly predictive model. The We're All Equal (WAE) worldview motivates the demographic parity test, and if the worldview does not hold, the demographic parity test tends to lower the utility of the model. The WYSIWYG worldview motivates the equalized odds test, and if the worldview does not hold, the equalized odds test allows models that have disparity amplification. Finally, regardless of the worldview, the predictive parity and calibration tests do not effectively prevent disparity amplification. Here, we assume that WAE and WYSIWYG do not hold simultaneously.

	We're All Equal (Worldview 1)	WYSIWYG (Worldview 2)
Demo. Parity (Definition 4)	✔ Theorem 1	Always suboptimal Theorem 2
Equal. Odds (Definition 5)	Amplification allowed Theorem 5	✔ Theorem 4
Predictive Parity (Definition 6)	Amplification allowed Theorem 6	
Calibration (Definition 7)	Not robust to post-processing Theorem 7	

if the WAE worldview holds, every construct optimal model satisfies demographic parity.

Proof. By the definition of demographic parity, the left-hand side of (3) is $d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) = 0$. Since the total variation distance is always nonnegative, demographic parity ensures the lack of disparity amplification.

If the WAE worldview holds, we have $Y' \perp Z$, so every optimal model satisfies $\hat{Y} \perp Z$. This implies demographic parity by Definition 4. □

The first part of Theorem 1 shows that we can guarantee that a model will not have disparity amplification by training it to pass the demographic parity test. However, this does not mean that demographic parity is appropriate for every situation. First, we remind the reader that the lack of disparity amplification does not mean that the model will be free of all issues related to discrimination. In particular, disparity amplification is only designed to catch the type of discrimination akin to *disparate impact*. If the WAE worldview holds, demographic parity is the only way to avoid disparity amplification, so it makes sense to enforce demographic parity. On the other hand, blindly enforcing demographic parity may introduce other forms of discrimination. For example, the U.S. Supreme Court held in *Ricci v. DeStefano* [31] that the prohibition against intentional discrimination can sometimes override the consideration of disparate impact, ruling that an employer unlawfully discriminated by discarding the results of a bona fide job-related test because of a racial performance gap.

Second, demographic parity can lower the utility of a model. If the WAE worldview does not hold, $d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1)$ is positive, and Theorem 2 shows that any model that satisfies demographic parity must be suboptimal. In fact, the more we deviate from the WAE worldview, the lower the maximum possible construct accuracy becomes.

Theorem 2. *If a model satisfies demographic parity, the construct accuracy of the model is at most* $1 - \frac{1}{2} d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1)$. *Moreover, there exists a distribution of* $\hat{Y}$ *that satisfies demographic parity and attains this construct accuracy.*

To prove this theorem, we will use Lemma 3.

Lemma 3. *Let* Y_0 *and* Y_1 *be categorical random variables with finite supports* $\mathcal{Y}_0$ *and* $\mathcal{Y}_1$. *Then,*

$$\sum_{y \in \mathcal{Y}_0 \cup \mathcal{Y}_1} \min\big(\Pr[Y_0{=}y], \Pr[Y_1{=}y]\big) = 1 - d_{\text{tv}}(Y_0, Y_1).$$

Proof of Lemma 3. First, we can express the total variation distance in terms of max and min.

$$\sum_{y \in \mathcal{Y}_0 \cup \mathcal{Y}_1} \max\big(\Pr[Y_0{=}y], \Pr[Y_1{=}y]\big) - \min\big(\Pr[Y_0{=}y], \Pr[Y_1{=}y]\big)$$
$$= \sum_{y \in \mathcal{Y}_0 \cup \mathcal{Y}_1} \big|\Pr[Y_0{=}y] - \Pr[Y_1{=}y]\big| = 2 d_{\text{tv}}(Y_0, Y_1).$$

In addition, we have

$$\sum_{y \in \mathcal{Y}_0 \cup \mathcal{Y}_1} \max\big(\Pr[Y_0{=}y], \Pr[Y_1{=}y]\big) + \min\big(\Pr[Y_0{=}y], \Pr[Y_1{=}y]\big)$$
$$= \sum_{y \in \mathcal{Y}_0 \cup \mathcal{Y}_1} \big(\Pr[Y_0{=}y] + \Pr[Y_1{=}y]\big) = 2.$$

Subtracting the first equation from the second gives us the desired result. □

Proof of Theorem 2. We first prove the upper bound on the construct accuracy. Let $\mathcal{Y}'$ and $\hat{\mathcal{Y}}$ be the supports of Y' and $\hat{Y}$, respectively. Then, by the law of total probability we have

$$\Pr[Y'{=}y', \hat{Y}{=}y' \mid Z{=}z] \le \min(\Pr[Y'{=}y' \mid Z{=}z], \Pr[\hat{Y}{=}y' \mid Z{=}z])$$

for all $y' \in \mathcal{Y}' \cup \hat{\mathcal{Y}}$ and $z \in \{0, 1\}$. We then sum this over y' and apply Lemma 3 to get

$$\begin{aligned}
&\Pr[Y'{=}\hat{Y} \mid Z{=}z] \\
&= \textstyle\sum_{y' \in \mathcal{Y}' \cup \hat{\mathcal{Y}}} \Pr[Y'{=}y', \hat{Y}{=}y' \mid Z{=}z] \\
&\le \textstyle\sum_{y' \in \mathcal{Y}' \cup \hat{\mathcal{Y}}} \min\big(\Pr[Y'{=}y' \mid Z{=}z], \Pr[\hat{Y}{=}y' \mid Z{=}z]\big) \\
&= 1 - d_{\text{tv}}(Y'|Z{=}z, \hat{Y}|Z{=}z) \\
&= 1 - d_{\text{tv}}(Y'|Z{=}z, \hat{Y}),
\end{aligned}$$

where the last equality follows from our assumption that the model satisfies demographic parity. Therefore, the construct accuracy can be bounded as

$$\begin{aligned}
&\tfrac{1}{2}\big(\Pr[Y'{=}\hat{Y} \mid Z{=}0] + \Pr[Y'{=}\hat{Y} \mid Z{=}1]\big) \\
&\le \tfrac{1}{2}\big(1 - d_{\text{tv}}(Y'|Z{=}0, \hat{Y}) + 1 - d_{\text{tv}}(Y'|Z{=}1, \hat{Y})\big) \\
&\le 1 - \tfrac{1}{2} d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1),
\end{aligned}$$

where the last inequality is an application of the triangle inequality.

Now we construct a random variable $\hat{Y}$ that satisfies demographic parity and attains this bound. When $Z{=}0$, we simply let $\hat{Y} = Y'$, making the first term in (4) equal to 1. When $Z{=}1$, we

constrain the marginal distribution of $(\hat{Y}|Z{=}1)$ to be the same as that of $(\hat{Y}|Z{=}0) = (Y'|Z{=}0)$, and we make the joint distribution of $(Y'|Z{=}1)$ and $(\hat{Y}|Z{=}1)$ a maximal coupling [24, pp. 19–20]. Then, by the theorem in [24, p. 19], such $\hat{Y}$ attains the value of $1 - d_{\text{tv}}(\hat{Y}|Z{=}1, Y'{=}1) = 1 - d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1)$ for the second term of (4). This means that the construct accuracy, which is the average of the two terms, is $1 - \frac{1}{2}d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1)$, which is what we want. Moreover, $(\hat{Y}|Z{=}1)$ and $(\hat{Y}|Z{=}0)$ have the same distribution, so $\hat{Y}$ satisfies demographic parity. □

Theorems 1 and 2 demonstrate that the WAE worldview, combined with the desire to avoid disparity amplification while retaining the utility of models, motivates the demographic parity test.

6.2 Equalized Odds and WYSIWYG

We now argue that a similar relationship exists between the equalized odds test and the WYSIWYG worldview.

Theorem 4. *If the WYSIWYG worldview holds, a model that passes the equalized odds test does not have disparity amplification under Definition 9. Moreover, if the WYSIWYG worldview holds, every construct optimal model satisfies equalized odds.*

Proof. Let $\mathcal{Y}'$ and $\hat{\mathcal{Y}}$ be the supports of Y' and $\hat{Y}$, respectively. Applying the WYSIWYG worldview to the definition of equalized odds, we get $\Pr[\hat{Y}{=}\hat{y} \mid Y'{=}y', Z{=}0] = \Pr[\hat{Y}{=}\hat{y} \mid Y'{=}y', Z{=}1] = \Pr[\hat{Y}{=}\hat{y} \mid Y'{=}y']$ for all $y' \in \mathcal{Y}'$ and $\hat{y} \in \hat{\mathcal{Y}}$. Therefore, we have

$$
\begin{aligned}
&d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) \\
&= \tfrac{1}{2} \textstyle\sum_{\hat{y}\in\hat{\mathcal{Y}}} \left|\Pr[\hat{Y}{=}\hat{y} \mid Z{=}0] - \Pr[\hat{Y}{=}\hat{y} \mid Z{=}1]\right| \\
&= \tfrac{1}{2} \textstyle\sum_{\hat{y}\in\hat{\mathcal{Y}}} \Big| \sum_{y'\in\mathcal{Y}'} \Pr[\hat{Y}{=}\hat{y} \mid Y'{=}y'] \\
&\qquad\qquad \cdot \big(\Pr[Y'{=}y' \mid Z{=}0] - \Pr[Y'{=}y' \mid Z{=}1]\big)\Big| \\
&\le \tfrac{1}{2} \textstyle\sum_{\hat{y}\in\hat{\mathcal{Y}}} \sum_{y'\in\mathcal{Y}'} \Pr[\hat{Y}{=}\hat{y} \mid Y'{=}y'] \\
&\qquad\qquad \cdot \left|\Pr[Y'{=}y' \mid Z{=}0] - \Pr[Y'{=}y' \mid Z{=}1]\right| \\
&= \tfrac{1}{2} \textstyle\sum_{y'\in\mathcal{Y}'} \Big(\left|\Pr[Y'{=}y' \mid Z{=}0] - \Pr[Y'{=}y' \mid Z{=}1]\right| \\
&\qquad\qquad \cdot \textstyle\sum_{\hat{y}\in\hat{\mathcal{Y}}} \Pr[\hat{Y}{=}\hat{y} \mid Y'{=}y']\Big) \\
&= \tfrac{1}{2} \textstyle\sum_{y'\in\mathcal{Y}'} \left|\Pr[Y'{=}y' \mid Z{=}0] - \Pr[Y'{=}y' \mid Z{=}1]\right| \\
&= d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1).
\end{aligned}
$$

This concludes the proof of the first statement.

For an optimal model, we have $\hat{Y} = Y' = Y$ by the WYSIWYG worldview. Because Y fully determines the value of $\hat{Y}$, Definition 5 implies that every optimal model satisfies equalized odds. □

On the other hand, our intuition is that when the observation process is biased, and WYSIWYG does not hold, treating the observation Y as accurate, as implicit with equalized odds, may lead to a failure to pass our construct-based criterion. We prove as much:

Theorem 5. *If the WYSIWYG worldview does not hold, a model passing the equalized odd test can still have disparity amplification.*

Proof. We show that there exists a joint distribution of Y', Y, $\hat{Y}$, and Z such that a model with equalized odds still has disparity amplification. Many models with equalized odds have nonzero output disparity, i.e., $d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) > 0$. Consider any such model. Since the WYSIWYG worldview does not hold, we have no guarantee that Y' will resemble Y in any way. Therefore, the equalized odds requirement does not restrict the distribution of Y', and the model can have disparity amplification if $d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1)$ is small enough. □

6.3 Predictive Parity

Under the WYSIWYG worldview, optimal models pass the predictive parity test, but any model that passes the test must satisfy $d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) \ge d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1)$, as can be seen from switching Y' and $\hat{Y}$ in the proof of the first part of Theorem 4. The inequality here is in the opposite direction of that in (2), so the predictive parity test does not place any upper bound on the output disparity of $\hat{Y}$ and guarantees that it is equal to that of Y' or amplified beyond this limit. In fact, the following theorem shows that, regardless of the worldview and the base rates of Y, even a model with almost the maximum output disparity can still pass the predictive parity test.

Theorem 6. *Let Y be a categorical random variable with finite support such that $\Pr[Y{=}y \mid Z{=}z]$ is positive for all y and z. Then, for any sufficiently small $\epsilon > 0$, there exists a model that passes the predictive parity test such that $d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) = 1 - \epsilon$.*

Proof. The main idea behind the proof is that the model simply outputs the value of Z. However, because predictive parity is not well-defined if $\Pr[\hat{Y}{=}\hat{y}, Z{=}z] = 0$ for any $\hat{y}$ and z, we must allow the model to output the other value with some very small probability. More specifically, we construct a model such that

$$
\Pr[\hat{Y}{=}\hat{y} \mid Z{=}z] = \begin{cases} 1 - \frac{\epsilon}{2}, & \text{if } \hat{y} = z \\ \frac{\epsilon}{2}, & \text{if } \hat{y} \neq z. \end{cases}
$$

We can choose which values our constructed model outputs, so assume without loss of generality that $\hat{Y} \in \{0, 1\}$.

Let $\mathcal{Y}$ be the support of Y. By the predictive parity test, we have $\Pr[Y{=}y \mid \hat{Y}{=}\hat{y}, Z{=}0] = \Pr[Y{=}y \mid \hat{Y}{=}\hat{y}, Z{=}1] = \Pr[Y{=}y \mid \hat{Y}{=}\hat{y}]$ for all $y \in \mathcal{Y}$ and $\hat{y} \in \{0, 1\}$. Let $p_{y\hat{y}} = \Pr[Y{=}y \mid \hat{Y}{=}\hat{y}]$. Our goal is to find the values of p_{y0} and p_{y1} that are consistent with the fixed observed probabilities $\Pr[Y{=}y \mid Z{=}0]$ and $\Pr[Y{=}1 \mid Z{=}1]$.

By the law of total probability, our model must satisfy

$$
\begin{pmatrix} \Pr[Y{=}y \mid Z{=}0] \\ \Pr[Y{=}y \mid Z{=}1] \end{pmatrix} = \begin{pmatrix} 1 - \frac{\epsilon}{2} & \frac{\epsilon}{2} \\ \frac{\epsilon}{2} & 1 - \frac{\epsilon}{2} \end{pmatrix} \begin{pmatrix} p_{y0} \\ p_{y1} \end{pmatrix}.
$$

Solving for p_{y0} and p_{y1}, we see that they converge to $\Pr[Y{=}y \mid Z{=}0]$ and $\Pr[Y{=}y \mid Z{=}1]$, respectively, as ϵ approaches zero. By assumption, these probabilities are positive. Since $\mathcal{Y}$ is finite, this means that there exists a small enough $\epsilon > 0$ such that $p_{y0}, p_{y1} > 0$ for all $y \in \mathcal{Y}$. Moreover, it is easy to verify that $\sum_{y\in\mathcal{Y}} p_{y0} = \sum_{y\in\mathcal{Y}} p_{y1} = 1$, making them valid probability distributions.

Now, when given $Y{=}y$ and $Z{=}z$, our model can output $\hat{Y}{=}\hat{y}$ with probability

$$
\Pr[\hat{Y}{=}\hat{y} \mid Y{=}y, Z{=}z] = \frac{p_{y\hat{y}} \cdot \Pr[\hat{Y}{=}\hat{y} \mid Z{=}z]}{\Pr[Y{=}y \mid Z{=}z]},
$$

where $\Pr[\hat{Y}{=}\hat{y} \mid Z{=}z]$ is either $\frac{\epsilon}{2}$ or $1 - \frac{\epsilon}{2}$ depending on whether $\hat{y} = z$. □

Because the predictive parity test allows models, such as the one we constructed in the above proof, that clearly amplify disparity, it is unsuitable for ensuring nondiscrimination as characterized by output disparity.

6.4 Calibration

Compared to the predictive parity test, the calibration test imposes an additional requirement that the output of the model must be the correct probability. Theorem 7 shows this additional requirement limits the model behavior by the disparity in observed values, ruling out the model described in the proof of Theorem 6.

Theorem 7. *If the WYSIWYG worldview holds, a model that passes the calibration test satisfies* $d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1) = |\mathbb{E}[\hat{Y} \mid Z{=}0] - \mathbb{E}[\hat{Y} \mid Z{=}1]|$. *Moreover, if the WYSIWYG worldview holds, every construct optimal model with binary* Y *satisfies calibration.*

Proof. Combining the definition of calibration with the WYSIWYG worldview, we get a binary Y' with $\Pr[Y'{=}1 \mid \hat{Y}{=}\hat{y}, Z{=}0] = \hat{y}$. Therefore, we have

$$\begin{aligned}\Pr[Y'{=}1 \mid Z{=}0] &= \textstyle\sum_{\hat{y}\in\hat{\mathcal{Y}}} \Pr[Y'{=}1 \mid \hat{Y}{=}\hat{y}, Z{=}0] \cdot \Pr[\hat{Y}{=}\hat{y} \mid Z{=}0] \\ &= \textstyle\sum_{\hat{y}\in\hat{\mathcal{Y}}} \hat{y} \cdot \Pr[\hat{Y}{=}\hat{y} \mid Z{=}0] \\ &= \mathbb{E}[\hat{Y} \mid Z{=}0],\end{aligned}$$

and a similar statement holds for $Z{=}1$.

Since Y' is binary, the construct disparity then becomes

$$\begin{aligned}d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1) &= |\Pr[Y'{=}1 \mid Z{=}0] - \Pr[Y'{=}1 \mid Z{=}1]| \\ &= |\mathbb{E}[\hat{Y} \mid Z{=}0] - \mathbb{E}[\hat{Y} \mid Z{=}1]|,\end{aligned}$$

which is what we want for the first statement.

To prove the second statement, note that an optimal model satisfies $\hat{Y} = Y' = Y$ by the WYSIWYG worldview. Then, for binary $Y \in \{0, 1\}$ it is easy to verify that calibration holds. □

Unlike Theorems 1 and 4, which bound the *total variation distance* between the outputs by the disparity in the construct, this theorem bounds only the difference in the *expected values* of the outputs. This contrast is significant because expected value, unlike total variation distances, are not robust to post-processing. We demonstrate this issue with an example where $Y' = Y$ and Z are independent and uniformly random binary variables and the model sets the value of $\hat{Y}$ as follows: if $Z = 0$, then $\hat{Y} = 0.5$; if $Z = 1$ and $Y = 1$, then $\hat{Y} = 0.5 + \epsilon$ for some small positive constant ϵ; and if $Z = 1$ and $Y = 0$, then $\hat{Y} = 0$ with probability $\frac{2\epsilon}{0.5+\epsilon}$ and $\hat{Y} = 0.5 + \epsilon$ otherwise. Some computation reveals that this model passes the calibration test, with all of the $Z = 0$ group receiving a prediction of 0.5 and the vast majority of the $Z = 1$ group receiving $0.5 + \epsilon$. However, in practice the predictions are often post-processed with a threshold because it is impossible to, say, admit half of a student. Therefore, although the inter-group difference in the model predictions is small, it can be amplified if the threshold is set between 0.5 and $0.5 + \epsilon$. In this case, the resulting decision is almost perfectly correlated with Z and exhibits disparity amplification.

As a result, in the rest of the paper we focus on equalized odds rather than predictive parity or calibration. We leave as future work the identification of a discrimination criterion and a worldview that together motivate the predictive parity or calibration test.

7 CONNECTION TO MISCLASSIFICATION

Here, we show that the definition of disparity amplification is closely related to that given by Zafar et al. [34] in their treatment of disparate misclassification rates. First, we motivate the issue of disparate misclassification rates with an example. Let Y' and Z be independent and uniformly random binary variables. If $\hat{Y} = Y' \oplus Z$, where $\oplus$ is the XOR, both protected groups are given the positive label exactly half of the time, so there is no output disparity. However, one group always receives the correct classification and the other always receives the incorrect classification, so the disparity in the misclassification rates is as large as it can be. This shows that a lack of disparity amplification does not imply a lack of disparity in misclassification rates.

Conversely, a lack of disparity in misclassification rates does not imply a lack of disparity amplification. To see this, modify the above example so that $\hat{Y} = Z$ instead. Now, both groups have half of its members misclassified since Z is independent of Y', so they have the same overall misclassification rate. On the other hand, we have $d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1) = d_{\text{tv}}(Y', Y') = 0$ and $d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) = d_{\text{tv}}(Z|Z{=}0, Z|Z{=}1) = 1$. Thus, $\hat{Y}$ has disparity amplification.

However, we can still find a connection between misclassification parity and disparity amplification. Let C be the indicator $\mathbb{1}(Y' = \hat{Y})$, and replace $\hat{Y}$ with C in the definition of output disparity (Definition 8). Since C is binary, the resulting expression $d_{\text{tv}}(C|Z{=}0, C|Z{=}1)$ is simply the difference in the misclassification rates. We would like to compare this value to some measure of disparity in the construct space. Since our standard measure of $d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1)$ does not necessarily justify inter-group differences in C, it may not be a correct measure to use. Exploring what measures provide justification for disparate misclassification rates is interesting future work.

8 HYBRID WORLDVIEWS

So far, we have assumed either the WAE or the WYSIWYG worldview. While these worldviews are interesting from a theoretical perspective, in practice it is unlikely that these worldviews hold.

In this section, we propose a family of more realistic worldviews for the case where Y' and Y are categorical. As we have depicted in Figure 1, worldviews describe the relationship between the construct and observed spaces. Because our definition of disparity amplification has to do with inter-group disparities, here we focus specifically on the inter-group disparities in Y' and Y. Note that the WAE worldview has the effect of assuming that none of the disparity in Y is explained by Y'. By contrast, under the WYSIWYG worldview, all of the disparity in Y is explained by Y'. Described below is the α-Hybrid worldview, which is a family of worldviews that occupy the space between the two extremes of WAE and WYSIWYG.

Worldview 3 (α-Hybrid). *Let* $\alpha \in [0, 1]$. *Under the* α-Hybrid *worldview, exactly an* α *fraction of the disparity in* Y *is explained by*

Y'. *More formally,*

$$d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1) = \alpha \cdot d_{\text{tv}}(Y|Z{=}0, Y|Z{=}1) \tag{5}$$

While the WAE worldview is equivalent to the 0-Hybrid worldview, the relationship between the WYSIWYG and 1-Hybrid worldviews is only unidirectional. Although the WYSIWYG worldview implies the 1-Hybrid worldview, there are plenty of ways to satisfy $d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1) = d_{\text{tv}}(Y|Z{=}0, Y|Z{=}1)$ even when the equality $Y' = Y$ does not hold. If we wanted to make the relationship bidirectional, we could instead have assumed that Y' can be broken down into two components, one of which satisfies WAE and the other WYSIWYG. However, this would mean that every component of Y' is either equal with respect to Z (WAE) or observable (WYSIWYG), whereas in practice many inter-group disparities in the construct space are not easily observable. Thus, to make the α-Hybrid worldview more realistic, we sacrifice one direction of the relationship between the WYSIWYG and 1-Hybrid worldviews.

Now we introduce the α-disparity test and prove that it corresponds to the α-Hybrid worldview. Unlike the demographic parity and equalized odds tests, the α-disparity test is parametrized and therefore can be applied to various real-world situations.

Definition 12 (α-Disparity Test). *A model passes the* α-disparity test *if*

$$d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) \leq \alpha \cdot d_{\text{tv}}(Y|Z{=}0, Y|Z{=}1). \tag{6}$$

Theorem 8. *If the α-Hybrid worldview holds, a model that passes the α-disparity test does not have disparity amplification under Definition 9. Moreover, if the α-Hybrid worldview holds, every construct optimal model satisfies the α-disparity test.*

Proof. To prove the first part of the theorem, we simply combine the inequality guaranteed by the α-disparity test under (6) with the equation that defines the α-Hybrid worldview under (5). We get

$$d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) \leq \alpha \cdot d_{\text{tv}}(Y|Z{=}0, Y|Z{=}1) = d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1),$$

which is what we want.

For the second part of the theorem, an optimal model has $Y' = \hat{Y}$, so we can substitute the Y' in (5) with $\hat{Y}$ to get

$$d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) = \alpha \cdot d_{\text{tv}}(Y|Z{=}0, Y|Z{=}1).$$

This is simply the equality in (6), so we are done. □

The α-disparity test is closely related to demographic parity and equalized odds. 0-disparity is satisfied if and only if the output disparity is zero, so it is equivalent to demographic parity. In addition, we can easily adapt the proof of Theorem 4 to show that equalized odds implies 1-disparity. However, because equalized odds imposes a condition for each possible value of Y, 1-disparity does not imply equalized odds. Although it may thus seem that equalized odds is stronger and better than 1-disparity, recent results by Corbett-Davies and Goel [7] show that the threshold rule, which they argue is optimal, does not lead to equalized odds in general. Therefore, there is a trade-off between the stronger fairness guarantee provided by equalized odds and the higher utility that is attainable under 1-disparity. Of course, the 1-disparity test has the additional benefit that it can be generalized to other values of α.

We end this section with theorems describing the consequences of enforcing the α-disparity test with a wrong value of α. These theorems are close analogues of Theorems 2 and 5, respectively.

Theorem 9. *If the α-Hybrid worldview holds, a model that passes the α'-disparity test, with $\alpha > \alpha'$, has a construct accuracy at most $1 - \frac{1}{2}(\alpha - \alpha') \cdot d_{\text{tv}}(Y|Z{=}0, Y|Z{=}1)$.*

Proof. By the reasoning in the proof of Theorem 2, we have for all $z \in \{0, 1\}$, $\Pr[Y'{=}\hat{Y} \mid Z{=}z] \leq 1 - d_{\text{tv}}(Y'|Z{=}z, \hat{Y}|Z{=}z)$, which can be rewritten as $\Pr[Y'{\neq}\hat{Y} \mid Z{=}z] \geq d_{\text{tv}}(Y'|Z{=}z, \hat{Y}|Z{=}z)$.

Thus, the construct *inaccuracy* of the model is

$$\begin{aligned}
&\tfrac{1}{2}\big(\Pr[Y'{\neq}\hat{Y} \mid Z{=}0] + \Pr[Y'{\neq}\hat{Y} \mid Z{=}1]\big) \\
&\geq \tfrac{1}{2}\big(d_{\text{tv}}(Y'|Z{=}0, \hat{Y}|Z{=}0) + d_{\text{tv}}(Y'|Z{=}1, \hat{Y}|Z{=}1)\big) \\
&\geq \tfrac{1}{2}\big(d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1) - d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1)\big) \\
&\geq \tfrac{1}{2}(\alpha - \alpha') \cdot d_{\text{tv}}(Y|Z{=}0, Y|Z{=}1),
\end{aligned}$$

where the second inequality is an application of the triangle inequality and the third follows from the definitions of the α-Hybrid worldview and the α'-disparity test.

Therefore, the construct accuracy, which is one minus the construct inaccuracy, is at most $1 - \frac{1}{2}(\alpha - \alpha') \cdot d_{\text{tv}}(Y|Z{=}0, Y|Z{=}1)$. □

Theorem 10. *If the α-Hybrid worldview holds, a model that passes the α'-disparity test, with $\alpha < \alpha'$, can still have disparity amplification.*

Proof. The α'-disparity test ensures that

$$d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) \leq \alpha' \cdot d_{\text{tv}}(Y|Z{=}0, Y|Z{=}1),$$

and if equality holds here, we have

$$d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) = \alpha' \cdot d_{\text{tv}}(Y|Z{=}0, Y|Z{=}1) > \alpha \cdot d_{\text{tv}}(Y|Z{=}0, Y|Z{=}1)$$

whenever $d_{\text{tv}}(Y|Z{=}0, Y|Z{=}1) \neq 0$. By the α-Hybrid worldview, the rightmost quantity equals $d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1)$, making the above inequality exactly that of disparity amplification (see (3)). □

9 A MORE GENERAL NOTION OF DISPARITY AMPLIFICATION

In this section, we present a more general definition of disparity amplification that is a broader discrimination criterion and is applicable to numerical Y'. Due to space constraints, proofs of theorems in this section are given in the supplementary material.

Definition 9 allows an output disparity if there *exists* an equally large disparity in Y', but it does not explicitly reflect the fact that we care about *how* the model came to exhibit the disparity. The only reason why we allow the disparity is that Y' is the right attribute to use. Thus, if the model does not use Y' at all, then there should be no output disparity. More formally, we want that if $Y' \perp \hat{Y}$, then $\hat{Y} \perp Z$.

Definition 13 generalizes this requirement and, unlike Definition 9, is applicable for both categorical and numerical Y' at the expense of limiting $\hat{Y}$ to be binary. The generalization deals with cases where $\hat{Y}$ is not independent of Y' by measuring how much $\hat{Y}$ depends upon Y'. For binary $\hat{Y}$, this dependence is captured by the likelihood function $\ell(y') = \Pr[\hat{Y}{=}1 \mid Y'{=}y']$, and we use the Lipschitz continuity of this function to measure the dependence.

Definition 13 (Disparity Amplification, Stronger). *For $\hat{Y} \in \{0,1\}$ and $\ell(y') = \Pr[\hat{Y}{=}1 \mid Y'{=}y']$, let ρ_ℓ^* be the smallest nonnegative ρ such that ℓ is ρ-Lipschitz continuous.*[1] *Then, a model exhibits* disparity amplification *if*

$$d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) > \rho_\ell^* \cdot d_{\text{em}}(Y'|Z{=}0, Y'|Z{=}1). \tag{7}$$

ρ_ℓ^* characterizes how much impact Y' can have on the output of the model. If the impact is small, we can conclude that the model is not using Y' much, so not much output disparity can be explained by Y'. On the other hand, if a small change in Y' can cause a large change in the probability distribution of $\hat{Y}$, then even a large output disparity can possibly be due to a small inter-group difference in Y'. In fact, the use of ρ_ℓ^* makes Definition 13 invariant to scaling in Y'. If a numerical Y' is increased by some factor, ρ_ℓ^* will decrease by the same factor, so the quantity on the right-hand side of (7) will not change.

We now show relationships between the new Definition 13 and the previous definition (Definition 9). First, we show that the old definition combined with a reasonable distance metric implies the new definition. The previous definition assumes that Y' is categorical, and in this case a natural distance metric for its support $\mathcal{Y}'$ is the indicator $d(u,v) = \mathbb{1}(u \neq v)$. With this distance metric, we can relate the total variation distance used in the right-hand side of (3) with the earthmover distance used in (7).

Theorem 11. *Let the construct Y' be categorical with support $\mathcal{Y}'$, which has distance metric $d(u,v) = \mathbb{1}(u \neq v)$. If a model has disparity amplification under Definition 9, the model has disparity amplification under Definition 13 as well.*

The proof relies upon a theorem using coupling [16, Theorem 4].

Second, we show that Theorems 1 and 4 still hold under the refined definition of disparity amplification. Since the definitions of optimality and the empirical tests have not changed, we focus strictly on the nondiscrimination portions of the theorems.

Theorem 12. *A model that passes the demographic parity test does not have disparity amplification under Definition 13.*

The proof of Theorem 12 is very similar to that of Theorem 1.

Theorem 13. *If the WYSIWYG worldview holds, then a model that passes the equalized odds test does not have disparity amplification under Definition 13.*

This proof uses Kantorovich duality [33, Equation 5.4].

We now discuss the tightness of the above result. In the extreme case where ℓ is a step function over real-valued y', ρ_ℓ^* is infinite, so we trivially have a lack of disparity amplification under Definition 13. Thus, to receive meaningful fairness guarantees from Theorem 13, we must make sure that ρ_ℓ^* is not too large. One way to achieve this is to apply the function ℓ to the construct space and reason about the transformed construct space. If any transformation of the construct space results in a finding of disparity amplification under Definition 13, then it is evidence that there could be a problem with the model with respect to discrimination.

[1] Technically, ρ_ℓ^* should be the *infimum* of all ρ such that ℓ is ρ-Lipschitz continuous, but it is not difficult to show then that ℓ is in fact ρ_ℓ^*-Lipschitz continuous.

Let $\tilde{y}' = \ell(y')$ be a value in the transformed construct space, and $\tilde{\ell}$ denote the likelihood function on this space. Then,

$$\tilde{\ell}(\tilde{y}') = \Pr[\hat{Y}{=}1 \mid \tilde{Y}'{=}\tilde{y}'] = \Pr[\hat{Y}{=}1 \mid Y'{=}y'] = \ell(y') = \tilde{y}',$$

so the transformation ensures that $\rho_{\tilde{\ell}}^* = 1$.

Connection to the α-Disparity Test. When Y' and Y are numerical, a natural extension of the α-disparity test (Definition 12) is

$$d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) \leq \rho_\ell^* \cdot \alpha \cdot d_{\text{em}}(Y|Z{=}0, Y|Z{=}1). \tag{8}$$

For this to work, Worldview 3 would have to change to use the earthmover distance rather than the total variation distance. Since the earthmover distance is defined over a distance metric, the parameter α is not very meaningful unless Y' and Y have the same scale. As a result, here we consider the case where Y' and Y are defined over the same metric space $(\mathcal{Y}, d)$.

Unfortunately, (8) is still not an empirical test because ρ_ℓ^* is defined in terms of Y'. As tempting as redefining ρ_ℓ^* in terms of Y is, Y' and Y can have vastly different likelihood functions despite having the same disparity, so this new empirical test will not guarantee the lack of disparity amplification under Definition 13. We leave as future work the discovery of an empirical test for numerical Y' and Y that corresponds to the α-Hybrid worldview.

10 CONCLUSION

We showed that demographic parity and equalized odds are related through our construct-based discrimination criterion of disparity amplification, arguing that the difference between the two empirical tests boils down to one's worldview. In addition, we proved that calibration is not robust to post-processing and that predictive parity allows a model with an arbitrarily large output disparity regardless of the worldview and the observed base rates.

Our work differs from much of the prior work in that we consider the construct as separate from the observed data. In particular, we interpreted the existing fairness definitions as acting on the observed data, whereas the discrimination criterion was viewed as a property of the construct. This bifurcation allowed us to handle the following issues simultaneously: (a) prohibitions against disparate impact have exceptions such as a business necessity, but (b) due to past discrimination, the observed data can be biased in an unjustified way. It is the second of these points that motivates our use of worldviews to characterize how biased the observed data is.

To illustrate how this might work in practice, let us revisit the examples in Section 3. In Example 1, there are reasons to believe that the observed recidivism rate is a racially biased measurement of the actual reoffense rate. In Example 2, for various socioeconomic reasons, some protected groups may have disproportionately many people who take longer than six years to graduate but are eventually considered successful in the university. The α-Hybrid worldview can characterize these real-world scenarios, and the value of α reflects one's beliefs about how much more biased the observed data is than the construct. Then, a practitioner can apply the α-disparity test as a substitute for demographic parity or equalized odds, with the value of α determined through social research and public dialogue.

REFERENCES

[1] Julia Angwin and Jeff Larson. 2016. ProPublica responds to company's critique of machine bias story. *ProPublica* (2016).

[2] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *California Law Review* 104 (2016), 671–732.

[3] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *ACM Conference on Fairness, Accountability, and Transparency*. 514–524.

[4] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *IEEE International Conference on Data Mining Workshops*. 13–18.

[5] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292.

[6] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163.

[7] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv* 1808.00023 (2018).

[8] Richard B Darlington. 1971. Another Look at "Cultural Fairness". *Journal of Educational Measurement* 8, 2 (1971), 71–82.

[9] Anupam Datta, Matt Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. 2017. Proxy Discrimination in Data-Driven Systems. *arXiv* 1707.08120 (2017).

[10] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science*. 214–226.

[12] Equal Employment Opportunities Commission. 1978. Uniform Guidelines on Employee Selection Procedures. 29 CFR Part 1607.

[13] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 259–268.

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

[15] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *ACM Conference on Fairness, Accountability, and Transparency*. 329–338.

[16] Alison L Gibbs and Francis Edward Su. 2002. On choosing and bounding probability metrics. *International Statistical Review* 70, 3 (2002), 419–435.

[17] Ben Green and Lily Hu. 2018. The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning. Presented at the Machine Learning: The Debates workshop at the 35th International Conference on Machine Learning.

[18] Susan S Grover. 1995. The business necessity defense in disparate impact discrimination cases. *Georgia Law Review* 30 (1995), 387–430.

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

[20] 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 *ACM Conference on Fairness, Accountability, and Transparency*. 181–190.

[21] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 35–50.

[22] Sampath Kannan, Aaron Roth, and Juba Ziani. 2019. Downstream effects of affirmative action. In *ACM Conference on Fairness, Accountability, and Transparency*. 240–248.

[23] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Innovations in Theoretical Computer Science*. 43:1–43:23.

[24] Torgny Lindvall. 2002. *Lectures on the coupling method*. Dover Publications.

[25] Adam Liptak. 2017. Sent to Prison by a Software Program's Secret Algorithms. *The New York Times* (2017).

[26] Lydia Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed Impact of Fair Machine Learning. In *International Conference on Machine Learning*. 3156–3164.

[27] Benjamin Mueller. 2018. Using Data to Make Sense of a Racial Disparity in NYC Marijuana Arrests. *The New York Times* (2018).

[28] Arvind Narayanan. 2018. Translation Tutorial: 21 Fairness Definitions and their Politics. Tutorial at the first Conference on Fairness, Accountability, and Transparency. Abstract available at https://facctconference.org/static/tutorials/narayanan-21defs18.pdf. Recording available at https://www.youtube.com/watch?v=jIXIuYdnyyk.

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

[30] John E Roemer. 2002. Equality of opportunity: A progress report. *Social Choice and Welfare* 19, 2 (2002), 455–471.

[31] Supreme Court of the United States. 2009. *Ricci v. DeStefano*. 557 U.S. 557.

[32] Supreme Court of Wisconsin. 2016. *State v. Loomis*. 881 N.W.2d 749.

[33] Cédric Villani. 2008. *Optimal transport: old and new*. Grundlehren der mathematischen Wissenschaften: Comprehensive Studies in Mathematics, Vol. 338. Springer-Verlag Berlin Heidelberg.

[34] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International Conference on World Wide Web*. 1171–1180.

[35] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Artificial Intelligence and Statistics*. 962–970.

[36] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–333.