

On the Generalizability of Fair Representations

Tosca Lechner¹, Shai Ben-David^{*2,1}, Ruth Urner^{*3,1}

¹Vector Institute, Toronto, Canada

²David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Canada

³Lassonde School of Engineering, EECs Department, York University, Toronto, Canada
tosca.lechner@vectorinstitute.ai, shai@uwaterloo.ca, ruth@eecs.yorku.ca

Abstract

The concept of a fair representation was introduced more than a decade ago and has inspired many follow-up works with the promising idea of a transferable module that guarantees fairness for many downstream tasks. However, it has been shown that this promise of guaranteed fairness is impossible in its proposed generality. That is, impossible without further restricting how and for which tasks the representation is being used. In this work, we consider a more attainable case in which classifiers are restricted, either structurally or through meeting certain accuracy-requirements (or both). We analyze whether the above impossibility results then still hold and how they could be overcome.

1 Introduction

As machine learning deployment has become ubiquitous, resulting automated decision rules are expected to comply with societal requirements, such as equitable treatment of protected sub-populations. The consideration of fairness for machine-learned classifiers was introduced in a seminal study by Dwork et al. over a decade ago (Dwork et al. 2012). It has since grown into a substantial sub-field of machine learning research. One major line of work in this area aims at providing fairness guarantees by mapping training data through a *fair representation* (Zemel et al. 2013). The idea of a fair representation (or fairness representation) is appealing: fully developed, it would provide a ready-to-use interface for users to pre-process their training data with the promise that any classification rule that would be trained on the so transformed data satisfies a chosen fairness requirement. The initial proposal has thus generated a flurry of follow-up studies developing various forms of fair representations, typically promoted with the suggestion the representation might provide fairness for multiple (if not all) downstream tasks.

However, it has been proven that the promise of guaranteed fairness for all, or even a somewhat diverse set of downstream tasks is too good to be true. A study published in 2022 showed that no useful representation can guarantee demographic parity or equalized odds fairness for all classification rules that are based on that representation (Lechner and Ben-David 2022). These results suggested that fair-

ness cannot simply be outsourced to some representation provider, but needs to be tested and evaluated on a case by case basis for each trained classifier. However, they have not diminished the general appeal of fair representations as a potentially easy to use tool and new variations of these representations continue to be developed (Sinigaglia et al. 2026; Luo et al. 2026).

In this work, we re-examine the question of the usefulness of fair representations beyond the question of fairness guarantees for *every* downstream task *independent of the downstream use of the representation*. First, we shift our focus away from worst-case fairness towards *accuracy driven* fairness guarantees. We analyze whether, while maybe not necessarily able to provide fairness guarantees *for all* downstream classifiers, it is possible for a representation to guarantee fairness for classifiers derived through the representation by optimizing for accuracy. This is arguably the type of classifier a user would most likely care about. Unfortunately, the conclusions of our analysis are rather pessimistic: fairness even of just the most accurate classifiers cannot always be guaranteed.

Our second focus is therefore to develop statistical tests that aim to certify whether a representation is suitable to yield fairness for a given task. Since, as mentioned above, fairness via a representation cannot be guaranteed for all downstream tasks, such testing is an important middle layer to protect against potential harms caused by blindly employing a representation. We show both upper and lower bounds for the fairness-testability of representations, ultimately coming again to a rather negative conclusion: Testing the fairness of a representation for a given task is not less data-expensive than learning a classifier and then testing the fairness of the classifier.

1.1 Outline

We start our investigation by outlining and phrasing precisely, various properties, or “desiderata” that users and developers may expect from fair representations: *complete fairness* for all downstream tasks and all classifiers; *accuracy induced fairness*, that is fairness guarantees for accurate classifiers; *avoidance of classification degradation*, namely that the loss of the best possible classifiers does not increase too much; and *obfuscating group identity* (which is a criterion that many fair representations are optimized for). We

^{*}These authors contributed equally.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

then show some basic connections between these desiderata in line with the previously established impossibility results for fair representations.

Next we further investigate the notion of accuracy induced fairness for equalized odds. Our analysis shows that in cases where a representation can provide such guarantees, the guarantee held already before applying the representation. More specifically, we show that if both the groups and the label classes are balanced, then accuracy implies equalized odds fairness (and no mapping through a representation is required). In other cases, fairness cannot be guaranteed by any representation either.

We round up our discussion by considering the sample complexity hardness of testing whether a representation is suitable for a given task. Here we prove hardness results showing that there can be no sample size bound for such testing that depends only on the desired accuracy and confidence parameters (and holds uniformly for all tasks).

1.2 Related Work

Fair representations were introduced over a decade ago (Zemel et al. 2013). This initial work generated a substantial line of research in the area of fairness in machine learning, with myriad alternative optimization objectives to achieve fair representations being proposed (Madras et al. 2018; Creager et al. 2019; Kim and Cho 2020; Zhao et al. 2020; Oneto et al. 2020; Gitiaux and Rangwala 2021, 2022; Oh et al. 2022; Kim et al. 2022; Liu et al. 2022; Shui et al. 2022; Shen, Wong, and Kankanhalli 2023; Deka and Sutherland 2023; Kleindessner et al. 2023; Kairouz et al. 2022; Jang et al. 2024; Sinigaglia et al. 2026). It is common in this line of research to motivate the notion of a fair representation by alluding to the possibility of guaranteeing fairness for many, if not all, downstream tasks (Kairouz et al. 2022; Kleindessner et al. 2023; Gitiaux and Rangwala 2021; Madras et al. 2018).

However, it has been proven that guaranteeing fairness with respect to demographic parity or with respect to equalized odds (the two notions that fairness representations are typically developed for) for all, or even a somewhat rich class of downstream tasks is impossible (Lechner and Ben-David 2022). A recent study developed compatibility tests for a proposed fair representation (Jovanovic et al. 2023). However the representation in that work is finite (all original feature vectors get mapped to one of finitely many new feature vectors), and thus those tests do not apply to typical representations that map onto an infinite new feature space. To the best of our knowledge, the question of testability in the general case has not been considered before.

The trade-off between accuracy and fairness through fair representation learning is acknowledged in the design of most optimization criteria for fair representations (which typically consist of a part aimed at obfuscating group identity and a part aimed at maintaining predictive quality) (Madras et al. 2018; Creager et al. 2019; Pouget et al. 2024). This trade-off has also been analyzed and quantified in earlier works (McNamara, Ong, and Williamson 2019; Zhao and Gordon 2022; Zhao et al. 2022; Ni and Huo 2026). Recent years have seen several (meta-)studies critiquing the va-

lidity of promises made in works that introduce new fair representations and have called for new or more rigorous evaluation frameworks (Jones et al. 2025; Pouget et al. 2024; Cerrato et al. 2024).

2 Setup

The data is originally collected in some original feature space $\mathcal{X}$. We consider prediction tasks (classification or regression) for a target variable taking values in a set $\mathcal{Y}$. We focus our discussion on binary classification tasks, where $\mathcal{Y} = \{0, 1\}$. A *classifier* is a function $h : \mathcal{X} \rightarrow \mathcal{Y}$.

A *learning task* is defined by a tuple $(\mathcal{X}, \mathcal{Y}, P, \ell)$ where P is a data-generating probability distribution over $\mathcal{X} \times \mathcal{Y}$ and $\ell : \mathcal{Y}^{\mathcal{X}} \times \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ is a loss function. We consider the corresponding *true* or *expected loss* of classifier h with respect to a distribution P , defined by $\mathcal{L}_P(h) = \mathbb{E}_{(x,y) \sim P}[\ell(h, (x, y))]$. We will focus the $0 - 1$ loss, $\ell(h, x, y) = \mathbf{1}[h(x) \neq y]$, for which we get $L_P(h) = \mathbb{P}_{(x,y) \sim P}[h(x) \neq y]$. We sometimes denote a task by only the distribution D (the other components left implicit).

The *Bayes classifier* for a probability distribution D is the function that minimizes the loss with respect to D . For binary classification, it is defined by $h_D^*(x) = \mathbf{1}[D[y = 1|x] \geq 0.5]$. We use $\text{opt}_D = L_D(h_D^*)$ to denote the *Bayes Loss*.

A *data representation*, or simply *representation*, is a function $\rho : \mathcal{X} \rightarrow \mathcal{Z}$, where $\mathcal{Z}$ is a new feature space. For a representation ρ and distribution P over $\mathcal{X} \times \mathcal{Y}$, we will use P_ρ to denote the distribution over $\mathcal{Z} \times \mathcal{Y}$ induced by ρ . For a representation $\rho : \mathcal{X} \rightarrow \mathcal{Z}$, we call a classifier h a ρ -*expressible classifier* if there exists a function $f : \mathcal{Z} \rightarrow \mathcal{Y}$ such that $h = f \circ \rho$. We denote the class of all ρ -expressible classifiers by $\mathcal{F}_\rho := \{h \in \mathcal{Y}^{\mathcal{X}} : \exists f : \mathcal{Z} \rightarrow \mathcal{Y} \text{ with } h = f \circ \rho\}$.

Given a representation ρ , we denote the minimum-error of any classifier expressible by over it by $\text{opt}_P(\rho) = \min_{h \in \mathcal{F}_\rho} L_P(h)$. The classifier that achieves this minimal loss is the *Bayes classifier over the representation*, $h_{\rho, P}^* = \mathbf{1}[P[y = 1|\rho(x)] \geq 0.5]$.

Fairness We assume that the original domain $\mathcal{X}$ is partitioned into two groups $\mathcal{A}$ and $\mathcal{B}$ by means of one of the features, which we name g , in the original feature space $\mathcal{X}$. Slightly abusing notation, we will also use g as a *group membership function* $g : \mathcal{X} \rightarrow \{\mathcal{A}, \mathcal{B}\}$. We use the notation $\mathcal{X}_{-g}$ and x_{-g} to denote the features space $\mathcal{X}$ or a feature vector x with feature g removed. The goal of learning or designing a fair representation is typically formalized through a *group fairness* requirement such as *Demographic Parity (DP)* or *Equalized Odds (EO)*. We here define these requirements as well as the corresponding “measures of unfairness” following prior work (Madras et al. 2018; Lechner and Ben-David 2022).

Definition 1 (Demographic Parity). A classifier $h : \mathcal{X} \rightarrow \mathcal{Y} = \{0, 1\}$ satisfies *Demographic Parity (DP)* with respect to distribution P over $\mathcal{X} \times \{0, 1\}$ and group membership $g : \mathcal{X} \rightarrow \{\mathcal{A}, \mathcal{B}\}$ if

$$\begin{aligned} \mathbb{P}_{(x,y) \sim P}[h(x) = 1 \mid g(x) = \mathcal{A}] \\ = \mathbb{P}_{(x,y) \sim P}[h(x) = 1 \mid g(x) = \mathcal{B}] \end{aligned}$$

We define the *measure of unfairness with respect to DP* as

$$\mu_P^{\text{DP}}(h) = |\mathbb{P}_{(x,y) \sim P}[h(x) = 1 \mid g(x) = \mathcal{A}] - \mathbb{P}_{(x,y) \sim P}[h(x) = 1 \mid g(x) = \mathcal{B}]|$$

Definition 2 (Equalized Odds). A classifier $h : \mathcal{X} \rightarrow \mathcal{Y} = \{0, 1\}$ satisfies *Equalized Odds (EO)* with respect to distribution P over $\mathcal{X} \times \{0, 1\}$ and group membership $g : \mathcal{X} \rightarrow \{\mathcal{A}, \mathcal{B}\}$ if for both $l \in \{0, 1\}$

$$\begin{aligned} \mathbb{P}_{(x,y) \sim P}[h(x) = 1 - l \mid g(x) = \mathcal{A} \wedge y = l] \\ = \mathbb{P}_{(x,y) \sim P}[h(x) = 1 - l \mid g(x) = \mathcal{B} \wedge y = l] \end{aligned}$$

We define the *measure of unfairness with respect to EO* as

$$\begin{aligned} \mu_P^{\text{EO}}(h) = \frac{1}{2} |\text{FPR}_P^{\mathcal{A}}(h) - \text{FPR}_P^{\mathcal{B}}(h)| \\ + \frac{1}{2} |\text{FNR}_P^{\mathcal{A}}(h) - \text{FNR}_P^{\mathcal{B}}(h)| \end{aligned}$$

where $\text{FPR}_P^G = \mathbb{P}_{(x,y) \sim P}[h(x) = 1 \mid g(x) = G \wedge y = 0]$ and $\text{FNR}_P^G = \mathbb{P}_{(x,y) \sim P}[h(x) = 0 \mid g(x) = G \wedge y = 1]$ are the false positive and false negative rates for groups $G \in \{\mathcal{A}, \mathcal{B}\}$ respectively.

For a fairness notion F , we denote the worst-case fairness of a representation ρ with respect to task P by

$$\mu_P^F(\rho) := \sup_{h \in \mathcal{F}_\rho} \mu_P^F(h).$$

Note that whenever a classifier h is perfectly accurate with respect to a distribution P then it is also perfectly EO fair with respect to P . Namely, $\mathcal{L}_P(h) = 0$ implies that $\mu_P^{\text{EO}}(h) = 0$.

3 Desiderata for Fair Data Representations

The hope of developing a “fair data representation” is that some (trusted) data curator will apply a mapping ρ to the raw data in such a way that any ρ -expressible classification rule will automatically satisfy some fairness criteria with respect to groups $\mathcal{A}$ and $\mathcal{B}$.

We start by listing (and phrasing precisely) properties that one might wish a fair data representation to have. We will define these desiderata with respect to a class $\mathcal{P}$ of distributions over $\mathcal{X} \times \mathcal{Y}$.

Complete fairness Let F be a fairness objective (such as EO or DP). For every task (distribution) $P \in \mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$ every ρ -expressible classifier $h : \mathcal{X} \rightarrow \mathcal{Y}$ satisfies F .

(ϵ, α) -accuracy induced fairness Let F be a fairness objective, μ_P^F be the corresponding measure of unfairness and $\epsilon, \alpha > 0$. For every task (distribution) $P \in \mathcal{P}$ every ρ -expressible classifier $h : \mathcal{X} \rightarrow \mathcal{Y}$ with loss $\mathcal{L}_P(h) \leq \epsilon$ has unfairness bounded by $\mu_P^F(h) \leq \alpha$.

(ϵ, α) -Bayes-accuracy induced fairness Let F be a fairness objective, μ_P^F be the corresponding measure of unfairness and $\epsilon, \alpha > 0$. For every task (distribution) $P \in \mathcal{P}$ every ρ -expressible classifier $h : \mathcal{X} \rightarrow \mathcal{Y}$ with loss $\mathcal{L}_P(h) \leq \text{opt}_P(\rho) + \epsilon$ has unfairness bounded by $\mu_P^F(h) \leq \alpha$. This notion guarantees fairness for approximately accurate classifiers and thus relevant for classifiers obtained by accuracy optimization.

γ -bounded classification degrading For every task (distribution) $P \in \mathcal{P}$ and every $\epsilon > 0$, the existence of a classifier $h : \mathcal{X} \rightarrow \mathcal{Y}$ with error at most ϵ implies the existence of a ρ -expressible classifier $h' : \mathcal{X} \rightarrow \mathcal{Y}$ with error at most $\epsilon + \gamma$. In other words, the Bayes error over the ρ representation is higher than the Bayes error over the original data by at most γ .

α -obfuscating group identity Let $0 < \alpha < 0.5$. We say that a representation ρ is α -obfuscating group identity with respect to a probability distribution P if for all functions $h \in \mathcal{F}_\rho$ (that is, for all ρ -representable functions $h : \mathcal{X} \rightarrow \{\mathcal{A}, \mathcal{B}\}$),

$$\begin{aligned} |\mathbb{P}_{x \sim P_X}[h(x) = \mathcal{A} \mid x \in \mathcal{A}] \\ - \mathbb{P}_{x \sim P_X}[h(x) = \mathcal{A} \mid x \in \mathcal{B}]| \leq \alpha \end{aligned}$$

A representation ρ is α -obfuscating with respect to a family of distributions $\mathcal{P}$, if it is α -obfuscating with respect to every P in $\mathcal{P}$.

3.1 Connections and Implications of the Desiderata

Complete fairness The first property in the above list, *complete fairness*, is what many fairness representations are aimed at. However, it has been proven in prior work that guaranteeing fairness for all tasks is impossible, with respect to any of the DP and EO notions, unless the representation is *trivial*.

Definition 3. We call a representation $\rho : \mathcal{X} \rightarrow \mathcal{Z}$ *trivial* if it maps all original feature vectors to the same point. That is, there is a $z \in \mathcal{Z}$ such that $\rho(x) = z$ for all $x \in \mathcal{X}$.

Observation 1. If a representation ρ is trivial it guarantees fairness (both with respect to DP and with respect to EO) for all downstream tasks.

Of course, a trivial representation is not useful for classification, since all ρ -expressible classifiers are constant functions. However, prior work has shown that trivial representations are actually *the only* representations that can guarantee complete DP fairness.

Theorem 2 (Slight strengthening of Theorem 1 in (Lechner and Ben-David 2022)). *Let $\mathcal{X}$ be a domain and $g : \mathcal{X} \rightarrow \{\mathcal{A}, \mathcal{B}\}$ a partition of $\mathcal{X}$ into two non-empty groups. Then for every non-trivial representation $\rho : \mathcal{X} \rightarrow \mathcal{Z}$, there exists a distribution P over $\mathcal{X} \times \mathcal{Y}$ such that the ρ -expressible Bayes classifier is maximally unfair with respect to demographic parity, $\mu_P^{\text{DP}}(h_{\rho,P}^*) = 1$.*

Proof. Note that for every non-trivial representation $\rho : \mathcal{X} \rightarrow \mathcal{Z}$, there exist points x_1 with $g(x_1) = \mathcal{A}$ and x_2 with $g(x_2) = \mathcal{B}$ such that $\rho(x_1) \neq \rho(x_2)$. Pick x_1, x_2 satisfying the above and define a distribution P over $\mathcal{X} \times \{0, 1\}$ by $P(x_1, 0) = P(x_2, 1) = 0.5$. Let h be a classifier that assigns the label 0 to every instance in $\{x : \rho(x) = \rho(x_1)\}$ and assigns the label 1 to every instance in $\{x : \rho(x) = \rho(x_2)\}$. Note that since $\rho(x_1) \neq \rho(x_2)$, h is ρ -expressible. Further, since $\mathcal{L}_P(h) = 0$, h is Bayes optimal for P while also being maximally unfair with respect to demographic parity: $\mu_P^{\text{DP}}(h) = 1$. $\square$

A similar impossibility has also been established for complete EO fairness.

Theorem 3 (Theorem 2 (Part 1) in (Lechner and Ben-David 2022)). *Let $\mathcal{X}$ be a domain and $g : \mathcal{X} \rightarrow \{\mathcal{A}, \mathcal{B}\}$ a partition of $\mathcal{X}$ into two groups. Then for every non-trivial representation $\rho : \mathcal{X} \rightarrow \mathcal{Z}$ that allows ρ -expressible classifiers that are not constant on both groups, there exists a distribution P over $\mathcal{X} \times \mathcal{Y}$ such that the set of ρ expressible classifiers contains functions h with $\mu_P^{\text{EO}}(h) = .5$.*

Accuracy induced fairness In light of these impossibilities for complete fairness, we will turn our focus in this work on *accuracy induced fairness*, and we will mostly focus on the EO fairness notion. The hope, as expressed in the property of (ϵ, α) -accuracy induced fairness, is that after applying the representation, every predictor that has a classification error of at most ϵ automatically has a bound, α , on its unfairness. Thus, if the representation is used, and a classifier of high accuracy is derived, fairness would also be (approximately) satisfied. We will explore this notion in more detail in the next section.

γ -bounded classification degradation The last two items on our list of desirable properties reflect how fair representations are often trained. Typically, works on fair representation introduce an optimization criterion consisting of two competing parts. One objective is to obfuscate group identity (as in our last property) and the other part is to maintain the attainability of highly accurate classification rules. We formalize this latter criterion with our notion of a representation being *γ -bounded classification degrading* (see list in the previous section for the definition). It requires that the Bayes loss does not increase (too much) through the application of the representation. If a task allowed highly accurate classifiers in the original representation, it is desirable to maintain that these exist. This requirement is immediately satisfied, if the representation ρ is redundant in the following sense:

Definition 4. We call a representation ρ *redundant* if the set $\mathcal{F}_\rho$ of all ρ -expressible classifiers from $\mathcal{X}$ to $\mathcal{Y}$ is identical to the set $\mathcal{Y}^{\mathcal{X}}$ of all possible classifiers from $\mathcal{X}$ to $\mathcal{Y}$.

It is easy to see that a representation $\rho : \mathcal{X} \rightarrow \mathcal{Y}$ is redundant if and only if it is an invertible (one-to-one) function.

Observation 4. *If a representation ρ is redundant, then it is γ -bounded classification degrading for any $\gamma > 0$ with respect to any set $\mathcal{P}$ of distributions over $\mathcal{X} \times \mathcal{Y}$.*

Conversely, if a representation is not redundant, then it will maximally increase the Bayes loss on some tasks.

Claim 5. *Let $\mathcal{X}$ be a domain set of size ≥ 2 . For every non-redundant data representation ρ over X , there is a task (namely, a probability distribution P over $X \times \{0, 1\}$) that is perfectly classifiable under the original data representation, but every ρ -expressible classifier has high error.*

Proof. If ρ is not redundant then there exist $x_1 \neq x_2 \in \mathcal{X}$ so that $\rho(x_1) = \rho(x_2)$. It follows that every ρ -based classifier h satisfies $h(x_1) = h(x_2)$. Let P be the distribution (over $X \times \{0, 1\}$) defined by $P(x_1, 1) = P(x_2, 0) = 0.5$. Note that over the original data $\mathcal{X}$, there exists a zero-error classifier

for P (any function $f : X \rightarrow \{0, 1\}$ for which $f(x_1) = 1$ and $f(x_2) = 0$). However, every ρ -based classifier h has loss $\mathcal{L}_P(h) = 0.5$. $\square$

Of course, despite the above observations, redundancy is not a desirable property for a fairness representation, since mapping through a redundant representation retains all the same unfair classification rules that the original representation can express. In other words, a redundant representation does not provide any advantages in terms of fairness objectives.

Moreover, it has been shown that no representation can simultaneously prevent γ -bounded classification degradation for arbitrarily small γ and provide complete EO-fairness for several tasks. More specifically,

Theorem 6 (A consequence of Theorem 3 of (Lechner and Ben-David 2022)). *Let $\mathcal{P}$ be a class of distributions that contains at least two different distributions P_1 and P_2 with Bayes loss 0 and at least one of the two having different success rates between the two groups. Then no representation $\rho : \mathcal{X} \rightarrow \mathcal{Z}$ can guarantee γ -bounded classification degrading for arbitrarily small $\gamma > 0$ and satisfy complete EO-fairness with respect to class $\mathcal{P}$.*

If a representation ρ guarantees complete fairness, even for very small and simple classes of distributions, then the representation inevitably leads to classification degradation. Not all tasks that were perfectly classifiable over the original feature space $\mathcal{X}$ will retain a perfect classifier after representation ρ is applied.

α -obfuscating group identity Lastly, we note that our last property on the list, *α -obfuscating group identity*, is often incorporated in some form to optimization criteria that aim for DP fairness.

Formally, this it can be justified by the following:

Observation 7. *A representation is α -obfuscating group identity with respect to $\mathcal{P}$ if and only if, every ρ -expressible classifier $h \in \mathcal{F}_\rho$ satisfies $\mu_P^{DP}(h) \leq \alpha$.*

Proof. Let $h : \mathcal{X} \rightarrow \{0, 1\}$. Define a function $\bar{h} : \mathcal{X} \rightarrow \{\mathcal{A}, \mathcal{B}\}$ by having $\bar{h}(x) = \mathcal{A}$ if and only if $h(x) = 1$. Note that h is ρ -expressible, if and only if so is $\bar{h}$. Note that by definition we get $\mu_P^{DP}(h) > \alpha$ if and only if

$$|\mathbb{P}_{\mathbf{x} \sim P_X}[\bar{h}(x) = \mathcal{A} | x \in \mathcal{A}] - \mathbb{P}_{\mathbf{x} \sim P_X}[\bar{h}(x) = \mathcal{A} | x \in \mathcal{B}]| > \alpha.$$

This implies the claim in the observation. $\square$

4 Accuracy Induced EO Fairness

The previous section showcased various examples on how accuracy, accuracy degradation and fairness might interact when employing a data representation. In this section we provide a more general look at these relations. We start by showing that for tasks where the groups have similar sizes, and for each of the groups the two label classes are relatively balanced, the EO unfairness can be upper bounded by the

error of a classifier. That is, in such cases training a highly accurate classifier will yield high EO fairness “for free”.

The bounds we establish in this section are based on the quantity $\min_{g,l}(P)$, that corresponds to the smallest non-zero mass of a group-label-subset for the task P :

$$\min_{g,l}(P) := \min_{\substack{l \in \{0,1\} \\ G \in \{\mathcal{A}, \mathcal{B}\}}} \mathbb{P}_{(x,y) \sim P}[g(x) = G, y = l],$$

where we use the notation $\min^+$ to denote the minimal non-zero quantity.

We start by observing an upper bound for the unfairness of a classifier in terms of its inaccuracy:

Theorem 8. *For every classifier $h \in \mathcal{Y}^{\mathcal{X}}$ we get:*

$$\mu_P^{\text{EO}}(h) \leq \min \left\{ \frac{\mathcal{L}_P(h)}{\min_{g,l}(P)}, 1 \right\}.$$

Proof. Let $h : \mathcal{X} \rightarrow \mathcal{Y}$ be an arbitrary classifier. Then:

$$\begin{aligned} \mu_P^{\text{EO}} &= \sum_{l \in \{0,1\}} |\mathbb{P}_{(x,y) \sim P}[h(x) = 1 - l | g(x) = \mathcal{A}, y = l] \\ &\quad - \mathbb{P}_{(x,y) \sim P}[h(x) = 1 - l | g(x) = \mathcal{B}, y = l]| \\ &\leq \sum_{\substack{l \in \{0,1\} \\ G \in \{\mathcal{A}, \mathcal{B}\}}} \mathbb{P}_{(x,y) \sim P}[h(x) = 1 - l | g(x) = G, y = l] \\ &= \sum_{\substack{l \in \{0,1\} \\ G \in \{\mathcal{A}, \mathcal{B}\}}} E_{G,l} = \frac{1}{\min_{g,l}(P)} \sum_{\substack{l \in \{0,1\} \\ G \in \{\mathcal{A}, \mathcal{B}\}}} \min_{g,l}(P) \cdot E_{G,l} \\ &\leq \frac{\mathcal{L}_P(h)}{\min_{g,l}(P)} \end{aligned}$$

where we use the notation

$$E_{G,l} = \mathbb{P}_{(x,y) \sim P}[h(x) = 1 - l | g(x) = G, y = l].$$

□

Thus, we get the following result on accuracy-induced fairness of a representation:

Corollary 9 (Implied (ϵ, α) -accuracy-induced fairness). *For every $\epsilon > 0$ and $\alpha > 0$, every representation ρ is (ϵ, α) -accuracy-induced fair w.r.t. the class of distributions P for which $\min_{g,l}(P) \geq \frac{\epsilon}{\alpha}$.*

The above statement shows that if the groups and label proportions are reasonably balanced in some task, then a certain level of accuracy-induced fairness is satisfied by any representation. For example, if $\min_{g,l}(P) = 0.2$, then the requirement of being $(\epsilon, 5\epsilon)$ -accuracy-induced-fair is always satisfied.

On the other hand, we will now show that balanced group and label classes sizes are not in itself sufficient to obtain *meaningful* success for accuracy induced fairness. The guarantee of Corollary 9 becomes vacuous if the Bayes loss becomes too high. High Bayes loss can be induced by the data representation, or can be due to the task having a high Bayes loss already over the original feature space $\mathcal{X}$.

Our next claim formalizes this by showing that no non-redundant representation can prevent the existence of highly unfair classifiers on some tasks.

Claim 10. *If a representation $\rho : \mathcal{X} \rightarrow \mathcal{Z}$ is neither redundant nor trivial, then there is a task (distribution over $\mathcal{X} \times \mathcal{Y}$) on which highly unfair ρ -representable classifiers exist, even when the two groups, and the two label classes within them, are balanced. Moreover, these highly unfair classifiers can be Bayes optimal among all ρ -representable classifiers.*

Proof. If a representation ρ is not redundant, then there exist at least two points x_1 and x_2 with $\rho(x_1) = \rho(x_2)$. If ρ is furthermore not trivial, then there exists at least one additional point $x_3 \in \mathcal{X}$ with $\rho(x_3) \neq \rho(x_1) = \rho(x_2)$.

Case 1: There exist x_1, x_2 with $\rho(x_1) = \rho(x_2)$ and $g(x_1) \neq g(x_2)$. Without loss of generality, let's assume that $g(x_1) = g(x_3) = \mathcal{A}$ and $g(x_2) = \mathcal{B}$. Now consider the distribution P over $\mathcal{X} \times \mathcal{Y}$ with $P(x_1, 0) = 0.25$, $P(x_3, 1) = 0.25$, $P(x_2, 0) = 0.25$, $P(x_2, 1) = 0.25$. Thus, in both groups, both labels are observed with probability 0.5 and the groups also have equal weight. Consider the ρ -representable classifier h with lowest classification error. We have $h(z_1) = h(\rho(x_1)) = h(\rho(x_2)) = 0$ and $h(z_2) = h(\rho(x_3)) = 1$. This classifier has false negative rate 1 for group $\mathcal{B}$ and false negative rate is 0 for group $\mathcal{A}$ (while the false positive rate is 0 for both groups). Thus $\mu_P^{\text{EO}}(h) = 0.5$. Note that both groups and label sets are balanced in the distribution P .

Case 2: For any two points x_1, x_2 with $g(x_1) \neq g(x_2)$ we have $\rho(x_1) \neq \rho(x_2)$. Let two points x_1, x_2 with $g(x_1) = \mathcal{A}$ and $g(x_2) = \mathcal{B}$ be given. Since ρ is not redundant, we can choose x_1 and x_2 so that there is a third point x_3 that gets the same representation as x_1 or x_2 . Without loss of generality we will assume $\rho(x_3) = \rho(x_1)$, and thus $g(x_3) = \mathcal{A}$. Now consider the distribution P with $P(x_1, 1) = 0.25$, $P(x_3, 0) = 0.25$, $P(x_2, 1) = P(x_2, 0) = 0.25$. Now the classifier h with $h(z_1) = h(\rho(x_1)) = h(\rho(x_3)) = 0$ and $h(z_2) = h(\rho(x_2)) = 1$, is Bayes optimal among ρ -representable classifiers. It has false negative rate 1 on group $\mathcal{A}$ and false negative rate 0 on group $\mathcal{B}$. Moreover it has false positive rate 1 on group $\mathcal{B}$ and false positive rate 1 on group $\mathcal{A}$. Thus $\mu_P^{\text{EO}}(h) = 0.5$. Note that both groups and label sets are balanced in the distribution P . □

A closer inspection of the example in the second case above reveals that for that specific task, every representation under which the minimal error classifier (i.e., the Bayes classifier) is fair, is accuracy-degrading.

Claim 11. *There is a task (distribution over $\mathcal{X} \times \mathcal{Y}$) such that for every representation $\rho : \mathcal{X} \rightarrow \mathcal{Z}$, that does not degrade the best possible accuracy by at least 0.25, no ρ -based classifier simultaneously achieves best possible fairness and best possible accuracy.*

Proof. Consider a task P defined on three points $x_1, x_2, x_3 \in \mathcal{X}$ with $g(x_1) = g(x_3) = \mathcal{A}$ and $g(x_2) = \mathcal{B}$, and $P(x_1, 1) = 0.25$, $P(x_3, 0) = 0.25$, $P(x_2, 1) = P(x_2, 0) = 0.25$. Now any classifier will either have false positive rate 1 or false negative rate 1 in group $\mathcal{B}$ (depending on which label point x_2 gets mapped to). To achieve optimal fairness, a classifier will need to match the false positive (or false negative) rate on group $\mathcal{A}$ and thus map both points x_1

and x_2 to the same label. Any such fairness-optimal classifier will therefore suffer loss 0.5 even though the Bayes optimal loss for this distribution is 0.25. Any Bayes optimal classifier h^* in the original representation is highly unfair with $\mu_P^{\text{EO}}(h^*) = 0.5$.

In this case, the only way for a representation ρ to ensure that the ρ -representable Bayes optimal is also optimally fair, is to either map $\rho(x_1) = \rho(x_2)$ or to map $\rho(x_1) = \rho(x_3)$. In both cases the ρ -based Bayes loss is 0.5 and thus the representation is not γ -classification degrading for any $\gamma < 0.25$. $\square$

The examples in the proofs of the above two claims show that a task having initially high Bayes loss or suffering classification degradation through the representation can render the “free fairness” guarantees of Theorem 8 and Corollary 9 vacuous. We will next quantify this relationship more carefully. We establish the following lower bound on the unfairness of the Bayes classifier over a given representation, in terms of its (in)accuracy:

Theorem 12. *For every representation ρ , we have the following accuracy-fairness trade-off for the ρ -induced Bayes classifier $h_{P,\rho}^*$ in terms of the unfairness of the original Bayes classifier h_P^* :*

$$\begin{aligned}\mu_P^{\text{EO}}(h_{P,\rho}^*) &\geq \mu_P^{\text{EO}}(h_P^*) - \frac{\mathcal{L}_P(h_{P,\rho}^*) + \mathcal{L}_P(h_P^*)}{\min_{g,l}(P)} \\ &= \mu_P^{\text{EO}}(h_P^*) - \frac{\Delta_P^\rho + 2\mathcal{L}_P(h_P^*)}{\min_{g,l}(P)}\end{aligned}$$

where $\Delta_P^\rho = \mathcal{L}_P(h_{P,\rho}^*) - \mathcal{L}_P(h_P^*)$ is the amount of classification degradation of the representation ρ .

The proof follows from application of Theorem 8 and can be found in the appendix.

Finally, we show that if the groups are not balanced, then even highly accurate classifiers can be very unfair (and their unfairness is due to mistreatment of the minority group in this case). For a distribution P over $\mathcal{X} \times \mathcal{Y}$, $G \in \{\mathcal{A}, \mathcal{B}\}$, and $l \in \{0, 1\}$, let P_G be the conditional distribution of $(x, y) \sim P$ conditioned on $g(x) = G$ and let $P_{G,l}$ be the conditional distribution of $(x, y) \sim P$ conditioned on $g(x) = G \wedge y = l$. Furthermore, for $G \in \{\mathcal{A}, \mathcal{B}\}$ and $l \in \{0, 1\}$, let $\bar{P}_{G,l}$ be the conditional distribution for $(x, y) \sim P$ conditioned on $g(x) \neq G \vee y \neq l$. Lastly, for $G \in \{\mathcal{A}, \mathcal{B}\}$ let $\bar{G} \in \{\mathcal{A}, \mathcal{B}\}$ be defined by $\bar{G} \neq G$ and for $l \in \{0, 1\}$ let $\bar{l} = 1 - l$.

Claim 13. *Let P be a task, such that there is a subgroup $G \in \{\mathcal{A}, \mathcal{B}\}$, $l \in \{0, 1\}$ with $P(\{(x, y) \in \mathcal{X} \times \mathcal{Y} : g(x) = G, y = l\}) = \min_{g,l}(P) < \epsilon$ and $d_{\text{TV}}(P_{G,l}, \bar{P}_{G,l}) = 1$. Then for any classifier $h_1 : \mathcal{X} \rightarrow \mathcal{Y}$ with $\mathcal{L}_P(h_1) \leq \gamma$, there exists also a classifier $h_2 : \mathcal{X} \rightarrow \mathcal{Y}$ with*

- $\mathcal{L}_P(h_2) \leq \gamma + \epsilon$
- $\mu_P^{\text{EO}}(h_2) \geq \frac{1}{4}$.

Proof. From $d_{\text{TV}}(P_{G,l}, \bar{P}_{G,l}) = 1$ we know, that there is a set $C \subset \mathcal{X}$ with $P_{G,l}(C) = 1$ and $\bar{P}_{G,l}(C) = 0$. Define $h_{C,y} : \mathcal{X} \rightarrow \mathcal{Y}$ by

$$h_{C,y}(x) = \begin{cases} h_1(x) & \text{if } x \in \mathcal{X} \setminus C \\ y & \text{if } x \in C. \end{cases}$$

Now consider the behaviour of h_1 for the group-label pair $(\bar{G}, l)$, i.e., consider the classification rate $r = \mathbb{P}_{(x,y) \sim P}[h_1(x) = \bar{l} | g(x) = \bar{G}, y = l]$. If $r > \frac{1}{2}$, define $h_2 = h_{C,l}$, otherwise define $h_2 = h_{C,\bar{l}}$. The classifier h_2 will behave exactly as h_1 outside of C , thus since $P(C) = P(\{(x, y) \in \mathcal{X} \times \mathcal{Y} : g(x) = G, y = l\}) = \min_{g,l}(P) < \epsilon$, we know $|\mathcal{L}_P(h_1) - \mathcal{L}_P(h_2)| \leq \epsilon$. Furthermore, we know that

$$\begin{aligned}\mu_P^{\text{EO}}(h_2) &\geq \frac{|\mathbb{P}_{(x,y) \sim P}[h_1(x) = \bar{l} | g(x) = \bar{G}, y = l] - r|}{2} \\ &\geq \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.\end{aligned}$$

This proves the claim. $\square$

With the same construction, one can also start with a (potentially Bayes optimal) classifier h_1 with $\mathcal{L}_P(h_1) \leq \gamma$ and change the classification rate r of the minority group to match the classification rate r of the other group $\bar{G}$. The resulting classifier in this setting would also satisfy $\mathcal{L}_P(h_3) \leq \gamma + \epsilon$ and either have fair false negative rates or fair false positive rates, potentially significantly improving fairness. Thus restricting the classifiers that are being used can have a large impact on fairness even if the changes in accuracy are small.

Summarizing discussion of this section We have shown in this section that a representation can provide EO fairness guarantees, however *only* under very specific conditions. Theorem 8 with Corollary 9 show that if the group sizes are balanced, and if the label classes in both groups are balanced, and a classifier has very small classification loss, then its EO unfairness is also bounded (note that such a high-accuracy classifier can only be found if the task did not suffer classification degradation through the representation ρ). The other results in this section, namely Claims 10, 11, and 13, and Theorem 12, show that as soon as one of the conditions listed above is not satisfied, a meaningful fairness guarantee is impossible.

In other words, accuracy induced fairness *improvements* through the application of a representation are only possible if at least one of the following holds:

- the Bayes loss of the original representation is high
- applying the representation ρ leads to significant classification degradation
- (at least) one of the group sizes is small

We note that such “fairness improvements” should rarely be desirable. If the Bayes loss is high already over the original feature representation, then (unless we are in the unlikely situation that accuracy is not important), the correct recourse should be to improve the data-collection pipeline. In such a situation we should choose features that are more predictive for our task and/or improve the measurements/collection of these features. Sending the (low quality) data through a representation can only lead to even more classification loss.

Similarly, if the representation leads to classification degradation, then applying such a representation to obtain fairness improvements, can only be justified if fairness is more important than accuracy in the given application.

While there are situations where this is the case, especially when the main downstream effect of a task concerns the distribution of resources, in other settings in which individuals mainly care about being classified accurately (such as medical diagnosis), unfairness should be mitigated by *improving* accuracy for the disadvantaged group, rather than by achieving fairness through worsening the accuracy for all groups.

The last case, namely improving fairness in the case one of the groups consists of a small minority, warrants a similar warning. If unfairness is largely due to mistreatment of a minority, a representation might “fix” the unfairness by lowering accuracy for the majority group. This will indeed yield a better score in terms of EO fairness. However, as before, we would argue that the correct way of improving fairness in such a situation, would be to *improve* the treatment of the minority group. This would require spending resources, namely, collecting more data or investing in better feature measurements for the minority group, so that the accuracy in that group can be lifted to the level of the majority.

5 Accuracy Induced EO Fairness for Multiple Tasks

One of the ideas of fair representations is the re-use of the representation for many tasks. However, a key result of Lechner and Ben-David (2022) is that for equalized odds fairness, that as soon as one considers tasks in which the ground-truth labels differ between groups *no two tasks with the same marginal are EO fairness compatible*: That is, no representation can simultaneously guarantee fairness of all classifiers expressed over the representation for both tasks, while also allowing for perfect accuracy for both tasks. We now consider the accuracy-driven fairness of a representation, rather than its worst case fairness. This is reasonable in situations in which a downstream user can be trusted to optimize for accuracy, rather than trying to discriminate adversarially.

We find that even in the case of accuracy-driven fairness, there are tasks, for which there cannot be a joint representation that allows for accurate classification, while still *guaranteeing fairness for all approximately accurate classifiers*. Thus, *any representation* that would satisfy fairness and accuracy-requirements of one task, would hurt either fairness or accuracy in the other task.

Lemma 14. *Let $\mathcal{X}$ be a finite domain set and $\eta > 0$. For every $0 < \epsilon < \frac{\eta - \eta^2}{2} < 0.5$ there exist two distributions P_1 and P_2 over $\mathcal{X} \times \{0, 1\}$ with equal marginals $P_{1,\mathcal{X}} = P_{2,\mathcal{X}}$, such that every function $h : \mathcal{X} \rightarrow \{0, 1\}$ that is approximately accurate with respect to P_1 , i.e., $\mathcal{L}_{P_1}(h) \leq \epsilon$ satisfies*

1. *h is also approximately accurate with respect to P_2 , i.e., $\mathcal{L}_{P_2}(h) \leq \epsilon + \eta$.*
2. *h is very EO-unfair with respect to P_2 , i.e., $\mu_{P_2}^{\text{EO}}(h) \geq \frac{1}{4}$.*

Proof. We start by defining $P_{\mathcal{X}} = P_{1,\mathcal{X}} = P_{2,\mathcal{X}}$ be a probability distribution over $\mathcal{X}$ with $P_{\mathcal{X}}(\{x \in \mathcal{X} : g(x) = \mathcal{A}\}) = 1 - \eta$ and $P_{\mathcal{X}}(\{x \in \mathcal{X} : g(x) = \mathcal{B}\}) = \eta$.

Let $f_1, f_2 : \mathcal{X} \rightarrow \{0, 1\}$ be any two functions that agree on group $\mathcal{A}$ and contradict each other on group $\mathcal{B}$. Namely,

for every $x \in \mathcal{X}$ with $g(x) \in \mathcal{A}$, $f_1(x) = f_2(x)$ and for every $x \in \mathcal{X}$ with $g(x) \in \mathcal{B}$, we have $f_1(x) = 1 - f_2(x)$.

We now define P_1 and P_2 as distributions over $\mathcal{X} \times \mathcal{Y}$ such that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we have $P_1((x, y)) = P_{\mathcal{X}}(x) \cdot \mathbf{1}[y = f_1(x)]$ and $P_2((x, y)) = P_{\mathcal{X}}(x) \cdot \mathbf{1}[y = f_2(x)]$.

Note that since $P_{\mathcal{X}}(\{x : f_1(x) \neq f_2(x)\}) = \eta$, for every function $h : \mathcal{X} \rightarrow \{0, 1\}$, $|\mathcal{L}_{(P_1)}(h) - \mathcal{L}_{(P_2)}(h)| \leq \eta$. This implies the first claim of the lemma.

For the second claim of the lemma, note that if $\mathcal{L}_{P_1}(h) \leq \epsilon$ then $P(\{x : f_1(x) = h(x) \wedge g(x) = \mathcal{B}\}) \geq \eta - \epsilon$. Note also that $\{x : f_1(x) = h(x) \cap g(x) = \mathcal{B}\} = \{x \in \mathcal{X} : h(x) \neq f_2(x) \wedge g(x) = \mathcal{B}\}$.

It follows that for every such h ,

$$\mathbb{P}_{(x,y) \sim P_2}[h(x) \neq f_2(x) | g(x) = \mathcal{B}] \geq \frac{\eta - \epsilon}{\eta}.$$

At the same time, for such h (satisfying $\mathcal{L}_{P_1}(h) \leq \epsilon$) we get

$$\begin{aligned} \mathbb{P}_{(x,y) \sim P_2}[h(x) \neq f_2(x) | g(x) = \mathcal{A}] \\ = \mathbb{P}_{(x,y) \sim P_2}[h(x) \neq f_1(x) | g(x) = \mathcal{A}] \leq \frac{\epsilon}{1 - \eta} \end{aligned}$$

Thus, we can lower bound the EO fairness of h by

$$\mu_{P_2}^{\text{EO}}(h) \geq \frac{1}{2} \left| \frac{\eta - \epsilon}{\eta} - \frac{\epsilon}{1 - \eta} \right| = \frac{1}{2} \left| \frac{\eta - \eta^2 - \epsilon}{\eta - \eta^2} \right| \leq \frac{1}{4}.$$

□

As a corollary, we find, that there are tasks which are not compatible for joint fair representation, not even if the fairness requirement for the task is limited to accuracy-driven fairness guarantees.

Corollary 15. *Let $0 < \gamma < \frac{\eta - \eta^2}{2} < 0.5$. There are tasks given by distributions P_1 and P_2 with the same marginal, such that every representation ρ that γ -bounded classification degrading with respect to P_1 , is not $(\gamma + \eta, \frac{1}{4})$ -Bayes-accuracy induced fair.*

The fact that this construction and thus the effect of the trade-off is symmetric is of particular interest. The construction shows that any classifier that is approximately accurate with respect to both tasks P_1 and P_2 , will be EO-unfair for at least one of them. Therefore, achieving accuracy-driven fairness, by making sure certain unfair classifiers cannot be expressed with regard to task P_1 , will inevitably remove all approximately fair and accurate classifiers for P_2 . This is because the tasks agree on the majority group but disagree on the minority group. Thus, if task-similarity is defined by accuracy and thus primarily by the majority group, the concerns of a minority group could be disregarded, and a transfer to such a seemingly similar setting would cause harm for the minority group.

6 Testing a Representation for a Given Task

6.1 Testing Complete Fairness

We first consider the question of whether the complete or worst-case fairness of a representation for a distribution is testable.

Towards this goal we define a notion of uniform testability in which data is sampled i.i.d. from a distribution P and the size of a test set that guarantees fairness with high probability does not depend on the underlying distribution P .

Definition 5. Given a fairness notion F . A representation ρ is considered F -fairness-testable with sample complexity $m_\rho^\mu : (0, 1)^2 \rightarrow \mathbb{N}$, if there exists a decision function $d : (\mathcal{X} \times \mathcal{Y})^m \rightarrow \{\text{Yes}, \text{No}\}$, such that for every $\alpha, \delta \in (0, 1)^2$, every $m \geq m_\rho^F(\alpha, \delta)$ and every task (given by a distribution) P if,

- If the representation is complete F -fair with respect to task P , then,

$$\mathbb{P}_{S \sim P^m}[d(S) = \text{Yes}] \geq 1 - \delta$$

- If for the representation ρ is not complete F -fair with respect to P with unfairness $\mu_P^F(\rho) \geq \alpha$, we have

$$\mathbb{P}_{S \sim P^m}[d(S) = \text{No}] \geq 1 - \delta.$$

Definition 6. We say a representation ρ partitions according to group-membership if for every x_1, x_2 in different groups, $\rho(x_1) \neq \rho(x_2)$.

Observation 16. A non-trivial representation ρ partitions according to group-membership if and only if there is $h \in \mathcal{F}_\rho$ with $h(x) = 1$ if and only if $g(x) = \mathcal{A}$.

That is, the representation violates the group identity obfuscation property in the strongest possible way.

Claim 17. Let ρ be a non-trivial representation that does not partition according to group-membership. Then ρ is not DP-fairness testable.

Proof. Consider any non-trivial representation ρ that does not partition according to group-membership. First, since the representation does not partition according to group-membership, there are two elements $x_a, x_b \in \mathcal{X}$, such that $g(x_a) = \mathcal{A}$, $g(x_b) = \mathcal{B}$ and $\rho(x_a) = \rho(x_b)$. Furthermore, since ρ is not trivial, there is an element $z_2 \neq \rho(x_a)$, such that there is $x_2 \in \mathcal{X}$ with $\rho(x_2) = z_2$. Without loss of generality, assume $g(x_2) = \mathcal{A}$. Now assume towards contradiction that ρ was indeed DP-fairness testable. Then there exists a sample complexity function m_ρ^{DP} and a decision function d , such that for every task P_1 with respect to which ρ is fair, with probability $3/4$ over $S_1 \sim P_1^m$ with $m = m_\rho^F(1/4, 1/4)$ we have $d(S_1) = \text{Yes}$. Furthermore, for every task P_2 with respect to which ρ is unfair with $\mu_{P_2}^{\text{DP}}(\rho) \geq 1/4$, we have with probability $3/4$ over $S_2 \sim P_2^m$ with $m = m_\rho^F(1/3, 1/3)$ we have $d(S_2) = \text{No}$.

Now define P_1 by $P_1(x_b, 1) = \sqrt[3]{7/8}$ and $P_1(x_a, 1) = 1 - \sqrt[3]{7/8}$. ρ is clearly fair with respect to P_1 as any classifier expressible over ρ is constant with respect to P_1 .

Similarly, define P_2 by $P_2(x_b, 1) = \sqrt[3]{7/8}$ and $P_2(x_2, 1) = 1 - \sqrt[3]{7/8}$. ρ is clearly unfair with respect

to P_2 with $\mu_{P_2}^{\text{DP}}(\rho) = 1$ as any classifier $h : \mathcal{Z} \rightarrow \mathcal{Y}$ with $h(\rho(x_b)) = 1$ and $h(\rho(x_2)) = 0$ has unfairness $\mu_{P_2}^{\text{DP}}(h) = 1$.

We now show a contradiction to ρ being DP-fairness-testable with sample complexity function m_ρ^{DP} and decision function d by showing that samples from distributions P_1 and P_2 are too hard to distinguish.

Let $S_{x_b} = ((x_b, 1), \dots, (x_b, 1))$ be defined as the sample of m repetitions of the example $(x_b, 1)$. With probability $7/8$, a sample $S_1 \sim P_1^m$ will be equal to S_{x_b} , this means that

$$\mathbb{P}_{\text{randomness of } d}[d(S_{x_b}) = \text{Yes}] \geq \frac{\frac{3}{4} - \frac{1}{8}}{\frac{7}{8}} = \frac{6 - 1}{7} = \frac{5}{7}.$$

Similarly, with probability $7/8$ a sample $S_2 \sim P_2^m$ will be equal to S_b . However, in order for the fairness-testing guarantee of d for ρ to hold, with probability $5/7$, we require $d(S_{x_b}) = \text{"No"}$. Thus, we get a contradiction to the assumption. $\square$

On the other hand, any trivial representation is fair. Furthermore, any representation that partitions according to group membership, allows for the expression of the classifier $h : \mathcal{X} \rightarrow \mathcal{Y}$ with $h(x) = 1$ if and only if $g(x) = \mathcal{A}$. This classifier clearly has maximum DP-unfairness for any distribution P . Thus such a representation is fairness testable, since the constant "No" predictor always satisfies the requirements for fairness-testability.

Claim 18. No non-trivial representation ρ that doesn't partition according to group-membership is μ^{EO} -fairness-testable.

Proof. The proof works exactly like the one of Claim 17. We note that the distribution P_1 defined in that proof is also EO-fair with respect to ρ . Furthermore, the distribution P_2 from that proof is EO-unfair with $\mu_{P_2}^{\text{EO}}(\rho) = \frac{1}{2}$. $\square$

The issue we run into when considering uniform testability is that group sizes may be very small. For every sample size m there is a small enough (roughly of order $\frac{1}{m}$) such that if the probability weight of a group w.r.t. P is smaller than that, then a sample of size m is unlikely to have instances from that group. In such a case the sample does not allow evaluating whether that group has been treated fairly. This difficulty of testing becomes more likely to occur when there are more groups, i.e., more sensitive attributes or sensitive attributes with more than two values.

One potential way of overcoming this would be through access to i.i.d. samples from the group (and label) conditional distributions P_G and $P_{G,1}$.

Definition 7. Given a fairness notion F . A representation ρ is considered group-conditional- F -fairness-testable with sample complexity $m_\rho^\mu : (0, 1)^2 \rightarrow \mathbb{N}$, if there exists a decision function $d : (\mathcal{X} \times \mathcal{Y})^m \rightarrow \{\text{Yes}, \text{No}\}$, such that for every $\alpha, \delta \in (0, 1)^2$ and every $m \geq m_\rho^F(\alpha, \delta)$ task (given by a distribution) P if,

- If the representation is F -fair with respect to task P , then,

$$\mathbb{P}_{S_A \sim P_A^m, S_B \sim P_B^m}[d(S_A, S_B) = \text{Yes}] \geq 1 - \delta$$

- If for the representation ρ is F -unfair with respect to P with $\mu_P^F(\rho) \geq \alpha$, we have

$$\mathbb{P}_{S_A \sim P_A^m, S_B \sim P_B^m}[d(S_A, S_B) = \text{No}] \geq 1 - \delta.$$

We assume, that any distribution considered for this test has positive weight on both groups, as otherwise the notion of fairness is neither formally defined nor is it a worthwhile consideration in such a scenario.

We now show that discrete representations are indeed conditional fairness testable, with their sample size depending on the representation's expressiveness.

Claim 19. *Discrete representations are group-conditional- μ^{DP} -fairness testable with $m_\rho^{\text{DP}}(\alpha, \delta) \in \Theta\left(\max\left\{\frac{|\sqrt{|Z_\rho^{\text{eff}}| + \log(1/\delta)}|}{\alpha^2}, \frac{|Z_\rho^{\text{eff}}|^{2/3} + \log(1/\delta)}{\alpha^{4/3}}\right\}\right)$, where $|Z_\rho^{\text{eff}}| = |\{z \in \mathcal{Z} : \exists x \in \mathcal{X} \text{ with } \rho(x) = z\}|$ represents the effective size of the range of ρ .*

Proof. We first note that

$$\begin{aligned} \mu_P^{\text{DP}}(\rho) &= \sup_{h \in \mathcal{F}_\rho} |\mu_P^{\text{DP}}(h)| \\ &= \sup_{h \in \mathcal{F}_\rho} |\mathbb{P}_{(x,y) \sim P}[h(1) = 1 | g(x) = \mathcal{A}] \\ &\quad - \mathbb{P}_{(x,y) \sim P}[h(1) = 1 | g(x) = \mathcal{B}]| \\ &= \sup_{h \in \mathcal{F}_\rho} |P_{\mathcal{A}}(\{x \in \mathcal{X} : h(x) = 1\}) \\ &\quad - P_{\mathcal{B}}(\{x \in \mathcal{X} : h(x) = 1\})| \\ &= d_{\mathcal{F}_\rho}(P_{\mathcal{A}}, P_{\mathcal{B}}), \end{aligned}$$

where $d_{\mathcal{F}_\rho}$ denotes the $\mathcal{F}_\rho$ -distance between two distributions. Thus, the question of whether $\mu_P^{\text{DP}}(\rho) \leq \alpha$ boils down to the question of whether $d_{\text{TV}} \mathcal{F}_\rho(P_{\mathcal{A}}, P_{\mathcal{B}}) \leq \alpha$. This is a known testing problem, which is known to have sample complexity $\Theta\left(\max\left\{\frac{|\sqrt{|Z_\rho^{\text{eff}}| + \log(1/\delta)}|}{\alpha^2}, \frac{|Z_\rho^{\text{eff}}|^{2/3} + \log(1/\delta)}{\alpha^{4/3}}\right\}\right)$ (Chan et al. 2014).

□

Observation 20. *No non-trivial representation ρ that doesn't partition according to groups is group-conditional- μ^{EO} -fairness-testable.*

The proof follows the same idea as Claim 18 with essentially the same construction, with slightly changed numbers to account for the additional tests and balanced group-sizes. We show hardness by allowing the proportions of labels to be arbitrarily small. The full proof can be found in the appendix.

Definition 8. Given a fairness notion F , a representation ρ is considered group-label-conditional- F -fairness-testable with sample complexity $m_\rho^F : (0, 1)^2 \rightarrow \mathbb{N}$, if there exists a decision function $d : (\mathcal{X} \times \mathcal{Y})^m \rightarrow \{\text{Yes}, \text{No}\}$, such that for every $\alpha, \delta \in (0, 1)^2$ and every $m \geq m_\rho^F(\alpha, \delta)$ and every task (given by a distribution) P :

- If the representation is F -fair with respect to task P , then with probability $1 - \delta$ over the generation of all samples $S_{\mathcal{A},1} \sim P_{\mathcal{A},1}^m, S_{\mathcal{B},1} \sim P_{\mathcal{B},1}^m, S_{\mathcal{A},0} \sim P_{\mathcal{A},0}^m, S_{\mathcal{B},0} \sim P_{\mathcal{B},0}^m$, we get $d(S_{\mathcal{A},1}, S_{\mathcal{B},1}, S_{\mathcal{A},0}, S_{\mathcal{B},0}) = \text{Yes}$.
- If for the representation ρ is F -unfair with respect to P with $\mu_P^F(\rho) \geq \alpha$, then with probability $1 - \delta$ over the generation of all samples $S_{\mathcal{A},1} \sim P_{\mathcal{A},1}^m, S_{\mathcal{B},1} \sim P_{\mathcal{B},1}^m, S_{\mathcal{A},0} \sim P_{\mathcal{A},0}^m, S_{\mathcal{B},0} \sim P_{\mathcal{B},0}^m$, we get

$$d(S_{\mathcal{A},1}, S_{\mathcal{B},1}, S_{\mathcal{A},0}, S_{\mathcal{B},0}) = \text{No}.$$

We will only consider distributions P in which all combinations (g, l) receive positive weight, since it is only in those settings that EO-fairness is defined.

Claim 21. *Discrete representations are group-label-conditional- μ^{EO} -fairness testable with $m_\rho^{\text{EO}}(\alpha, \delta) \in \Theta\left(\max\left\{\frac{|\sqrt{|Z_\rho^{\text{eff}}| + \log(1/\delta)}|}{\alpha^2}, \frac{|Z_\rho^{\text{eff}}|^{2/3} + \log(1/\delta)}{\alpha^{4/3}}\right\}\right)$.*

The proof of this claim follows the same idea as the proof of Claim 19, which reduces the question of estimating the fairness of a representation to distribution testing over a finite domain. The full proof can be found in the appendix.

An implication of Claims 19 and 21 is that if a representation is not discrete or allows for the expression of infinitely many different feature vectors, the worst-case fairness of a representation cannot be tested.

For example many feature vectors in medical settings are continuous or require a high precision for correct diagnosis.

In the next section, we consider ways of overcoming this hardness by restricting the class of potentially usable classifiers beyond what can be expressed by the representation.

6.2 Restricting the Expressible Classifiers

As we have seen in the previous section, when testing the fairness of a representation for a given downstream task, guaranteeing the fairness of every potential classifier that may be defined over that representation, may be too hard if the representations are too expressive, in particular if it assumes continuous values.

However, in many such settings it may be feasible to further restrict the family of potential models. For example, in many situations which require continuous values, only easily interpretable classifiers, such as sparse linear classifiers or those which have decision boundaries expressible through low-degree polynomials over few features matter. In such cases we can improve the results for testing the fairness of a representation.

Restriction of function class for downstream use We consider the case in which the use of a representation has been restricted to a class $\mathcal{H} \subset \mathcal{Y}^{\mathcal{Z}}$. In this case, the resulting classifiers for a representations ρ are given by $\mathcal{H}_\rho := \{\rho \circ h : h \in \mathcal{H}\}$. Similar to the fairness of ρ , we define the (un)fairness of F the (un)fairness of $\mathcal{H}_\rho$ as the least fair expressible classifier, i.e., $\mu_P^F(\mathcal{H}_\rho) = \sup_{h \in \mathcal{H}_\rho} \mu_P^F(\rho)$. The testability of a representation-based class $\mathcal{H}_\rho$ is defined analogous to the previous subsection. We say a representation-based class $\mathcal{H}_\rho$ partitions according to group membership

if there is $h \in \mathcal{H}_\rho$ with $h(x) = \mathbf{1}[g(x) = \mathcal{A}]$ or $h(x) = \mathbf{1}[g(x) = \mathcal{B}]$.

Observation 22. *No non-trivial representation-based class $\mathcal{H}_\rho$ (that doesn't partition according to group-membership) is μ^{DP} -fairness-testable or μ^{EO} -fairness-testable.*

Proof. The proofs are equivalent to the proofs Claims 17 and 18. More precisely, in the DP case from the assumptions, it follows that there exists an expressible classifier $h \in \mathcal{H}_\rho$ and points x_a, x_b, x_2 with $h(x_a) = h(x_b) \neq h(x_2)$ and $g(x_a) = g(x_2) = \mathcal{A}$ and $g(x_b) = \mathcal{B}$. The proof then follows from the same argument as in the proof of Claim 17. Furthermore, as noted in the proof of Claim 18 the construction also suffices to get the impossibility result for testing EO fairness. $\square$

Theorem 23. *If $\text{VC}(\mathcal{H}_\rho)$ is finite the following two statements are true (for d denoting this VC-dimension):*

- $\mathcal{H}_\rho$ is group-conditional-DP fairness testable with sample complexity: $O\left(\frac{d+\log(1/\delta)}{\alpha^2}\right)$.
- $\mathcal{H}_\rho$ is group-label-conditional-EO fairness testable with sample complexity: $O\left(\frac{d+\log(1/\delta)}{\alpha^2}\right)$.

The proofs follow the uniform convergence arguments of section 3 in (Kifer, Ben-David, and Gehrke 2004).

Lower bounds on the sample complexities of these testing tasks can be readily shown by replacing the effective domain sizes in Claim 17 and Claim 18 by the VC-dimension of the classes.

Note that in this case we are left with a small gap between the upper and lower bounds. We leave closing these gaps to future work.

7 Discussion

In this work we re-examined the role of fair representations after recent impossibility results, to see if those results can be meaningfully overcome to achieve what fair representations seem to promise: A way to transfer fairness guarantees across tasks, without the need to spend additional data or computational resources on fairness for each individual task. We first examined the possibility of weakening fairness guarantees to only account for approximately accurate classifiers. This may be appropriate in settings in which representation users can be trusted not to misuse a representation for the purpose of adversarial discrimination, but where unfairness comes more as an unwanted byproduct of bias in the data. We showed that restricting to this less restrictive guarantee is no cure-all to overcome the issues of transferability. We find that in order for a representation to offer any non-trivial fairness guarantee, even in this weaker, accuracy-driven setting, the transferability to different tasks is severely limited.

We then discussed another potential venue to allow a potential transfer or re-use of representations; We consider the option that users of a representations (or third parties) would test whether a representation achieves fairness for a given task at hand. Here, our results are still rather pessimistic:

While it is possible to achieve fairness testability, if one has sufficient access to data from all subgroups, the representations need to be of limited expressiveness, or further restrictions on the use of classification rules needs to be made (also limiting down-stream expressiveness).

Our work also highlights the importance of group imbalance: Non-trivial fairness improvements through restriction of the representations in terms of accuracy-driven EO fairness can only be achieved if there is a significant imbalance in group proportions. However, it is exactly in situations where group proportions are highly imbalanced that we get hardness results for both the fairness transferability and the fairness testability of representations.

Below we offer some further thoughts about how our results on testability and accuracy-driven fairness could be viewed.

Testability of Accuracy-driven fairness. As accuracy-driven fairness cannot be guaranteed in cases with high group imbalance, future work may aim to develop tests for accuracy-driven fairness of a representation.

However, analogous to our results for the testability of complete fairness, the reliability of such tests can only be expected if there is access to sufficiently large samples from each subpopulation. This may not be feasible in situations with high group imbalance.

Testing fairness of a representation vs testing fairness of a classifier One question to ask when it comes to testability, is whether it would be better to test fairness of a trained classifier rather than the fairness of a representation. To test the fairness of a trained classifier, one must of course first learn said classifier, which may be computationally and statistically expensive. Therefore, testing a representation for fairness in advance seems appealing. However, we note that given the sample complexity bounds for evaluating the fairness of a representation, this task is not statistically easier than training the classifier. In comparison, doing an additional fairness evaluation of the learned classification rule, seems rather inexpensive. Thus, in terms of data-efficiency of testing the fairness of a representation in advance does not seem to save any additional costs, but rather make the problem harder.

There may still be reasons to prefer testing the fairness of representation in advance. In settings in which a decision rule needs to be applied consistently and trained in advance, a decision maker may not want to reveal their decision rule to outsiders in order to limit potential manipulation in an ongoing round of decision making. In such cases, they only may settle for revealing that a predefined class $\mathcal{H}_\rho$ of candidate classifiers rather than the trained classifier. Adherence to the restriction to $\mathcal{H}_\rho$ could still falsified in on-going decision making via finite evidence sets, if the class $\mathcal{H}_\rho$ indeed has finite VC-dimension, thus allowing recourse for individuals (Ben-David et al. 2025).

Acknowledgements

Shai Ben-David's research is supported by an NSERC Discovery grant and through a CIFAR AI Chair. Ruth

Urner’s research is supported by an NSERC Discovery grant. Tosca Lechner was supported by a Vector Postdoctoral Fellowship.

References

- Ben-David, S.; Gourdeau, P.; Lechner, T.; and Urner, R. 2025. What’s Individual About Individual Fairness? *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1): 379–390.
- Cerrato, M.; Köppel, M.; Wolf, P.; and Kramer, S. 2024. 10 Years of Fair Representations: Challenges and Opportunities. *CoRR*, abs/2407.03834.
- Chan, S.-O.; Diakonikolas, I.; Valiant, G.; and Valiant, P. 2014. Optimal algorithms for testing closeness of discrete distributions. *SODA ’14*, 1193–1203. USA: Society for Industrial and Applied Mathematics. ISBN 9781611973389.
- Creager, E.; Madras, D.; Jacobsen, J.; Weis, M. A.; Swersky, K.; Pitassi, T.; and Zemel, R. S. 2019. Flexibly Fair Representation Learning by Disentanglement. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA*, Proceedings of Machine Learning Research, 1436–1445. PMLR.
- Deka, N.; and Sutherland, D. J. 2023. MMD-B-Fair: Learning Fair Representations with Statistical Testing. In Ruiz, F. J. R.; Dy, J. G.; and van de Meent, J., eds., *International Conference on Artificial Intelligence and Statistics, 25–27 April 2023, Palau de Congressos, Valencia, Spain*, Proceedings of Machine Learning Research, 9564–9576. PMLR.
- Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. S. 2012. Fairness through awareness. In Goldwasser, S., ed., *Innovations in Theoretical Computer Science ITCS, 2012*, 214–226. ACM.
- Gitiaux, X.; and Rangwala, H. 2021. Learning Smooth and Fair Representations. In Banerjee, A.; and Fukumizu, K., eds., *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13–15, 2021, Virtual Event*, Proceedings of Machine Learning Research, 253–261. PMLR.
- Gitiaux, X.; and Rangwala, H. 2022. SoFair: Single Shot Fair Representation Learning. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23–29 July 2022*, 687–695. ijcai.org.
- Jang, T.; Gao, H.; Shi, P.; and Wang, X. 2024. Achieving Fairness through Separability: A Unified Framework for Fair Representation Learning. In Dasgupta, S.; Mandt, S.; and Li, Y., eds., *International Conference on Artificial Intelligence and Statistics, 2–4 May 2024, Palau de Congressos, Valencia, Spain*, Proceedings of Machine Learning Research, 28–36. PMLR.
- Jones, C.; Ribeiro, F. D. S.; Roschewitz, M.; Castro, D. C.; and Glocker, B. 2025. Rethinking Fair Representation Learning for Performance-Sensitive Tasks. In *The Thirtieth International Conference on Learning Representations, ICLR 2025, Singapore, April 24–28, 2025*. OpenReview.net.
- Jovanovic, N.; Balunovic, M.; Dimitrov, D. I.; and Vechev, M. T. 2023. FARE: Provably Fair Representation Learning with Practical Certificates. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *International Conference on Machine Learning, ICML 2023, 23–29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, 15401–15420. PMLR.
- Kairouz, P.; Liao, J.; Huang, C.; Vyas, M.; Welfert, M.; and Sankar, L. 2022. Generating Fair Universal Representations Using Adversarial Models. *IEEE Trans. Inf. Forensics Secur.*, 17: 1970–1985.
- Kifer, D.; Ben-David, S.; and Gehrke, J. 2004. Detecting Change in Data Streams. In *(e)Proceedings of the Thirtieth International Conference on Very Large Data Bases, VLDB 2004*, 180–191. Morgan Kaufmann.
- Kim, D.; Kim, K.; Kong, I.; Ohn, I.; and Kim, Y. 2022. Learning fair representation with a parametric integral probability metric. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *International Conference on Machine Learning, ICML 2022*.
- Kim, J.; and Cho, S. 2020. Fair Representation for Safe Artificial Intelligence via Adversarial Learning of Unbiased Information Bottleneck. In Espinoza, H.; Hernández-Orallo, J.; Chen, X. C.; ÓhÉigeartaigh, S. S.; Huang, X.; Castillo-Effen, M.; Mallah, R.; and McDermid, J. A., eds., *Proceedings of the Workshop on Artificial Intelligence Safety, co-located with 34th AAAI Conference on Artificial Intelligence, SafeAI@AAAI 2020*.
- Kleindessner, M.; Donini, M.; Russell, C.; and Zafar, M. B. 2023. Efficient fair PCA for fair representation learning. In Ruiz, F. J. R.; Dy, J. G.; and van de Meent, J., eds., *International Conference on Artificial Intelligence and Statistics, 25–27 April 2023, Palau de Congressos, Valencia, Spain*, Proceedings of Machine Learning Research, 5250–5270. PMLR.
- Lechner, T.; and Ben-David, S. 2022. Inherent Limitations of Multi-Task Fair Representations. In Chandar, S.; Pascanu, R.; and Precup, D., eds., *Conference on Lifelong Learning Agents, CoLLAs*, volume 199 of *Proceedings of Machine Learning Research*, 583–603. PMLR.
- Liu, J.; Li, Z.; Yao, Y.; Xu, F.; Ma, X.; Xu, M.; and Tong, H. 2022. Fair Representation Learning: An Alternative to Mutual Information. In Zhang, A.; and Rangwala, H., eds., *KDD ’22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 – 18, 2022*, 1088–1097. ACM.
- Luo, Y.; Hoag, A.; Wang, X.; Thomas, P. S.; and Grabowicz, P. A. 2026. Fair Representation Learning with Controllable High Confidence Guarantees via Adversarial Inference. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. S. 2018. Learning Adversarially Fair and Transferable Representations. In Dy, J. G.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, July 10–15,*

2018, *Proceedings of Machine Learning Research*, 3381–3390. PMLR.

McNamara, D.; Ong, C. S.; and Williamson, R. C. 2019. Costs and Benefits of Fair Representation Learning. In Conitzer, V.; Hadfield, G. K.; and Vallor, S., eds., *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES*, 263–270. ACM.

Ni, Y.; and Huo, X. 2026. Kernel-based Equalized Odds: A Quantification of Accuracy-Fairness Trade-off in Fair Representation Learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Oh, C.; Won, H.; So, J.; Kim, T.; Kim, Y.; Choi, H.; and Song, K. 2022. Learning Fair Representation via Distributional Contrastive Disentanglement. In Zhang, A.; and Rangwala, H., eds., *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, 1295–1305. ACM.

Oneto, L.; Donini, M.; Pontil, M.; and Maurer, A. 2020. Learning Fair and Transferable Representations with Theoretical Guarantees. In Webb, G. I.; Zhang, Z.; Tseng, V. S.; Williams, G.; Vlachos, M.; and Cao, L., eds., *7th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2020*.

Pouget, A.; Jovanovic, N.; Vero, M.; Staab, R.; and Vechev, M. T. 2024. Back to the Drawing Board for Fair Representation Learning. *CoRR*, abs/2405.18161.

Shen, X.; Wong, Y.; and Kankanhalli, M. S. 2023. Fair Representation: Guaranteeing Approximate Multiple Group Fairness for Unknown Tasks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1): 525–538.

Shui, C.; Chen, Q.; Li, J.; Wang, B.; and Gagné, C. 2022. Fair Representation Learning through Implicit Path Alignment. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *International Conference on Machine Learning, ICML 2022*.

Sinigaglia, A.; Sartor, D.; Ceccon, M.; and Susto, G. A. 2026. Simple and Effective Specialized Representations for Fair Classifiers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zemel, R. S.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, JMLR Workshop and Conference Proceedings, 325–333. JMLR.org.

Zhao, H.; Coston, A.; Adel, T.; and Gordon, G. J. 2020. Conditional Learning of Fair Representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhao, H.; Dan, C.; Aragam, B.; Jaakkola, T. S.; Gordon, G. J.; and Ravikumar, P. 2022. Fundamental Limits and Tradeoffs in Invariant Representation Learning. *Journal of Machine Learning Research*, 23(340): 1–49.

Zhao, H.; and Gordon, G. J. 2022. Inherent Tradeoffs in Learning Fair Representations. *Journal of Machine Learning Research*, 23(57): 1–26.