

# A Different View of Fair Data Representation

## Abstract

We investigate notions of fair data representation as reflected in the fairness of classification predictions of an agent accessing training data represented that way. The resulting notions rely of course on the notion of fair classification being applied, but also on the agent’s objective driving that classification. We distinguish three major types of such objectives; malicious (driven by a bias against a group of subjects), accuracy driven (ignoring fairness considerations), and fairness driven.

The main questions we analyze under this taxonomy of representation fairness notions are: 1) Can there be a generic fair representation (independent of the classification task)? 2) Can the effect of a given feature in a representation on fairness be determined independently of the other features involved? 3) What is the effect of training sample sizes on the fairness based on such representations?

We show that, for central notions of classification fairness (e.g., odds equality), the answer to the first question is negative for all the three different objectives. This is in contrast to previous results suggesting a positive answer when the underlying notion of fairness is demographic parity. We furthermore show that there are tasks which are fairness incompatible, i.e. there is no representation that prevents unfairness and allows accurate representation for both tasks simultaneously.

For the second question we also give negative answers for all three classification objectives under various fairness notions. In particular, we show that in the accuracy driven case a feature can have opposing fairness effects, dependent on the other available features.

Finally, we address the effect of sampling size on unfairness, as a property of the representation, in the accuracy driven case. We show that there are situations where the the Bayes optimal classifier is fair, but sample sizes that guarantee high accuracy do not suffice to guarantee fairness.

## 1 Introduction

Automated decision making has become more and more successful over the last few decades and has therefore been used in an increasing number of domains, either to fully automate decision making, or to support human decision makers. This includes many sensitive domains which significantly impact people’s livelihoods, such as loan applications, university admissions, recidivism predictions, or insurance rate settings. It has been found that many such decision tools have, often unintentionally, biases against minority groups, and therefore

lead to discrimination. In response to these concerns, the machine learning research community has been devoting effort to developing precise notions of fair decision making, and coming up with algorithms for implementing fair machine learning.

While many papers in this domain propose algorithmic solutions to fairness related issues, the main contributions of this paper are conceptual. We feel that unlike many other facets of machine learning, the field of fairness in machine learning is still in need of investigating its fundamental concepts and setup. Some basic questions are still far from being satisfactorily elucidated; What should be considered fair decision making? (various mutually inconsistent notions have been proposed, but the real life application of picking between them still lacks clarifications). What is a fair data representation? To what extent should accuracy or other practical utilities and costs be compromised for achieving such goals? and so on. The answers to these questions are not generic. They vary with the principles and the goals guiding the agents involved (decision makers, subjects of such a decision, policy regulators, etc.), as well as with what can be assumed regarding the underlying learning setup. We view those as the primary issues facing this area, deserving explicit research attention (in addition to the more common algorithmic and optimization aspects). In this paper we address one of these questions - the nature of fair data representation.

Machine learning tools base their classification output on the data they access for training. The features used for representing that data play a major role in the outcome of an algorithm trained on such inputs. This realization has spurred research on notions of fair data representation, which is the subject of this paper.

What is a fair data representation? Most previous works on fair representation aim to find representations that can be considered fair across many classification prediction tasks (often termed ‘transferable’ or ‘flexibly fair’). Consequently, they impose some level of *demographic parity* or independence of the feature values from protected group membership. Such properties have the advantage that they can be verified for a feature regardless of the way it may be used by an agent (Zhao and Gordon 2019; Creager et al. 2019; Madras et al. 2018; Oneto et al. 2019). A downside of this approach is that, once there is significant correlation between group membership and the target classification, it is in conflict with

88 accuracy. Any classifier that is based on such fair features  
89 will fail to accurately match a classification that is highly  
90 correlated with the groups it requires parity for.

91 We address this question from the perspective of its effect  
92 on the classification rules that an agent using data represented  
93 that way may come up with. Such a view takes into consid-  
94 eration (and therefore is a function of and adapted to) two  
95 setup characteristics:

- 96 1. The type of agent using the data, depending on the ob-  
97 jective of the agent; malicious (driven by a bias against  
98 a group of subjects), accuracy driven (ignoring fairness  
99 considerations), or fairness driven.
- 100 2. The applied notion of group fairness of classification de-  
101 cisions. In this paper we primarily focus on the notion of  
102 odds equality.

103 Applying our notion of fairness of representations to  
104 ground-truth-related notions of group fairness in the spirit of  
105 odds equality (rather than demographic parity), allows for  
106 fair representations that are aligned with accurate classifica-  
107 tion prediction (including when group membership is highly  
108 correlated with the target class).

109 Dependent on those two considerations we examine three  
110 questions concerning the choice of features for a fair data  
111 representation:

- 112 1. Can there be a generic fair representation (independent of  
113 the classification task the representation will be used for)?
- 114 2. Can the effect of a given feature in a representation on the  
115 fairness of the representation be determined independently  
116 of the other features involved?
- 117 3. What is the effect of training sample sizes on the fairness  
118 of a classifier trained on data under the representation?

119 We show that, for central notions of classification fair-  
120 ness (e.g., odds equality), the answer to the first question is  
121 negative for all three types of agent objectives, which are  
122 malicious, fairness driven, and accuracy driven (Observa-  
123 tion 4, Observation 5, Theorem 2). This is in contrast with  
124 previous works cited above suggesting a positive answer  
125 when the underlying notion of fairness is demographic parity.  
126 We furthermore show that there are tasks which are fairness  
127 incompatible (Theorem 3). Namely, pair of tasks for which  
128 there is no representation that guarantees fair classification  
129 and allows accurate representation for both tasks simultane-  
130 ously.

131 For the second question we also give negative answers  
132 for all three classification objectives under various fairness  
133 notions. In particular, we consider the fairness implications  
134 of feature deletion. A well known real world example of this  
135 is the "ban the box" policy which disallowed employers using  
136 criminal history in hiring decisions (Doleac and Hansen  
137 2016). It was observed that such a policy could lead to less  
138 fair decision making even though it was motivated by fairness  
139 considerations. Evaluating the fairness of a representation  
140 through the classification decisions they imply provides a  
141 framework for analysing such issues. We show that it is not  
142 possible to determine the impact a single feature has on the  
143 fairness of a representation if other available features are not  
144 considered. This can be shown for many fairness notions for

145 representation that consider a decision maker who actively  
146 tries to discriminate (Observation 2, Observation ?? and Ob-  
147 servation ??) and for decision makers who try to actively  
148 pursue fair classification (Observation 3). Furthermore, we  
149 show that when decision makers optimize accuracy and are  
150 indifferent to fairness, feature deletion of the same feature  
151 can have opposing effects depending on the context of other  
152 available features: When considering the fairness of the most  
153 accurate decision-rule for a given feature set, we show that  
154 the unfairness can increase when deleting a feature for a given  
155 set of available features, while deleting the same feature can  
156 decrease unfairness, if the set of other available features is  
157 different (Theorem 1).

158 One should note that while the data representation and task  
159 determines the Bayes classifier, in most cases the decision  
160 maker does not have access to the Bayes classifier. Instead,  
161 it usually bases its decision on some training sample. The  
162 last question we investigate is the effect of training sample  
163 sizes on unfairness, as a property of the representation. We  
164 show that there are situations where the the Bayes optimal  
165 classifier is fair, but sample sizes that guarantee good approx-  
166 imation of that optimal classifier in terms of accuracy do not  
167 suffice to guarantee a similar level of fairness (Theorem 4).  
168 It follows that in such scenarios, an accuracy driven agent  
169 may end up with an unfair classifier if it settles for training  
170 samples of sizes sufficient to guarantee accuracy. We give a  
171 bound of this effect in terms of the probability mass of the  
172 least likely group-label combination (Theorem 5) and suggest  
173 importance sampling as a countermeasure for unfairness due  
174 to under-sampling.

175 Finally, we look into the details of accuracy-driven and  
176 adversarial fairness. We give necessary and sufficient condi-  
177 tions on the alignment of conditional distributions given  
178 labels and groups for accuracy-driven fairness (Theorem 8)  
179 and adversarial fairness (Theorem 7) for the equalized odds  
180 notion of group fairness. For both of those cases we bound  
181 the unfairness in terms of distribution distances of condition-  
182 als (Theorem 9). We furthermore connect adversarial fairness  
183 for various group-fairness notions to independence consider-  
184 ations between the feature-vector, ground-truth-labeling and  
185 group-membership of random instances (Theorem 6, Theo-  
186 rem 10). We end this discussion by showing that there are  
187 also group-fairness notions for which adversarial fairness is  
188 not always achievable (Theorem 10), i.e., we show that there  
189 can be no adversarially fair representation with respect to  
190 predictive rate parity, if the success rates between groups is  
191 not equal.

192 Our paper is organized as follows: Section 2 introduces  
193 the formal setup. Section 3 introduces our taxonomy of fair  
194 representations. Section 4 discusses the impact on a single  
195 feature on the fairness of a representation. Section 5 addresses  
196 the question of transferable fair representation. Section 6  
197 discusses unfairness due to under-sampling as a separate  
198 source of unfairness. Section 7 gives some characterization  
199 of accuracy-driven and adversarially fair representations. In  
200 Section 8 we present related work. Section 9 concludes this  
201 paper with some final remarks.

## 2 Formal Setup

**Domain set**  $X$  is the set of instances that we wish to classify (say people that may be applying for a loan)

**Data generating probability distribution** We assume that there is some fixed probability distribution  $P$  over  $X \times \{0, 1\}$ . As common in PAC learning theory, we assume that the training data is generated i.i.d. by that distribution.

**Ground truth labeling rule**  $t : X \rightarrow [0, 1]$ . We will think of the label 1 as denoting ‘qualified’ and the label 0 as ‘unqualified’ and  $t(x) = P[y = 1|x]$ . For concreteness, we focus here on the case of deterministic labeling (that is  $t : X \rightarrow \{0, 1\}$ ). Most of our discussion can readily be extended to the probabilistic labeling case as well.

**Group membership** : We assume that  $X$  is partitioned by sets  $A$  and  $D$  We further use the function  $G : X \rightarrow \{A, D\}$  to indicate the group-membership of an instance.

**Feature Set**  $\mathcal{F} = \{f_1, \dots, f_n\}$  where each  $f_i : X \times \{0, 1\} \rightarrow Y_i$  is a feature (we think of  $Y_i$  as being finite, maybe a discretization of an infinite set of values).

**Feature-induced cells**  $\mathcal{F}$  includes an equivalence relation  $\sim_{\mathcal{F}}$  over  $X \times \{0, 1\}$ . Namely,  $x \sim_{\mathcal{F}} y$  if for every  $f \in \mathcal{F}$ ,  $f(x) = f(y)$ . Let  $\mathcal{C}_{\mathcal{F}}$  be the set of the  $\sim_{\mathcal{F}}$  equivalence classes. We call each such equivalence class a ‘cell’. That is, cells are formed by grouping together individuals with the same representation.

**Ground truth scoring** We define the *ground truth score function*  $s_t : \mathcal{C}_{\mathcal{F}} \rightarrow [0, 1]$ .  $s_t^P(C)$  is the probability, w.r.t.  $P$ , of  $x \in C$  having the true-label 1. i.e.,

$$s_t^P(C) = \mathbb{E}_{x \sim P}[t(x)|x \in C]$$

In cases where the distribution is non-ambivalent we will use the abbreviated notation  $s_t$  instead of  $s_t^P$ .

**Decision rule** A feature based decision rule is a function  $h : \mathcal{C}_{\mathcal{F}} \rightarrow \{0, 1\}$ . We denote the hypothesis class of all feature based decision rules as  $H_{\mathcal{C}_{\mathcal{F}}}$ .

**Learner** We consider learners that get as input a labeled sample and output a decision rule (a.k.a. label predictor). The learner (or decision maker) does not know what  $P$  or  $t$  are. Furthermore, rather than seeing the identity of each instance in the training sample (or the test data), the decision maker only sees the feature based representation  $F(x) = (f_1(x), \dots, f_n(x))$ .

### 2.1 Loss minimizing decision maker

A decision maker uses a training sample to pick predictor  $h$  aiming to minimize some loss function. In this work we consider weighted 0/1 losses. Namely, for some given parameter  $0 \leq \alpha \leq 1$ , false negative instances incur a loss  $\alpha$  and false positives  $(1 - \alpha)$ :

$$L_P^\alpha(h) = \alpha P(\{x : t(x) = 1, h(x) = 0\}) + (1 - \alpha)P(\{x : t(x) = 0, h(x) = 1\})$$

For a threshold  $\alpha = 0.5$  we will use the notation  $L_P = L_P^{0.5}$ . We denote the set of threshold classifiers w.r.t. the ground-truth scoring function  $s_t$  by

$$\mathcal{H}_{\mathcal{C}_{\mathcal{F}}, s_t}^{\text{thres}} = \{h : \mathcal{C}_{\mathcal{F}} \rightarrow \{0, 1\} : \text{for some } \alpha, h(C) = 0 \text{ iff } s_t(C) < \alpha\}$$

**Lemma 1.** (Corbett-Davies and Goel 2019) *The predictor in  $\mathcal{H}_{\text{cells}}$  that minimizes  $L_P^\alpha$  is the Bayes Optimal predictor  $t_{P,F}^\alpha$  that for a cell  $C \in \mathcal{C}_{\mathcal{F}}$  assigns the label 1 if  $s_t(C) > \alpha$  and 0 otherwise.*

## 3 Taxonomy of fair representation notions

In this section we propose several notions for the fairness of a representation. Since we focus on data representation for classification/decision-making purposes, our notions of fair representation are induced by notions of fair decision making. To define a notion of fair representation, we therefore need to answer two questions 1) When is a decision rule considered fair? and 2) Which decision rule will a decision maker arrive at?. The first question has been addressed in the fairness literature in various ways by the introduction of group fairness notions. Our discussion will mainly focus on the equalized odds notion of fairness (Hardt, Price, and Srebro 2016). For the second question we distinguish between three different motivations of decision makers: accuracy-driven, fairness-driven and malicious. We will start now by giving a definition of some common notions of group fairness.

### Group fairness notions (for decision making)

**Definition 1** (Group fairness; Equalized odds). *Group fairness is a property of a classifier that takes into account the classifiers predictions and group membership. We say a group fairness notion is ground-truth-related, if it also considers the ground truth label. The notion of group-fairness we will focus on in this paper is the ground-truth-related notion of odds equality as introduced by (Hardt, Price, and Srebro 2016).*

*A classifier  $h$  is considered fair w.r.t. to odds equality ( $L^{EO}$ ) and a distribution  $P$  if  $h(x) \perp\!\!\!\perp G(x)|t(x)$ . The respective unfairness is given by the sum of differences in false positive rate and false negative rate: For  $g \in \{A, D\}$*

$$FNR_g(h, t, P) = \frac{P(\{x \in X_{g,1} : h(x) = 0\})}{P(X_{g,1})}$$

$$FPR_g(h, t, P) = \frac{P(\{x \in X_{g,0} : h(x) = 1\})}{P(X_{g,0})}$$

$$L_P^{EO}(h) = |FNR_A - FNR_D| + |FPR_A - FPR_D|$$

*If we say a classifier is fair, without referring to any particular group-fairness notion, we mean fairness w.r.t. equalized odds.*

Another widely used notion of group-fairness that we will refer to throughout the paper is demographic parity.

**Definition 2.** *A classifier  $h$  is considered fair w.r.t. to demographic parity ( $L^{DP}$ ) and a distribution  $P$  if  $h(x) \perp\!\!\!\perp G(x)$ . The respective unfairness is given by difference in positive classification rates between groups*

$$L_P^{DP}(h) = \left| \frac{P(h^{-1}(1) \cap A)}{P(A)} - \frac{P(h^{-1}(1) \cap D)}{P(D)} \right|$$

278 **Fairness of a representation** We will now give our defini-  
 279 tions of representation fairness in terms of a general group  
 280 fairness notion  $L^{\text{fair}}$  with unfairness measure  $L_P^{\text{fair}}$ . We wish  
 281 to provide fairness guarantees with respect to this notion to  
 282 the decision rules a decision maker arrives at. We therefore  
 283 introduce different fairness notions for representations for dif-  
 284 ferent classification objectives that would inform the decision  
 285 rule that would be used.

286 We start by considering a *malicious decision maker* who  
 287 tries to actively discriminate against one group. To protect  
 288 against this kind of decision maker, we need to give a guar-  
 289 antee such that based on the feature set it is not possible  
 290 to discriminate against one group. This corresponds to the  
 291 notion of adversarial fairness.

292 **Definition 3** (Adversarial fairness). *A feature set  $\mathcal{F}$  is con-*  
 293 *sidered to be adversarial fair w.r.t. the distribution  $P$  and*  
 294 *group fairness objective  $L^{\text{fair}}$ , if every classifier  $h \in H_{C_{\mathcal{F}}}$  is*  
 295 *group-fair. We define the worst case unfairness of a feature*  
 296 *set  $\mathcal{F}$  by  $U_{\text{adv}}(\mathcal{F}) = \max_{h \in H_{C_{\mathcal{F}}}} L_P^{\text{fair}}(h)$ .*

297 Furthermore, we consider an *accuracy-driven decision*  
 298 *maker*, who aims to label instances correctly and is agnostic  
 299 about fairness. For this kind of decision maker, we only  
 300 need to make sure that optimizing for correct classification  
 301 results in a fair classifier. In order to achieve this, we need  
 302 to provide a guarantee for the fairness of the Bayes optimal  
 303 classifier. We define this to be the accuracy-driven fairness of  
 304 a representation.

305 **Definition 4** (Accuracy-driven fairness). *We define*  
 306 *accuracy-driven fairness in two steps:*

307 1. *A feature set  $\mathcal{F}$  is considered to be accuracy-driven*  
 308 *fair w.r.t. the fairness objective  $L^{\text{fair}}$ , the distribution*  
 309  *$P$  and weight  $\alpha$  if every classifier  $h \in H_{C_{\mathcal{F}}}$  with*  
 310  *$L_P^\alpha(h) = \min_{h \in H_{C_{\mathcal{F}}}} L_P^\alpha(h)$  is group-fair. The accuracy-*  
 311 *driven (un)fairness w.r.t.  $L^{\text{fair}}$ ,  $P$  and  $\alpha$  is defined by*  
 312  *$U_{\text{acc}}^\alpha = L_P^{\text{fair}}(t_{P,F}^\alpha)$ .*

313 2. *We say a feature set  $\mathcal{F}$  is accuracy-driven fair w.r.t.  $L^{\text{fair}}$*   
 314 *and  $P$  if for every weight  $\alpha \in [0, 1]$  the feature set  $\mathcal{F}$  is*  
 315 *accuracy-driven fair w.r.t. to  $P$  and  $\alpha$ . Namely,*

$$U_{\text{acc}}(\mathcal{F}) = \max_{h \in H_{C_{\mathcal{F}}}, \alpha \in [0, 1]: L_P^\alpha(h) = \min_{h \in H_{C_{\mathcal{F}}}} L_P^\alpha(h)} L_P^{\text{fair}}(h).$$

316 We will see in Section 6, that in cases where the decision  
 317 maker does not have access to the distribution  $P$ , but only to a  
 318 labelled sample, this requirement is actually not sufficient for  
 319 guaranteeing that an accuracy-driven decision maker arrives  
 320 at a fair decision. In this case, we would need to further  
 321 require that any classifier who is close to the Bayes optimal  
 322 classifier in terms of accuracy is also close to the Bayes  
 323 optimal classifier in terms of fairness. This can be phrased as  
 324 another fairness criterion for a representation ( $\lambda$ -robustness,  
 325 see Appendix).

326 Lastly, we also consider a *fairness-driven decision maker*  
 327 who actively tries to find a fair and accurate decision rule,  
 328 while maintaining some accuracy guarantees. For such a deci-  
 329 sion maker a representation should allow for fair and accurate  
 330 decision rules. If a representation fulfills this requirement,  
 331 we call it fairness-enabling.

329 **Definition 5** ( $(\epsilon, \eta)$ -fairness-enabling representation). *A fea-*  
 330 *ture set  $\mathcal{F}$  is considered to be  $(\epsilon, \eta)$ -fairness-enabling w.r.t. a*  
 331 *fairness objective  $L^{\text{fair}}$ , if there exists a classifier  $h \in H_{C_{\mathcal{F}}}$*   
 332 *that such that  $L_P^\alpha(h) \leq \epsilon$  and  $L_P^{\text{fair}}(h) \leq \eta$ .*

333 Our discussion focuses primarily on the case of malicious  
 334 and indifferent decision makers. These notions of fair repre-  
 335 sentation can be defined with respect to any group-fairness  
 336 notion. In our paper we will mainly focus on the equalized  
 337 odds notion of fairness (Hardt, Price, and Srebro 2016).

## 4 Fairness of a feature set vs. fairness of a feature

340 In this section we discuss feature deletion and its impact on  
 341 the fairness of a representation. We mainly focus on accuracy-  
 342 driven fairness w.r.t. equalized odds. In particular, we show  
 343 that the deletion of a feature  $f$  can lead to an increase in  
 344 accuracy-driven fairness for some set of other given features  
 345  $\mathcal{F}$  and that the deletion of the *same* feature  $f$  can lead to a  
 346 decrease in accuracy-driven fairness for another set of other  
 347 available features  $\mathcal{F}'$ . We start by giving an example of the  
 348 above phenomena in a real-life scenario. Quite surprisingly,  
 349 we show this phenomena holds for a general class of fea-  
 350 tures that satisfy some non-triviality properties (That on the  
 351 one hand do not reveal too much information about group  
 352 membership and labels (non-disclosing), and on the other  
 353 hand does not reveal identity when label and group informa-  
 354 tion is given ( $k$ -anonymity (Samarati and Sweeney 1998))).  
 355 Lastly, we will discuss the impact of a single feature on  
 356 the fairness of a representation for adversarial fairness and  
 357 fairness-enabling representations.

### 4.1 University Admissions Example

358 Let us assume that a university has to deploy an automated  
 359 classifier to admit students to their program. They have access  
 360 to some features for each applicant. They would typically  
 361 have access to tens or maybe even hundreds of features, but  
 362 for illustration purposes we consider a simplified version,  
 363 with very few features.  
 364

365 1. One feature is group membership:

$$\text{Advantaged} = A.$$

$$\text{Disadvantaged} = D$$

366 2. Another is the feature which indicates if an applicant has  
 367 volunteered before or not.

$$\text{Volunteered previously} = V.$$

$$\text{Not volunteered previously} = N$$

368 3. The third is the feature which indicates if an applicant is  
 369 old or young.

$$\text{Old} = O.$$

$$\text{Young} = Y$$

370 One might be tempted to think that using feature 1 in the  
 371 classifier is unfair. We show that this is not always the case.  
 372 More specifically, we give two cases: In Situation 1, using the  
 373

368 group membership feature actually harms the fairness of the  
 369 classifier. In Situation 2 however, we see the opposite effect!  
 370 Contrary to popular intuition, using group membership in  
 371 this situation, actually helps the fairness of the classifier. Let  
 372 us see how this happens.

373 In both situations, an applicant needs to have a score of  
 374 greater than 0.5 to be admitted.

**Situation 1** The classifier has access to features 1 and 3.  
 The following are the cells with weights and scores in the  
 following format:

$$\text{Cell}(\text{Feature 1, Feature 3}) = (\text{Weight, Score})$$

$$C(A, O) = \left(\frac{1}{4}, 0.65\right)$$

$$C(D, O) = \left(\frac{1}{4}, 0.5\right)$$

$$C(A, Y) = \left(\frac{1}{4}, 0.65\right)$$

$$C(D, Y) = \left(\frac{1}{4}, 0.5\right)$$

375 One sees that feature 3 (age) has no effect on an individuals  
 376 ability to succeed in university. It is not hard to observe that  
 377 using only feature 3 implies that each applicant is admitted  
 378 to university. However, using both feature 1 and 3 results  
 379 in every member of group *A* being accepted, while each  
 380 member of group *D* is rejected. Easy calculations show us  
 381 that the former case (Unfairness = 0.46) is more fair than the  
 382 latter (Unfairness = 2), lending support to our intuition that  
 383 using feature 1 (group membership) is unfair.  
 384

385 **Situation 2** The classifier has access to features 1 and 2.  
 386 The following are the cells with weights and scores in the  
 387 following format:

$$\text{Cell}(\text{Feature 1, Feature 2}) = (\text{Weight, Score})$$

$$C(A, V) = \left(\frac{3}{8}, 0.8\right)$$

$$C(D, V) = \left(\frac{1}{8}, 0.47\right)$$

$$C(A, N) = \left(\frac{1}{8}, 0.2\right)$$

$$C(D, N) = \left(\frac{3}{8}, 0.51\right)$$

388 We see that people in the disadvantaged group have a much  
 389 smaller fraction of people who have volunteered. This is also  
 390 something observed in real life: students in underprivileged  
 391 schools do not have as much access to volunteering and  
 392 other opportunities. Hence, we observe a large difference  
 393 between the scores of people in the disadvantaged group who  
 394 have not volunteered, and advantaged group who have not  
 395 volunteered. This is not surprising: an advantaged person who

has not made use of the ample volunteering opportunities  
 available to them is less likely to succeed in university, than  
 a disadvantaged person who has not volunteered, because  
 the disadvantaged person had little to no opportunities to  
 volunteer.

Using only feature 2 (volunteering) groups disadvantaged  
 applicants with high scores ( $C(D, N)$ ), and advantaged  
 applicants with lower scores ( $C(A, N)$ ). Both these cells are  
 rejected as a consequence. Adding feature 1 (group member-  
 ship) into the mix allows the classifier to separate disadvan-  
 tagged applicants with high scores ( $C(D, N)$ ), and advantaged  
 applicants with lower scores ( $C(A, N)$ ). We can therefore  
 now accept applicants in ( $C(D, N)$ ) and reject applicants  
 from ( $C(A, N)$ ). Easy calculations show us that the former  
 case (Unfairness = 0.96) is less fair than the latter (Unfairness  
 = 0.43), contradicting our intuition that using feature 1 (group  
 membership) is unfair.

**Two "unfair" features together?** We see that using either  
 the feature of group membership (*A* and *D*), or of volun-  
 teering decreases the fairness w.r.t. using no feature. This  
 might make someone think that both the group membership  
 feature and volunteering features are unfair. Therefore, one  
 would think that if one is using one of these two features, one  
 should not use the other. Using one of them is unfair enough,  
 and using both of them together would be worse. However,  
 this intuition is proved false. Using both features results in a  
 more fair classifier compared to using just one of the features.

Overall, we see that the effect of adding features on the fair-  
 ness of the classifier depends on the set of other features avail-  
 able to the classifier. More specifically, we see that, in one  
 situation, adding the group membership feature decreased  
 fairness, and in another situation it increased fairness.

## 4.2 Result

As mentioned previously, this section shows that the deletion  
 of a feature can lead to either an increase or a decrease in  
 accuracy-enforced fairness. Moreover, the deletion of the  
*same* feature can have opposing effects on the accuracy-  
 enforced fairness of different representations. These results  
 hold for a general class of features that satisfy some non-  
 triviality properties.

We start by detailing the non-triviality feature properties  
 that suffice for this phenomenon.

### Non-Triviality properties

**Definition 6.** We define the following two non-triviality re-  
 quirements for a feature:

1. **Non-disclosing** We will call a feature non-disclosing  
 if this feature does not reveal too much information  
 about group-membership and label. That is, a feature  
*f* is non-disclosing if there are two distinct values  $y_1$   
 and  $y_2$ , such that *f* assigns each of these values to at  
 least one instance of each  $X_{A,0}, X_{A,1}, X_{D,1}, X_{D,0}$ . i.e.  
 $f^{-1}(y_1) \cap X_i \neq \emptyset$  and  $f^{-1}(y_2) \cap X_i \neq \emptyset$  for every  
 $X_i \in \{X_{A,0}, X_{A,1}, X_{D,1}, X_{D,0}\}$

449 2.  **$k$ -anonymity** A feature  $f$  is  $k$ -anonymous if knowing this  
 450 feature, group-membership and label, will only reveal identity  
 451 of an individuals up to a set of at least  $k$  individuals.  
 452 Namely, for every combination of value of this feature,  
 453 group membership and class label, there are either no in-  
 454 stances satisfying this combination or there are at least  $k$   
 455 many such instances.

456 **Overview of Result** We show that for every feature that  
 457 is 6-anonymous and non-disclosing there is a context (i.e.,  
 458 a set of features) for which deleting this feature increases  
 459 accuracy-driven fairness, and a context for which deleting  
 460 this feature decreases accuracy-driven fairness. This shows  
 461 that in most situations we cannot answer the question of how  
 462 fair a feature is without regarding the context of the other  
 463 features available to the classifier.

464 **Theorem 1.** (Context-relevance for fairness of features) For  
 465 every 6-anonymous non-disclosing feature  $f$ , there exists a  
 466 probability function  $P$  over  $X$  and feature sets  $\mathcal{F}$  and  $\mathcal{F}'$   
 467 such that:

- 468 • The accuracy-driven fairness w.r.t  $L^{EO}$ ,  $P$  and  $\alpha = 0.5$  of  
 469  $\mathcal{F} \cup \{f\}$  is greater than that of  $\mathcal{F}$ , i.e.

$$U_{acc}^\alpha(\mathcal{F} \cup \{f\}) < U_{acc}^\alpha(\mathcal{F})$$

470 Thus, deleting  $f$  in this context will increase unfairness.

- 471 • The accuracy-driven fairness w.r.t  $L^{EO}$ ,  $P$  and  $\alpha = 0.5$  of  
 472  $\mathcal{F}' \cup \{f\}$  is less than that of  $\mathcal{F}'$ , i.e.

$$U_{acc}^\alpha(\mathcal{F}' \cup \{f\}) > U_{acc}^\alpha(\mathcal{F}')$$

473 Thus, deleting  $f$  in this context will decrease unfairness.

474 We can also show that this is the case when natural condi-  
 475 tions between  $f$  and  $P$  hold, such as  $\{f\}$  being adversarial  
 476 fair w.r.t. to  $P$  and equalized odds. For a more general theo-  
 477 rem, details and proofs of this theorem we refer the reader to  
 478 the appendix.

### 479 4.3 The fairness of a feature for different notions 480 of fairness

481 We will now briefly discuss the fairness of a single feature for  
 482 the case of a malicious or a fairness-driven decision maker.  
 483 In contrast to the accuracy-driven case, adding features has a  
 484 monotone effect on the fairness in those cases. That is, adding  
 485 any feature in the malicious case, will only give the decision  
 486 maker more information and thus give the decision maker  
 487 more chances of discrimination. Similarly, in the fairness  
 488 driven case, any feature will only give the decision maker  
 489 another option for fair decision making.

490 **Observation 1.** For any feature  $f$  and any featureset  $\mathcal{F}$  we  
 491 have  $U_{adv}(\mathcal{F}) \leq U_{adv}(\mathcal{F} \cup \{f\})$ . Similarly, if the representa-  
 492 tion  $\mathcal{F}$  is  $(\epsilon, \eta)$ -fairness-enabling, the representation  $\mathcal{F} \cup \{f\}$   
 493 is also  $(\epsilon, \eta)$ -fairness-enabling.

494 *Proof.*  $\mathcal{H}_{\mathcal{C}_{\mathcal{F}}} \subset \mathcal{H}_{\mathcal{C}_{\mathcal{F} \cup \{f\}}}$ .  $\square$

495 The above observation of course also implies that the fair-  
 496 ness of  $\{f\}$  as a feature set is a bound on the fairness of  
 497 any representation that includes that feature  $\mathcal{F} \cup \{f\}$ , i.e.

498  $U_{adv}(\{f\})$  lower bounds  $U_{adv}(\mathcal{F} \cup \{f\})$ . However, we find  
 499 that in both the malicious and the fairness-driven case, the  
 500 impact of fairness of adding a single feature cannot be de-  
 501 termined without considering the context of other available  
 502 features. To illustrate this we make the following two obser-  
 503 vations.

504 **Observation 2.** For every distribution  $P$  and feature  $f$ , there  
 505 exists a feature set  $\mathcal{F}$ , such that adding  $f$  will not impact the  
 506 fairness of the distribution, e.g.  $U_{adv}(\mathcal{F}) = U_{adv}(\mathcal{F} \cup \{f\})$ .  
 507 Furthermore, there exist distributions  $P$ , features  $f$  and  $\mathcal{F}'$ ,  
 508 such that  $U_{adv}(\mathcal{F}') = 0$  and  $U_{adv}(\{f\}) = 0$ , but  $U_{adv}(\mathcal{F}' \cup$   
 509  $\{f\}) = 1$ .

510 **Observation 3.** For every distribution  $P$  and every feature  
 511  $f$ , there exists a feature set  $\mathcal{F}$ , such that  $\mathcal{F} \cup \{f\}$  is  $(\eta, \epsilon)$ -  
 512 fairness-enabling, if and only if  $\mathcal{F}$  is  $(\eta, \epsilon)$ -fairness-enabling.  
 513 Furthermore, there exists a distribution  $P$ , a feature  $f$  and  
 514 a feature set  $\mathcal{F}'$ , such that both  $\mathcal{F}'$  and  $\{f\}$  are not  $(\epsilon, \eta)$ -  
 515 fairness-enabling for any  $\epsilon, \eta < \frac{1}{2}$ , but such that  $\mathcal{F}' \cup \{f\}$  is  
 516  $(0, 0)$ -fairness-enabling.

517 While this section focused on fairness with respect to equal-  
 518 ized odds, we note that many of these results can be replicated  
 519 for other notions of fairness. For a more general version of  
 520 Observation 2, which takes into account other fairness notions,  
 521 like equalized odds, we will refer the reader to the  
 522 Appendix.

523 Thus, in general for any of motivation of a decision maker,  
 524 we cannot determine the impact of a single feature without  
 525 considering the context.

## 526 5 Can there be a generic fair representation?

527 In this part we address the question of whether there can be  
 528 a representation, that is fair independent of task. A task is  
 529 defined by a distribution  $P$  and consists of two components,  
 530 the marginal  $P_X$  and the ground truth labeling  $t$ .

531 We start by considering scenarios in which the marginals  
 532 shift between two tasks, e.g. two openings for different jobs,  
 533 for which a different pool of people would apply. Such a dis-  
 534 tribution shift can likely affect one group more than another  
 535 and would thus affect the classification rates of both groups  
 536 differently.

537 **Observation 4.** For every representation  $\mathcal{F}$  and every  
 538 ground truth labeling function  $t$ , that is not constant on ei-  
 539 ther group, there exists a distribution  $P$  with ground truth  
 540 labeling  $t$ , such that  $\mathcal{F}$  is adversarial unfair w.r.t. to  $P$  and  
 541  $L^{EO}$ . The same holds for any non-constant labeling function  
 542  $t$  for the (seemingly task-independent) demographic parity  
 543 notion of fairness.

544 Thus, when a shift in marginal occurs between tasks, any  
 545 fairness guarantee for previous tasks does not necessarily  
 546 imply a fairness guarantee for a new task. Now let us con-  
 547 sider situations, in which tasks share a marginal distribution  
 548 and only the ground-truth labelling differ between them. For  
 549 fairness-notions that are independent from the ground-truth,  
 550 like demographic parity, the question of the existence of gen-  
 551 eral fair representation under this additional restriction can be  
 552 answered to the positive; we only need demographic parity  
 553 in every cell to achieve that goal (see Appendix). However,

554 in the case of a ground-truth-dependent notions of fairness  
555 like equalized odds we show that there is no representation  
556 that guarantees fairness with respect to all ground-truth label-  
557 ings. This is true for all three decision objectives. We start  
558 by making the following observation for the fairness-driven  
559 case.

560 **Observation 5.** For every marginal distribution  $P_X$  and  
561 every representation  $\mathcal{F}$ , for which there exists a classi-  
562 fier  $h_{DP} \in \mathcal{H}_{\mathcal{C}_{\mathcal{F}}}$  with demographic parity that labels ex-  
563 actly half the mass as 1, i.e.  $P_X(h_{DP}^{-1}(1)) = \frac{1}{2}$ , there is a  
564 task given by distribution  $P$  with marginal  $P_X$ , such that  
565  $\min_{h' \in \mathcal{H}} L_P(h') \leq \frac{1}{4}$  such that any  $h \in \mathcal{H}_{\mathcal{C}_{\mathcal{F}}}$  based on the  
566 representation is either unfair with respect to  $P$  and equal-  
567 ized odds or has loss  $L_P(h) \geq \frac{1}{2}$ .

568 We get a similar result for the other two classification  
569 objectives. In the next theorem, we will consider a feature  
570  $f$  to be non-trivial with respect to a marginal  $P_X$ , if it  
571 splits the support of  $P_X$  according to something other than  
572 group-membership, i.e.  $P_X(f^{-1}(1) \cap A) \notin \{0, P(A)\}$  or  
573  $P_X(f^{-1}(1) \cap D) \notin \{0, P_X(D)\}$ .

574 **Theorem 2.** For every marginal distribution  $P_X$  and every  
575 representation  $\mathcal{F}$  containing at least one feature that is non-  
576 trivial w.r.t.  $P_X$ , there exists a task given by a distribution  $P$   
577 with marginal  $P_X$ , such that  $\mathcal{F}$  is neither accuracy-driven  
578 nor adversarial fair w.r.t.  $P$ .

579 We will now look at a slightly more restricted setting and  
580 analyse the case of multi-task learning, where instead of ask-  
581 ing for a representation that is fair for every task, we only  
582 consider fairness with respect to a fixed (finite) set of tasks  
583 that we want to learn. We again make the distinction of the  
584 three motivations for decision makers. For a decision maker  
585 who is accuracy-driven or fairness-driven, it is always possi-  
586 ble to have such a fair multi-task representation with respect  
587 to equalized odds. To see this, we consider a representation,  
588 that allows perfect classification for all considered (finitely  
589 many) tasks. Such a representation can be achieved by includ-  
590 ing the ground-truth labeling for each task as a feature.  
591 Since every classifier with perfect accuracy is also fair w.r.t.  
592 equalized odds, this representation would be considered fair  
593 in the accuracy-driven and the fairness-driven case. However,  
594 when the goal is to prevent unfair classification, i.e. if we  
595 want to address the malicious case, it is not always possible  
596 to find a representation that fulfills both fairness and accuracy  
597 requirements. In the next theorem, we show that there are  
598 fairness-incompatible tasks, i.e., tasks for which there is no  
599 joint representation which prevents unfair classification while  
600 enabling accurate prediction for both tasks.

601 **Definition 7 (Success-rate).** For a task given by distribution  
602  $P$ , the success-rate of a group  $G$  is defined by  $SR_G(P) =$   
603  $\frac{P(X_{G,1})}{P(G)}$ . We say a distribution  $P$  has equal success rates if  
604  $SR_A(P) = SR_D(P)$ .

605 **Theorem 3.** Given two tasks with distributions  $P_1$  and  $P_2$   
606 with the same marginal  $P_X = P_{1,X} = P_{2,X}$ , such that their  
607 success rates  $SR_A(P_1), SR_D(P_1), SR_A(P_2), SR_D(P_2) \in$   
608  $(0, 1)$  are not equal to 0 or 1 and have different ratios between  
609 groups, i.e.  $\frac{SR_A(P_1)}{SR_D(P_1)} \geq \max\{\frac{SR_A(P_2)}{SR_D(P_2)}, \frac{1-SR_A(P_2)}{1-SR_D(P_2)}\}$ , then  
610 there is no representation  $\mathcal{F}$ , such that

- $\mathcal{F}$  is worst case fair w.r.t.  $P_1$  and  $P_2$  and  $L^{EO}$  611
- $\mathcal{F}$  allows for perfect accuracy w.r.t. to  $P_1$ , i.e. there is a 612  
 $h \in \mathcal{H}_{\mathcal{C}_{\mathcal{F}}}$ , such that  $L_{P_1}(h) = 0$ . 613

614 Therefore, if the goal is to prevent discrimination from  
615 a possibly adversarial decision maker, while also enabling  
616 accurate prediction, each task requires its task-specific feature  
617 representation.

## 6 Effect of sampling on group fairness 618

619 In previous sections, we have examined the group-unfairness  
620 of feature based loss-minimizing classifiers,  $t_{P,F}^{\alpha}$ . However,  
621 in reality, a classifier is derived based on finite samples and  
622 can only approximate such loss minimizers. In this section,  
623 we turn our attention to the effect of sampling errors on the  
624 fairness of feature based learnt classifiers.

625 Our main finding on this issue is that while under some  
626 conditions the goals of maximizing accuracy and minimizing  
627 unfairness go hand in hand, when these conditions fail, sam-  
628 ple sizes that suffice for achieving high accuracy may still  
629 result in learnt classifiers whose unfairness is significantly  
630 high.

631 We consider learning through empirical risk minimization.  
632 The sampling method we consider is i.i.d. sampling of a  
633 selected sample size. All of our results are stated with respect  
634 to the following sample-based classifier.

635 **Definition 8 (empirical  $\alpha$ -threshold predictor).** We define  
636 the empirical  $\alpha$ -threshold predictor  $t_F^{\alpha}(S) : \mathcal{C}_{\mathcal{F}} \rightarrow \{0, 1\}$   
637 as a function that given access to a labeled sample  $S \subset$   
638  $(\mathcal{C}_{\mathcal{F}} \times \{0, 1\})^m$  of some sample size  $m$ , assigns label 1 to a  
639 cell if and only if more than  $\alpha$  fraction of samples in the cell  
640 have label 1, i.e.,

$$t_F^{\alpha}(S)(C) = \begin{cases} 1 & \text{if } |S \cap (C \times \{0, 1\})| > 0 \\ & \text{and } \frac{|S \cap C \times \{1\}|}{|S \cap C \times \{0, 1\}|} > \alpha \\ 0 & \text{otherwise} \end{cases}$$

641 Note that, if there is no sample point in a cell  $C$  the cell is  
642 assigned label 0.

643 We show that the empirical  $\alpha$ -threshold predictor is indeed  
644 an Empirical Risk Minimizer.

645 **Lemma 2.** For a sample  $S$ ,  $t_F^{\alpha}(S)$  is an Empirical Risk Min-  
646 imizer of the class  $\mathcal{H}_{cells}$ .

647 We can now state the main theorem of this section.

648 **Theorem 4.** There is a feature set  $\mathcal{F}$  such that for every  
649  $\delta > 0$  and every  $\epsilon > 0$ , there is a probability distribution  
650  $P$  over the domain  $X$  and a sample size  $M \in \mathbb{N}$  such that  
651 for every sample  $S \sim P^M$  the empirical risk minimizing  
652 classifier  $t_F^{\alpha}(S)$

- $\mathcal{F}$  is accuracy-driven fair w.r.t.  $P$  653
- $t_F^{\alpha}(S)$  has estimation error  $L_P^{\alpha}(t_F^{\alpha}(S)) - L_P^{\alpha}(t_{P,F}^{\alpha}) < \epsilon$  654  
and 655
- $t_F^{\alpha}(S)$  has group unfairness  $L_P^{fair}(t_F^{\alpha}(S)) = 1$  656

657 with probability at least  $1 - \delta$  over the sample generation.

658 Just as how the learnt classifier can have worse fairness  
659 than the Bayes classifier, the learnt classifier could also have  
660 better fairness than the Bayes classifier.

661 **Observation 6.** *There are situations in which a sample-*  
662 *based classifier is more fair than the respective Bayes classi-*  
663 *fier, with significant probability over the sampling.*

664 When the feature set and distribution satisfy some prop-  
665 erties, any classifier with loss close to the loss of  $t_{P,F}^\alpha$  also  
666 has unfairness close to the unfairness of  $t_{P,F}^\alpha$ . When these  
667 properties are satisfied, any accuracy guarantee we get for  
668 learning from a sample implies a fairness guarantee.

**Theorem 5.** *For every distribution  $P$ , every threshold  $\alpha$  and  
every feature set  $\mathcal{F}$ , for every feature based classifier  $h$*

$$L_P^{\text{fair}}(h) - L_P^{\text{fair}}(t_{P,F}^\alpha) \leq \lambda(L_P^\alpha(h) - L_P^\alpha(t_{P,F}^\alpha))$$

669 where  $\lambda = \max\left\{\frac{1}{\alpha}, \frac{1}{1-\alpha}\right\} \cdot \max_{(g,l) \in \{A,D\} \times \{0,1\}} \frac{1}{P(X_{g,l})}$

670 The property that allows for accuracy guarantee to imply a  
671 fairness guarantee is that all groups and labels are balanced.  
672 That is, no group-label set  $X_{g,l}$  has low measure. Based on  
673 an i.i.d. sample, we can detect if there is a set  $X_{g,l}$  with low  
674 measure. We can use the fraction of the sample belonging  
675 to the set  $X_{g,l}$  as an estimate of the probability measure of  
676  $X_{g,l}$ . With high probability over i.i.d. samples of size  $m$ , this  
677 estimate is accurate up to order of  $1/\sqrt{m}$ .

678 If we find that in an i.i.d. sample, a set  $X_{g,l}$  occurs with low  
679 frequency, we might be in a situation where an ERM classifier  
680 based on i.i.d. sampling would have high unfairness despite  
681 having good accuracy. We could prevent this by changing our  
682 sampling strategy. One way is by increasing the sample size  
683 of i.i.d. sampling. Another way is to shift away from i.i.d.  
684 sampling into a more targeted sampling and collect more data  
685 points from the low probability set.

## 7 Analysing Worst-case and Accuracy-driven fairness

686 In this section we characterize accuracy-driven and adversarial  
687 representation fairness w.r.t. the odds equality notion of  
688 classification fairness. We will start by introducing a prop-  
689 erty we call zero-group knowledge. it is aimed to prevent  
690 an adversary from inferring the group membership from the  
691 representation, when given access to the ground-truth labels.  
692 To ensure that an adversarial agent won't be able to infer  
693 group-membership, one would of course require the repre-  
694 sentation to have demographic parity. However, in situations  
695 where label information is correlated with group member-  
696 ship, demographic parity of all features will hurt classification  
697 accuracy. In such cases, zero-group-knowledge might be a  
698 better tool for concealing group-information.

699 We will then see that this property is closely related to  
700 adversarial fairness.

701 **Definition 9** (Zero-group-knowledge). *A feature set  $\mathcal{F}$  has  
702 zero-group-knowledge w.r.t.  
703 a distribution  $P$ , if for  $x \sim P$ , knowing the feature vec-  
704 tor  $F(x)$  will not reveal more information about the group  
705 membership  $G(x)$  than knowing just the ground truth,  $t(x)$ .  
706 Namely,  $G(x) \perp\!\!\!\perp F(x) | t(x)$ .*

707 It turns out that this property is equivalent to adversarial  
708 fairness with respect to equalized odds.

711 **Theorem 6.** *A feature set  $\mathcal{F}$  has zero-group knowledge w.r.t.*  
712  *$P$  if it has adversarial fairness w.r.t to  $P$  and the group-*  
713 *fairness measure  $L^{EO}$ .*

714 A similar observation has been made and shown by  
715 Zhang et al (Zhang, Lemoine, and Mitchell 2018), relating  
716 the optimization criteria for the goal of concealing group-  
717 membership and preventing unfair classification with respect  
718 to equalized odds in a representation learning setting with  
719 GANs.

720 We will now give a characterization of accuracy-driven and  
721 worse-case fairness in terms of the conditional distributions  
722 given label and group-membership. In the following we will  
723 denote the conditional probabilities given label  $l$  and group  
724  $G$  as  $P_{G,l}$ . We will see that a feature set is adversarial fair,  
725 if and only if the conditional probabilities are aligned. It has  
726 already been shown in (Zhao et al. 2019) that if conditional  
727 probabilities are aligned over a representation, every classifier  
728 based on that representation is fair. We go a step further  
729 here, by noting, that this is indeed a necessary condition for  
730 adversarial fairness.

731 **Theorem 7.** *A feature set  $\mathcal{F}$  is adversarial fair w.r.t. distri-*  
732 *bution  $P$  if and only if for each cell  $C \in \mathcal{C}_{\mathcal{F}}$  and for each*  
733  *$l \in \{0, 1\}$  we have  $P_{A,l}(C) = P_{D,l}(C)$ .*

734 We now give a similar statement for accuracy enforced  
735 fairness. Here, the same statement holds, if instead of con-  
736 sidering the probability distributions over the set of cells  $\mathcal{C}_{\mathcal{F}}$ ,  
737 we consider the set of cells that results from merging all cells  
738 of the same score:

739 **Definition 10** (Score-induced cells). *For a set of cells  $\mathcal{C}_{\mathcal{F}}$ ,*  
740 *the corresponding set of score-induced cells  $\mathcal{C}_{\mathcal{F}_{s_t}}$  is the set*  
741 *of cells that is obtained by merging all cells with the same*  
742 *score together. More formally, each feature set and scoring*  
743 *function, induce an equivalence relation  $\sim_{\mathcal{F},s_t}$ , such that*  
744  *$x \sim_{\mathcal{F},s_t} y$  if and only if there are cells  $C_x, C_y \in \mathcal{C}_{\mathcal{F}}$  such*  
745 *that  $x \in C_x, y \in C_y$  and  $s_t(C_x) = s_t(C_y)$ . The set  $\mathcal{C}_{\mathcal{F}_{s_t}}$  is*  
746 *then defined as the set of  $\sim_{\mathcal{F},s_t}$  equivalence classes.*

747 **Theorem 8.** *A feature set  $\mathcal{F}$  is accuracy-driven fair w.r.t.*  
748 *distribution  $P$  if and only if for each cell in the score-induced*  
749  *$C \in \mathcal{C}_{\mathcal{F}_{s_t}}$  and for each  $l \in \{0, 1\}$  we have  $P_{A,l}(C) =$*   
750  *$P_{D,l}(C)$ .*

751 We can now bound the unfairness in terms of accuracy-  
752 driven and adversarial fairness of a representation by the  
753 distribution distance of conditional probabilities. For this we  
754 take the  $\mathcal{H}$ -distance as introduced by (Ben-David et al. 2010).

755 **Definition 11** ( $\mathcal{H}$ -distance). *Given two distributions  $P$  and*  
756  *$Q$  over  $X$ , we define their  $\mathcal{H}$ -distance by*

$$d_{\mathcal{H}}(P, Q) = \sup_{1_B \in \mathcal{H}} |P(B) - Q(B)|,$$

757 where  $1_B$  denotes the indicator function of set  $B$ .

758 **Theorem 9.** *We can bound adversarial fairness and accu-*  
759 *racy enforced fairness of a feature set  $\mathcal{F}$  w.r.t.  $P$  and  $L^{EO}$*   
760 *by the  $d_{\mathcal{C}_{\mathcal{F}}}$ -difference and  $d_{\mathcal{C}_{\mathcal{F}_{s_t}}}$ -difference of conditional*  
761 *distributions respectively:*

$$U_{\text{adv}}(\mathcal{F}) \leq d_{\mathcal{H}_{\text{cells}}}(P_{A,1}, P_{D,1}) + d_{\mathcal{H}_{\mathcal{C}_{\mathcal{F}}}}(P_{A,0}, P_{D,0})$$



$$U_{acc}(\mathcal{F}) \leq d_{\mathcal{H}_{\mathcal{C}_{\mathcal{F}}, s_t}^{thres}}(P_{A,1}, P_{D,1}) + d_{\mathcal{H}_{\mathcal{C}_{\mathcal{F}}, s_t}^{thres}}(P_{A,0}, P_{D,0})$$

763 Furthermore, we can lower bound the adversarial fairness of  
764 a representation by

$$d_{\mathcal{H}_{cells}}(P_{A,l}, P_{D,l}) \leq U_{adv}(\mathcal{F}) \text{ for every } l \in \{0, 1\}$$

765 Note that for both bounds there exist probability distribu-  
766 tions  $P$  such that equality holds in all cases. Furthermore we  
767 note that since the  $\mathcal{H}$ -distance between two distributions can  
768 be estimated, if  $\mathcal{H}$  has a finite VC-dimension (Ben-David et al.  
769 2010), we can estimate both the upper and the lower bound  
770 with a sample size dependent on  $|\mathcal{C}_{\mathcal{F}}|$ , when given access to  
771 i.i.d. samples from  $P_{A,1}, P_{D,1}, P_{D,0}$  and  $P_{A,0}$  each.

## 772 7.1 Other fairness notions

773 So far we have demonstrated our approach mainly with re-  
774 spect to the odds equality notion of fair decision making. In  
775 this subsection we briefly discuss the implications on fair  
776 representations of using other notions of decision making  
777 fairness.

778 For some of the common such notions, it is easy to see  
779 that adversarial representation fairness is achievable and easy  
780 to characterize: For *demographic parity*, requiring cells rep-  
781 resentation induced cell have the same proportions for all  
782 groups is both necessary and sufficient for achieving adver-  
783 sarial fairness of the representation. This is probably an  
784 already known insight and we provide a formal statement  
785 and proof in the appendix.

786 However, some other acceptable notions of fairness do  
787 not always allow a adversarial fair representation. One such  
788 notion is *predictive rate parity*.

**Definition 12.** (*Predictive rate parity (PRP)*) A classifier  
 $h$  is considered PRP fair w.r.t. to a marginal data distri-  
bution  $P$  and true classification  $t$  if the random variable  
 $t(x)$  is independent of the group membership,  $G(x)$  given  
the classification  $h(x)$ . The respective measure of unfair-  
ness is given by the sum of differences in positive predictive  
rate and negative predictive rate, where the positive pre-  
dictive rate for a hypothesis  $h$  and group  $G$  is defined as  
$$PPR_G(h) = \frac{P(h^{-1}(1) \cap X_{G,1})}{P(h^{-1}(1) \cap G)}$$
 and the negative predictive  
rate is defined as  $NPR_G(h) = \frac{P(h^{-1}(0) \cap X_{G,0})}{P(h^{-1}(0) \cap G)}$ , i.e.

$$L_P^{Pred}(h) = |PPR_A(h) - PPR_D(h)| + |NPR_A(h) - NPR_D(h)|$$

789 Theorem 10 below shows that there are distributions for  
790 which there is no representation that has adversarial fairness  
791 with respect to predictive rate parity. In cases, where such  
792 a adversarial representation is achievable, however, we can  
793 characterize it by the following natural requirement on the  
794 representation.

795 **Definition 13.** A feature set  $\mathcal{F}$  has calibration parity w.r.t.  
796 a distribution  $P$  if for every cell  $C \in \mathcal{C}_{\mathcal{F}}$  both groups have  
797 equal success probability. Equivalently, one can say that for  
798 a random instance  $x \in P$  the ground truth labeling  $t(x)$  and  
799 the group membership  $G(x)$  are statistically independent,  
800 when the feature vector  $F(x)$  of  $x$  is known, i.e.  $G(x) \perp\!\!\!\perp$   
801  $t(x) | F(x)$ .

We can now state the main theorem of this subsection.

**Theorem 10.** A feature set  $\mathcal{F}$  has calibration parity w.r.t.  $P$   
if it has adversarial fairness w.r.t  $P$  and the group-fairness  
measure  $L^{Pred}$ . The other direction does not hold. In particu-  
lar, adversarial fairness w.r.t.  $P$  and  $L^{Pred}$  is only possible, if  
 $P$  has equal success rates for both groups.

## 808 8 Related Work

809 Most recent works on fair representation learning focus on  
810 learning fair intermediate representations that is then used  
811 by a learning algorithm to build a decision rule (Zemel et al.  
812 2013; Madras et al. 2018; Zhao et al. 2019; Adel et al. 2019;  
813 Zhang, Lemoine, and Mitchell 2018). Most of the papers in  
814 this area use the demographic parity notion of fairness (Ed-  
815 wards and Storkey 2016; Madras et al. 2018; Zemel et al.  
816 2013). Our work uses the equalized odds notion of fair-  
817 ness (Hardt, Price, and Srebro 2016). Some other papers on  
818 fair representation also use equalized odds fairness (Zhang,  
819 Lemoine, and Mitchell 2018; Beutel et al. 2017; Song et al.  
820 2019).

821 The connection between the motivation of the decision  
822 maker using the representation and the study of its fairness  
823 has been made before (Madras et al. 2018; Zhang, Lemoine,  
824 and Mitchell 2018). These papers identify two motivations.  
825 The first is malicious which is the intent to discriminate  
826 without regard for accuracy. The second is accuracy-driven  
827 which is the intent to maximize accuracy. Our results are  
828 expressed according to these motivations. Additionally, we  
829 provide results for the motivation of fairness-driven which is  
830 the intent to provide the most fair outcome while maintaining  
831 some level of accuracy.

832 Many algorithmic approaches to learning fair representa-  
833 tions have been proposed (Zhang, Lemoine, and Mitchell  
834 2018; Madras et al. 2018; Edwards and Storkey 2016; Song  
835 et al. 2019). Providing algorithmic approaches for fair repre-  
836 sentation learning is not our goal in this paper. We instead aim  
837 to answer questions about what we can and cannot achieve  
838 for a fair representation. One question in this spirit is the  
839 fairness-accuracy trade-off due to fair representations. This  
840 question has been studied for demographic parity (Zhao and  
841 Gordon 2019; McNamara, Ong, and Williamson 2019). A  
842 lower bound on the error of any classifier based on a represen-  
843 tation that guarantees demographic parity, when the success  
844 rates are different between groups, is shown in Zhao et al.  
845 (Zhao and Gordon 2019). The existence of situations when  
846 we can build a more accurate and more fair classifier based  
847 on the original representation than any classifier built using a  
848 learnt demographic parity satisfying representation is shown  
849 in McNamara et al. (McNamara, Ong, and Williamson 2019).

850 With demographic parity, there are works aiming to con-  
851 struct fair representations that can be used for different tasks  
852 (Creager et al. 2019; Madras et al. 2018; Oneto et al. 2019).  
853 Representations for different ground truths are considered in  
854 Madras et al. (Madras et al. 2018), and Oneto et al. (Oneto  
855 et al. 2019). Representations for different fairness notions  
856 - demographic parity w.r.t. different subgroups are consid-  
857 ered in Creager et al. (Creager et al. 2019). We prove that  
858 for a ground truth dependent fairness notion like equalized

859 odds, we cannot construct fair representations that transfer  
860 to different ground truths, other than in trivial situations. We  
861 also note that representations that are fair with respect to  
862 seemingly task-independent notions like demographic parity,  
863 are also not transferable to other tasks, if there is a change in  
864 marginals between tasks. Such a scenario is likely to occur in  
865 a job-application setting where the pool of applicants might  
866 be different between jobs.

867 The question of feature deletion has also been considered  
868 in real world examples, such as in the "ban the box" pol-  
869 icy which disallowed employers using criminal history in  
870 hiring decisions (Doleac and Hansen 2016). The effect of  
871 allowing or disallowing features on fairness has been studied  
872 before, for example in Grgić-Hlača et al. (Grgić-Hlaca et al.  
873 2018). However, the effect of a feature on fairness, has been  
874 discussed in isolation (as in (Grgić-Hlaca et al. 2018)). In  
875 contrast, we show that fairness of a feature should not be  
876 considered in isolation, but should also take into account the  
877 remaining features available.

878 Under-sampling has been identified as a source of unfair-  
879 ness (Chen, Johansson, and Sontag 2018; Balashankar and  
880 Lees 2019). Both these papers note the possibility of learning  
881 accurate classifiers that are unfair, even though there is a  
882 classifier that is both fair and accurate. This is done using an  
883 assumption on how generalization error behaves as a func-  
884 tion of the sample size in Chen et al. (Chen, Johansson, and  
885 Sontag 2018). In Balashankar et al. (Balashankar and Lees  
886 2019), this is done by noting that the lower bound on sample  
887 complexity for good generalization across multiple groups is  
888 strictly higher than the upper bound on sample complexity  
889 for good overall generalization. We prove that this situation  
890 of learning accurate but unfair classifiers due to sampling can  
891 occur, even with significant probability over the sampling. In-  
892 creasing sample sizes and performing targeted sampling have  
893 been suggested as ways to prevent unfairness from sampling  
894 (Chen, Johansson, and Sontag 2018; Balashankar and Lees  
895 2019; Zelaya, Missier, and Prangle 2019).

## 896 9 Conclusion

897 In this paper we introduced a general taxonomy of notions of  
898 fair representation, taking into consideration both different  
899 objectives of decision makers using the representation, and  
900 different group fairness notions. Within this taxonomy we  
901 addressed several questions about fair representations.

902 We first considered the question of "fairness of a feature",  
903 which has been used in legal scenarios. We showed that for  
904 notions of decision-making fairness other than demographic  
905 parity, the fairness of a single feature is an ill defined notion.  
906 Namely, the impact of a feature on the fairness of a decision  
907 cannot be determined without considering the other features  
908 of the representation.

909 Next, we addressed the existence of generic fair represen-  
910 tations and of fair transfer learning. We show that even seem-  
911 ingly task-independent fairness notions like demographic  
912 parity are vulnerable to shifts in marginals between tasks. We  
913 have also proved that for the odds equality notion of fairness,  
914 there is no fixed representation that is fair with respect to all  
915 ground-truth labelings, even when the marginal distribution

is fixed. Furthermore, when representations are aimed to pre- 916  
vent an adversary from unfair decision making, we show that 917  
there exist fairness incompatible tasks, i.e. tasks that do not 918  
allow an adversarial fair representation which also enables 919  
accurate predictions of class labels for both tasks. These in- 920  
sights stand in contrast to the impression arising from recent 921  
papers (Madras et al. 2018; Oneto et al. 2019) that claim to 922  
learned transferable fair decisions. Such results are inevitably 923  
limited to ground-truth oblivious notions of fairness (such as 924  
demographic parity). 925

926 Following this, we address the issue of unfairness due 927  
to under-sampling. We analyze this issue as a property of 928  
the distribution and show that it differs from the unfairness 929  
of a representation due to an unfair Bayes classifier. We 930  
show that there are distributions and representations such 931  
that an empirical risk minimizing algorithm is unfair with 932  
high probability over the sample generation, while also being 933  
accurate with high probability. We show that this can occur 934  
even if the Bayes classifier based on the representation is 935  
fair. This suggests that further stability criteria need to be 936  
taken into account when considering fair representations for 937  
accuracy-driven decision makers who base their decision on 938  
a finite sample. As a solution for ensuring fairness we suggest 939  
importance sampling.

940 Lastly, we provide a characterization of accuracy-driven 941  
and adversarial fairness of a representation with respect to 942  
the equalized odds notion of fairness in terms of differences 943  
between conditional distribution given labels and groups. 944  
Furthermore we show that some fairness notions, like pre- 945  
dictive rate parity, do not always allow an adversarially fair 946  
representation. 947

unsrt

## 948 References

- 949 Adel, T.; Valera, I.; Ghahramani, Z.; and Weller, A. 2019. 950  
One-network adversarial fairness. In *AAAI*. 951
- Balashankar, A.; and Lees, A. 2019. Fairness Sample Com- 952  
plexity and the Case for Human Intervention. *arXiv preprint* 953  
*arXiv:1910.11452* .
- Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, 954  
F.; and Vaughan, J. W. 2010. A Theory of Learning from 955  
Different Domains. *Mach. Learn.* 79(1–2). 956
- Beutel, A.; Chen, J.; Zhao, Z.; and Chi, E. H. 2017. Data 957  
Decisions and Theoretical Implications when Adversarially 958  
Learning Fair Representations. *CoRR* abs/1707.00075. 959
- Chen, I.; Johansson, F. D.; and Sontag, D. 2018. Why is my 960  
classifier discriminatory? In *NIPS*. 961
- Corbett-Davies, S.; and Goel, S. 2019. The measure and 962  
mismeasure of fairness: A critical review of fair machine 963  
learning. *CoRR* abs/1808.00023. 964
- Creager, E.; Madras, D.; Jacobsen, J.-H.; Weis, M.; Swersky, 965  
K.; Pitassi, T.; and Zemel, R. 2019. Flexibly Fair Representa- 966  
tion Learning by Disentanglement. In *ICML*. 967
- Doleac, J. L.; and Hansen, B. 2016. Does “ban the box” help 968  
or hurt low-skilled workers? Statistical discrimination and 969  
employment outcomes when criminal histories are hidden. 970  
Technical report, National Bureau of Economic Research. 971

972 Edwards, H.; and Storkey, A. J. 2016. Censoring Representa-  
973 tions with an Adversary. In *ICLR*.

974 Grgic-Hlaca, N.; Bilal Zafar, M.; P. Gummadi, K.; and Weller,  
975 A. 2018. Beyond Distributive Fairness in Algorithmic Deci-  
976 sion Making: Feature Selection for Procedurally Fair Learn-  
977 ing. In *AAAI*.

978 Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of oppor-  
979 tunity in supervised learning. In *NIPS*.

980 Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018.  
981 Learning Adversarially Fair and Transferable Representa-  
982 tions. In *ICML*.

983 McNamara, D.; Ong, C. S.; and Williamson, R. C. 2019.  
984 Costs and benefits of fair representation learning. In *Proceed-*  
985 *ings of the 2019 AAAI/ACM Conference on AI, Ethics, and*  
986 *Society*, 263–270.

987 Oneto, L.; Donini, M.; Maurer, A.; and Pontil, M. 2019.  
988 Learning fair and transferable representations. *arXiv preprint*  
989 *arXiv:1906.10673* .

990 Samarati, P.; and Sweeney, L. 1998. Protecting Privacy when  
991 Disclosing Information: k-Anonymity and Its Enforcement  
992 through Generalization and Suppression. Technical report.

993 Song, J.; Kalluri, P.; Grover, A.; Zhao, S.; and Ermon, S.  
994 2019. Learning controllable fair representations. In *The*  
995 *22nd International Conference on Artificial Intelligence and*  
996 *Statistics*, 2164–2173.

997 Zelaya, V.; Missier, P.; and Prangle, D. 2019. Parametrised  
998 data sampling for fairness optimisation. *KDD XAI* .

999 Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C.  
1000 2013. Learning fair representations. In *ICML*.

1001 Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mit-  
1002 igating Unwanted Biases with Adversarial Learning. In  
1003 *AAAI/ACM Conference on AI, Ethics, and Society*.

1004 Zhao, H.; Coston, A.; Adel, T.; and Gordon, G. J. 2019.  
1005 Conditional Learning of Fair Representations. *CoRR*  
1006 abs/1910.07162.

1007 Zhao, H.; and Gordon, G. J. 2019. Inherent Tradeoffs in  
1008 Learning Fair Representations. *CoRR* abs/1906.08386.  
otherwise