Don't Throw it Away! The Utility of Unlabeled Data in Fair Decision Making

Miriam Rateike*
MPI for Intelligent Systems
Tübingen, Germany
Saarland University
Saarbrücken, Germany
mrateike@tue.mpg.de

Ayan Majumdar*
MPI for Software Systems
Saarland University
Saarbrücken, Germany
ayanm@mpi-sws.org

Olga Mineeva
ETH Zürich
Zürich, Switzerland
MPI for Intelligent Systems
Tübingen, Germany
omineeva@ethz.ch

Krishna P. Gummadi
MPI for Software Systems
Saarbrücken, Germany
gummadi@mpi-sws.org

Isabel Valera
Saarland University
MPI for Software Systems
Saarbrücken, Germany
ivalera@cs.uni-saarland.de

ABSTRACT

Decision making algorithms, in practice, are often trained on data that exhibits a variety of biases. Decision-makers often aim to take decisions based on some ground-truth target that is assumed or expected to be unbiased, i.e., equally distributed across socially salient groups. In many practical settings, the ground-truth cannot be directly observed, and instead, we have to rely on a biased proxy measure of the ground-truth, i.e., *biased labels*, in the data. In addition, data is often *selectively labeled*, i.e., even the biased labels are only observed for a small fraction of the data that received a positive decision. To overcome label and selection biases, recent work proposes to learn stochastic, exploring decision policies via i) online training of new policies at each time-step and ii) enforcing fairness as a constraint on performance. However, the existing approach uses only labeled data, disregarding a large amount of unlabeled data, and thereby suffers from high instability and variance in the learned decision policies at different times. In this paper, we propose a novel method based on a variational autoencoder for practical fair decision-making. Our method learns an unbiased data representation leveraging both labeled and unlabeled data and uses the representations to learn a policy in an online process. Using synthetic data, we empirically validate that our method converges to the *optimal (fair) policy* according to the ground-truth with low variance. In real-world experiments, we further show that our training approach not only offers a more stable learning process but also yields policies with higher fairness as well as utility than previous approaches.

*Both authors contributed equally to this research.

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533199

CCS CONCEPTS

• **Computing methodologies → Machine learning algorithms**; **Online learning settings**; • **Social and professional topics**;

KEYWORDS

fairness, decision making, label bias, selection bias, variational autoencoder, fair representation

ACM Reference Format:
Miriam Rateike, Ayan Majumdar, Olga Mineeva, Krishna P. Gummadi, and Isabel Valera. 2022. Don't Throw it Away! The Utility of Unlabeled Data in Fair Decision Making. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3531146.3533199

1 INTRODUCTION

The extensive literature on fair machine learning has focused primarily on studying the fairness of the predictions by classification models [1, 17, 23, 56, 57] deployed in critical decision-making scenarios. Consider a university admissions process where the goal is to admit students based on their *true potential*, which may not be directly observable. We assume that a student's ground-truth potential is independent of and unbiased by, their assignment to various socially salient groups - defined by sensitive characteristics (race, gender) protected by anti-discrimination laws [7]. That is, we hold it as self-evident that students of different socially salient groups are endowed with a similar (equal) distribution of potential. In practice, ground-truth potential cannot be measured directly and remains unobserved. Instead, we rely on *proxy labels*, which we assume to contain information about the ground truth. However, due to structural discrimination, these proxy labels are often biased measures of ground truth. For example, prevailing societal discrimination may result in students with similar ground-truth potential but different assigned genders having a very different distribution of university grades (proxy labels). This phenomenon is termed *label bias* [52] and has been studied extensively in [1, 23, 57, 58]. These fairness studies assume that independent and identically distributed (i.i.d.) labeled data is available for training. However, in decision-making scenarios, data may also suffer from *selection* *bias* [36]. That is, we only observe the labels of a small fraction of the data which received positive decisions. For example, we observe (biased) university grades of only the admitted students, resulting in biased, non-i.i.d. labeled data.

In decision-making scenarios affected by *both label bias and selection bias*, Kilbertus et al. [31] show that to learn the optimal policy, it is necessary to move from learning fair predictions (e.g., predicting grades) to learning fair decisions (e.g., deciding to admit students). To tackle label bias, the authors introduce fairness constraints in the optimization problem. To address selection bias, they propose to learn *stochastic, exploring* decision policies in an online learning process, where a new decision policy is learned at each time-step. To get unbiased loss estimates from non-i.i.d. labels, the authors further rely on inverse propensity scoring (IPS) [27].

However, the approach ignores unlabeled data. A large fraction of data in the learning process may remain unlabeled due to receiving a negative decision (e.g., students denied admission). Using only labeled data, the approach suffers from high instability and variance in the learning process. In particular, i) the method may give very different outcomes to the same individual, depending on the random initializations of the learning process, and ii) the method may give the same individual very different outcomes at different points in time.

In this paper, we propose a novel online learning process for fair decision-making that leverages both labeled and unlabeled data. Our method learns fair representations of all data using latent variable models in an attempt to capture the unobserved and unbiased ground truth information. In turn, these representations are used to learn a policy that approximates the optimal fair policy (according to the unobserved ground truth). Importantly, as shown in our experiments, our approach leads to a stable, fair learning process, achieving decision policies with similar utility and fairness measures across time and training initializations.

Our primary contributions in this paper are listed below:

(1) We propose a novel two-phase decision-making framework that utilizes both labeled and unlabeled data to learn a policy that converges to the optimal (fair) policy with respect to the unobserved and unbiased ground truth.
(2) We present a novel policy learning framework, FairAll that relies on a VAE (a latent variable model) architecture to significantly reduce the need for bias correction of selective labeling.
(3) Through theoretical analyses and empirical evaluation on synthetic data, we show that the VAE from our FairAll framework is able to learn an unbiased data representation that captures information from the ground truth.
(4) Through extensive evaluations based on real-world data, we show that FairAll, compared to prior work, offers a significantly more effective and stable learning process, achieving higher utility and fairness.

1.1 Related Work

Fair Classification. There exists a variety of approaches for fair classification to tackle biased labels. *In-processing methods* optimize for correct predictions under additional fairness constraints [1, 17, 56, 57]. This requires formulating differentiable fairness constraints and often lead to unstable training [13]. *Pre-processing methods* instead utilize representation learning first to learn a fair data representation. This representation is then used for downstream predictive tasks [58]. Different methods for fair representation learning have been brought forward, including variational autoencoders (VAEs) [38, 40], normalizing flows [5], and generative adversarial networks [53]. To enforce independence between the learned representation and the sensitive attribute, some methods condition deep generative models on the sensitive features [14, 21, 39, 40], revert to disentanglement [14], perform adversarial training [21, 39, 50] or add regularization, like Maximum-Mean-Discrepancy [21, 38]. While most work on fair representation learning focuses on satisfying group fairness notions [14, 38], some have also considered individual fairness [46] and counterfactual fairness [21]. Recently, contrastive learning for fair representations has attracted much attention [41]. However, it requires the definition of a similarity measure and meaningful data augmentations. This is non-trivial, especially for tabular data. While some recent work [4] exists, further research is needed.

Although all of the works above use fully labeled training data, some have studied fair classification in the presence of partially labeled data [38, 59, 60]. Further, all of these works assume access to a *biased proxy label* and not the ground truth. Zafar et al. [55] considered analyzing fairness notions separately, assuming access to a ground-truth label. Further, the notions of biased observed proxies and unbiased, unobserved ground-truth (denoted as construct spaces) were discussed in [16, 19]. Note that all of these studies also assume i.i.d. data. But, in most real-world scenarios, the semi-labeled data is not i.i.d. (selection bias [36]). Wick et al. [52] perform an initial fairness-accuracy analysis of classifiers with respect to label and selection bias. Our work, similar to [31] aims to tackle both label and selection bias while transitioning from a static classification setting to online decision-making.

Fair Online Decision Making. Recent works [8, 31] have started exploring fairness in online decision-learning processes in the presence of partially labeled non-i.i.d. data. In such settings, convergence to the optimal policy requires exploration and stochastic policies [31]. Kilbertus et al. [31] use an extra fairness constraint in the loss to trade-off between utility and fairness. Additionally, they correct for selection bias in training using *inverse propensity scoring* (IPS) [27] on the entire loss function, which, unfortunately, can introduce additional variance. Bechavod et al. [8] derive an oracle-efficient bandit algorithm to learn an accurate policy while explicitly controlling the exploration-exploitation trade-off, and thus the variance. However, both approaches disregard a major portion of the data that receives the negative decision and remain unobserved. We show how this data contain useful information about the underlying data distribution. In our approach, we posit utilizing this unlabeled data to reduce the need for IPS. We empirically validate how this helps in faster convergence to an optimal decision policy while providing high utility and fairness during training.[1]

[1]In Section 5 we compare our method to [31]. Note that [8] is a theoretical work that provides neither experimental results nor an implementation and thus prevents us from comparing to them as the baseline.

(a) Ground truth generative process.

(b) FairAll latent variable model.

Figure 1: (a) Ground truth data generative process and (b) our FairAll generative (solid) and inference (dashed) models. Sensitive attribute S, non-sensitive attribute $\mathbf{X}$, proxy label $\tilde{Y}$, ground truth label Y and decision D. Observed (unobserved) random variables grey (white).

2 BACKGROUND AND PROBLEM SETTING

Let us consider a university admission decision-making process inspired by [35], which we will use as a running example. We use uppercase letters for random variables and lowercase letters for their assignments. With p, we optionally refer to a probability distribution or a probability mass function. Let S be a random variable indicating a *sensitive attribute* of an individual describing their membership in a socially salient group (e.g., gender). For simplicity we assume binary $S \in \{-1, 1\}$. Let $\mathbf{X} \in \mathbb{R}^n$ be a set of n *non-sensitive features* that are observed (e.g., high school grades), and may be influenced by S. The university aims to take an admission decision $D \in \{0, 1\}$ based on a *ground truth target* Y (e.g., intellectual potential) [16, 19]. For simplicity we assume $Y \in \{0, 1\}$. Importantly, throughout this paper, we assume that $Y \perp S$ (e.g., potential is equally distributed across social groups), such that an optimal policy decides $D \perp S$.[2]

2.1 Label Bias and Selection Bias

Label Bias. In practice, the ground truth Y often remains unobserved (as it cannot be directly measured). Instead, as shown in Figure 1a, we observe a different label $\tilde{Y}$ (e.g., semester grades) that is assumed to contain information on Y along with measurement noise. We refer to this label as the *proxy label*. For simplicity, we assume $\tilde{Y} \in \{0, 1\}$. A data-generative process exhibits *label bias*, if $\tilde{Y} \not\perp S$, i.e., the proxy target is biased by the sensitive attribute [52]. For example, the same potential may result in higher grades for one demographic group over another due to existing structural discrimination. Figure 1a highlights our assumed data generative process with biased labels. Recall that we aim to take decisions according to $Y \perp S$. However, in the biased label scenario, both $\mathbf{X}$ and $\tilde{Y}$ are biased by S. A policy that maps $\mathbf{X}$ (and potentially S) to $\tilde{Y}$ will thus – in the absence of fairness constraints – take biased decisions.

Selection Bias. In practice, algorithms often also need to learn from partially labeled data, where labels $\tilde{Y}$ are observed only for a particular (usually positive) decision. This is called the *selective labels problem* [36]. For example, a university only knows whether a student gets good semester grades if it accepts the student in the first place. Let these decisions be taken according to policy π, which may be biased and not optimal. For example, for an individual with features $(\boldsymbol{x}, s)$, a decision d may be taken according to $d \sim \pi(\boldsymbol{x}, s)$.[3] Then a labeled data point $(\boldsymbol{x}, s, \tilde{y})$ observed under a policy π is *not an i.i.d. sample* from the true distribution $p(\mathbf{X}, S, \tilde{Y})$. Instead, the data is sampled from the distribution induced by the probability of a positive decision $\pi(d = 1 \mid \boldsymbol{x}, s)$ such that: $p_\pi(\mathbf{X}, S, \tilde{Y}) \propto p(\tilde{Y} \mid \mathbf{X}, S)\pi(D = 1 \mid \mathbf{X}, S)p(\mathbf{X}, S)$ [31].

Corbett-Davies et al. [12] have shown that deterministic decisions (e.g. taken by thresholding) are optimal with respect to $\tilde{Y}$ for i.i.d. data. However, Kilbertus et al. [31] demonstrated that, if labels are not i.i.d., we require *exploration*, i.e., stochastic decision policies. Such policies map features to a strictly positive distribution over D. This implies that the probability of making a positive decision for any individual is never zero. Exploring policies are trained in an *online* fashion, where the policy is updated at each time step t as π_t. Moreover, we typically learn a policy from labeled data by minimizing a loss that is a function of the revealed labels. However, if labels are non-i.i.d., it is necessary to perform bias correction to get an unbiased loss estimate. A common technique for such bias correction is *inverse propensity score* (IPS) weighting [27]. It divides the loss for each labeled datum by the probability with which it was labeled, i.e., received a positive decision under policy π. However, this bias correction may lead to high variance in the learning process, when this probability is small.

2.2 Measures of Interest: Utility, Fairness, and Their Temporal Stability

Assuming an incurred cost c for every positive decision (e.g., university personnel and facility costs) [12], the decision-maker aims to *maximize its profit* (revenue − costs), which we call *utility*. We define utility $U = D(Y - c)$ as a random variable that can take on three values $U \in \{-c, 1 - c, 0\}$ depending on decision D. A correct positive decision results in a positive profit of $1 - c$ (admitting students with high potential leads to more success and funding), an incorrect positive decision results in a negative profit $-c$ (sunk facility costs), and a negative decision (rejecting students) in zero profit. The utility of a policy is then defined as the expected utility U with respect to population $p(\mathbf{X}, S, Y)$ and policy π:

[2] Note, depending on which label Y refers to, $Y \perp S$ may not always hold in practice. See Section 6 for a discussion of this assumption.

[3] A policy always takes as input features of an individual. Here, we assume the features to be $(\mathbf{X}, S)$. However, they could also be only $\mathbf{X}$ or some feature representation.

Definition 2.1 (Utility of a policy [31]). *Given utility as a random variable $U = D(Y-c)$, we define the utility* UT *of a policy* π *as the expected overall utility* $\mathrm{UT}(\pi) := \mathbb{E}_{\boldsymbol{x},s,y\sim p(\mathrm{X},S,Y)}[\pi(D=1 \mid \boldsymbol{x}, s)(Y-c)]$, *where decision and label are* $D, Y \in \{0,1\}$*, and* $c \in (0,1)$ *is a problem specific cost of taking a positive decision.*

Note, we defined UT with respect to ground truth target Y. However, as mentioned above, in most practical settings, we only observe proxy $\tilde{Y}$ and can thus only report $\widetilde{\mathrm{UT}}$, i.e., the expected utility measured with respect to proxy $\tilde{Y}$.

As detailed above, we are interested in taking decisions according to Y, where $Y \perp S$ (e.g., potential is equally distributed across sensitive groups). A policy that takes decisions based on Y satisfies *counterfactual fairness* [35] and *demographic parity* [17] . This follows directly from the fact that Y is a non-descendant of S. Any policy π that is a function of the non-descendants of S (namely Y) is demographic parity and counterfactually fair [35].[4] The notion of DP fairness for a policy π requires the proportion of decision d to be the same across all social groups :

Definition 2.2 (Demographic Parity Unfairness of a Policy [17]). *We define the demographic parity unfairness (*DPU*) of a policy* π *with respect to sensitive attribute* $S \in \{-1,1\}$ *and decision* $D \in \{0,1\}$*:*

$$\begin{aligned}\mathrm{DPU}(\pi) = \big|&\mathbb{E}_{\boldsymbol{x}\sim p(\mathrm{X}|S=1)}[\pi(D=1 \mid \boldsymbol{x}, S=1)] \\ &- \mathbb{E}_{\boldsymbol{x}\sim p(\mathrm{X}|S=-1)}[\pi(D=1 \mid \boldsymbol{x}, S=-1)]\big|\end{aligned}$$

Correspondingly, a policy is counterfactually fair if it assigns the same decision to an individual in the observed (or, factual) world as well as in a *counterfactual world*, in which the individual belongs to a different sensitive group[5].

Definition 2.3 (Counterfactual Unfairness of a Policy [35]). *The counterfactual unfairness (*CFU*) of policy* π *with respect to a factual individual belonging to* $S = s$ *with features* $\boldsymbol{x}^F$ *and the decision* $D \in \{0,1\}$ *can be defined as:*

$$\begin{aligned}\mathrm{CFU}(\pi) =&\mathbb{E}_{(\boldsymbol{x}^F,s)\sim p(\mathrm{X},S),\,\boldsymbol{x}^{CF}\sim p(\mathrm{X}|\boldsymbol{x}^F, do(S=s'))} \\ &\left|\pi(D=1 \mid \boldsymbol{x}^F, s) - \pi(D=1 \mid \boldsymbol{x}^{CF}, s')\right|\end{aligned}$$

Here, $\boldsymbol{x}^F$ *refers to the non-sensitive features of an individual in the factual world with sensitive attribute* s*, and* $\boldsymbol{x}^{CF}$ *refers to the non-sensitive features of the same individual in a counterfactual world, where its sensitive attribute is* s' *with* $s' \neq s$.

Note, from [35] that satisfying counterfactual fairness implies satisfying demographic parity but not vice-versa. Further, counterfactual analysis requires hypothetical interventions on S and exact knowledge of the causal generation process. While estimation techniques for real-world data exist [30, 47], in this paper, we only analyze for synthetic data (with access to the true exogenous variables and the structural equations). See Appendix C.4 for more details.

The above metrics allow assessing the performance of one particular policy. However, the online policy learning process outputs, over T training steps, the set of policies $\Pi_{t=1}^{T} := \{\pi_1 \dots \pi_T\}$. Assume we wish to stop the learning process from time t_1. Can we reliably deploy any policy $\pi_{t\geq t_1}$? Inspired by prior work on temporal fairness [10, 22], we propose a new notion of *temporal variance* (TV) for a policy learning process. TV indicates how much a metric M (e.g., utility, fairness) varies for the set of policies across some time interval $[t_1, t_2]$.

Definition 2.4 (Temporal Variance of a Policy Learning Process). *We define the temporal variance (*TV*) of the outcome of a policy learning process* $\Pi_{t_1}^{t_2} := \{\pi_{t_1} \dots \pi_{t_2}\}$ *in time interval* $[t_1, t_2]$ *with respect to metric* M *as:*

$$\mathrm{TV}_M(\Pi_{t_1}^{t_2}) = \sqrt{\frac{1}{t_2 - t_1}\sum_{t=t_1}^{t_2}\left[(\mathrm{M}(\pi_t) - \mu_M)^2\right]}$$

where $\mu_M = \frac{1}{t_2-t_1}\sum_{t=t_1}^{t_2}\mathrm{M}(\pi_t)$ *denotes the temporal average for the metric* M *over the time interval* $[t_1, t_2]$.

High TV denotes an unstable learning process, where policies of different time steps achieve different utility and fairness levels for a fixed group of people. For example, policy π_{t_1} may treat the same group of individuals very different compared to π_{t_1+1}. A low TV on the other hand indicates a stable learning process, where policies of different time steps achieve similar utility and fairness levels. Hence, it is safe to stop the learning process any time after t_1. Note, μ_M measures the average metric value (e.g. utility, fairness) over the time interval $t = [t_1, t_2]$.

2.3 Variational Autoencoder

Deep generative models (DGMs), like Variational Autoencoders (VAEs) [33, 45], Normalizing Flows, [44] and Generative Adversarial Networks [20] are latent variable models (LVMs) that *estimate complex data distributions* by capturing hidden structures in the latent space Z.

VAE is one of the most prominent DGMs. It jointly learns a probabilistic generative model $p_\theta(\mathbf{X} \mid \mathbf{Z})$ (*decoder*) and an approximate posterior estimator $q_\phi(\mathbf{Z} \mid \mathbf{X})$ (*encoder*). Encoder and decoder are parameterized by neural networks. As the marginal likelihood $p(\boldsymbol{x}) = \int p(\boldsymbol{x}, z)dz$ is intractable, a VAE is trained by maximizing the evidence lower bound (ELBO) of the observations $\boldsymbol{x} \sim p(\mathrm{X})$, consisting of the expected log-likelihood and the posterior-to-prior KL divergence:

$$\begin{aligned}\log p(\boldsymbol{x}) \geq& \underbrace{\mathbb{E}_{z\sim q_\phi(\mathrm{Z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x} \mid z)\right]}_{\text{Exp. log-likelihood}} - \underbrace{\mathrm{KL}\left(q_\phi(\mathrm{Z} \mid \boldsymbol{x}) \| p(\mathrm{Z})\right)}_{\text{KL divergence}} \\ =& \mathrm{ELBO}(\theta, \phi; \boldsymbol{x})\end{aligned}$$

3 LEARNING TO DECIDE FAIR

Let us assume that the data $p(\mathbf{X}, S, Y)$ has a generative process as shown in Figure 1a[6]. Recall that a decision-maker ideally aims to take decisions d according to the unbiased ground truth Y, i.e., $d \sim p(Y)$[7] [31]. However, in practice, Y remains unobserved. Consider

[4]Since $Y \perp S$, a policy that decides according to Y also satisfies equal opportunity [35].
[5]For example, where the individual had been growing up with a different sensitive identity like gender.
[6]Note, Figure 1a is the same as the causal model presented as the *Scenario 3: University success* in [35].
[7] Note, this is formulated abusing notation for simplicity. Here, decision D is a deterministic function of Y, i.e., for a data point $d = y$, where $y \sim \pi(Y)$.

for now a setting with label bias but no selection bias. We have access to i.i.d. samples from the underlying distribution and the observable label is a biased proxy $\tilde{Y}$. As per Figure 1a, observed features $\mathbf{X}$ and proxy labels $\tilde{Y}$ both contain information about Y, but are biased by S (label bias), i.e., $\mathbf{X}\not\perp S$, $\tilde{Y}\not\perp S$. Hence, a policy that takes decisions $d_{\text{prox}} \sim p(\tilde{Y} \mid \mathbf{X}, S)$ from such biased observed data is unfair.

Assuming access to only biased data, we posit using a conditional latent variable model for fair decision making. As we theoretically show, with the help of a conditional latent variable model (LVM), it is possible to learn a latent representation $\mathbf{Z}$ that: i) is independent of the sensitive S, i.e., $\mathbf{Z}\perp S$ and ii) captures the information contained in Y, up to the noise of the observed features and the approximation error of the LVM.

We assume observed features $\mathbf{X}$ and proxy labels $\tilde{Y}$ are generated by Y, S and an independent noise variable $\mathcal{E} \in \mathbb{R}$.

Lemma 1. *Assume the observed $\{\mathbf{X}, \tilde{Y}\}$ is a bijective function of the ground-truth Y, sensitive S and noise $\mathcal{E}$ with $S, Y, \mathcal{E}$ being pairwise independent. Then, the conditional data entropy is $H(\mathbf{X}, \tilde{Y} \mid S) = H(Y) + H(\mathbf{X}, \tilde{Y} \mid S, Y) = H(Y) + H(\mathcal{E})$.*

So, the conditional data distribution $p(\mathbf{X}, \tilde{Y} \mid S)$ captures the information of the unobserved ground-truth Y, up to the extent of noise $\mathcal{E}$. Next, we consider approximating the underlying data distribution with LVMs.

Lemma 2. *Given a latent variable model conditional on S and input data $\{\mathbf{X}, \tilde{Y}\}$, having encoder $q_\phi(\mathbf{Z} \mid \mathbf{X}, \tilde{Y}, S)$ and decoder $p_\theta(\mathbf{X}, \tilde{Y} \mid \mathbf{Z}, S)$, the mutual information between latent variable $\mathbf{Z}$ and the conditional data distribution $p(\mathbf{X}, \tilde{Y} \mid S)$ is $I(\mathbf{Z}; \mathbf{X}, \tilde{Y} \mid S) = H(\mathbf{X}, \tilde{Y} \mid S) - \Delta$ with approximation error Δ.*

Hence, the information captured by the latent $\mathbf{Z}$ reduces the uncertainty about the conditional distribution $p(\mathbf{X}, \tilde{Y} \mid S)$ up to the error. We refer to Appendix A.2 for the detailed proof following [2]. Combining the two lemmas, we get:

$$I(\mathbf{Z}; \mathbf{X}, \tilde{Y} \mid S) = \underbrace{H(Y)}_{\text{information of } Y} + \underbrace{H(\mathcal{E})}_{\text{noise}} - \underbrace{\Delta}_{\text{approx. error}} \quad (1)$$

The two lemmas together show that using a conditional LVM to model the observed data $p(\mathbf{X}, \tilde{Y} \mid S)$ allows us to learn a latent variable $\mathbf{Z}$ that captures the information of the unobserved ground-truth Y, up to the extent of noise (note that Δ is also dependent on $\mathcal{E}$). Consequently, a policy that learns to make decisions using the latent $\mathbf{Z}$ with respect to the proxy $\tilde{Y}$ would, in fact, make decisions based on the information contained in Y (up to the effect of noise $H(\mathcal{E})$).

As pointed out in Section 2, following [35], a policy deciding based on Y satisfies *counterfactual fairness* and *demographic parity*. Hence, a policy π mapping from $\mathbf{Z}$ to $\tilde{Y}$ tackles label bias and satisfies both fairness notions without the need for additional constraints (up to the distortion due to $H(\mathcal{E})$ and Δ). Following, in Section 4, we propose a pipeline to learn a fair policy using unbiased representations $\mathbf{Z}$ from *non-i.i.d.* data that suffer from both label and selection bias.

4 OUR APPROACH

In this section, we propose a novel online fair policy learning framework FairAll for tackling both *biased* and *selective* labels. Our FairAll framework consists of: i) a fair representation learning step that relies on a VAE-based model (illustrated in Figure 1b) trained on both labeled and unlabeled data and; ii) a policy learning approach that leverages the learned fair representations to approximate the optimal fair policy according to the ground truth Y. Both steps of our framework, i.e., the VAE and the policy, are continually optimized as more data becomes available through the development of previous policies (i.e., *in an online manner*). Note, in taking decisions based on a fair representation, our policy mitigates *label bias*; in learning a stochastic policy in an online manner, we allow for exploration during training, which mitigates *selection bias* [31]. Specifically, we correct *label bias* by conditioning the VAE on sensitive S, while we correct *selection bias* by weighting our online learning loss with IPS.

In the following, we first detail how to use both labeled and unlabeled data to learn a fair representation and then describe decision learning with policy π. Lastly, we present an overview of our entire fair online policy learning pipeline and propose a method to further exploit unlabeled population information.

4.1 Learning a Fair Representation

Following the result in Eq. 1, we aim to learn latent $\mathbf{Z}$ that is both informative of Y and independent of the sensitive attribute S, i.e., $\mathbf{Z}\perp S$. Consider an online setting in which we have access to a dataset $\mathbf{A}_t$ (*applicants*) at each time step t. The partitioning of $\mathbf{A}_t$ into labeled data $\mathbf{A}_t^L$ (*accepted applicants*) and unlabeled $\mathbf{A}_t^{UL}$ (*rejected applicants*) is invoked by policy π_t. To ease notation, we will, in the following, consider a particular time step t and omit the subscript.

Let us recall that for each data observation, we only observe the proxy label $\tilde{Y}$ if the previous policy made the positive decision $D = 1$ (labeled data), and the actual value of the ground truth remains unobserved. However, we can leverage the fact that the utility with respect to the proxy label, $\tilde{U} = D(\tilde{Y} - c)$, is always observed.[8] This allows us to learn an unbiased latent representation $\mathbf{Z}$ from both labeled and unlabeled data.

VAE-Based Fair Representation Learning. Specifically, we build on previous work on semi-supervised and conditional VAEs [32, 44, 49] to approximate the conditional distribution $p_\theta(\mathbf{X}, \tilde{U}|S, D) = \int p_\theta(\mathbf{X}, \tilde{U}, \mathbf{Z}|S, D)dZ$ (generative model), and the posterior over the fair latent representation

$$q_{\omega,\phi}(\mathbf{Z}|\mathbf{X}, S, D = 1) = \int q_\phi(\mathbf{Z}|\mathbf{X}, S, \tilde{u}, D = 1) q_\omega(\tilde{u}|\mathbf{X}, S, D = 1) d\tilde{u}.$$

The inference model contains an encoder q_ϕ and a separate classifier model q_ω. Note, we condition the inference model on $d = 1$ (see Appendix C.1 for an overview) and thus introduce the classifier to predict the label for any unlabeled data point. We optimize the model parameters (θ, ϕ, and ω) by minimizing the following

[8] As per Section 2.2, utility U can take three values dependent on the decision, hence providing a value for accepted and rejected applicants.

objective function:

$$J(\theta,\phi,\omega) = \alpha \underbrace{\mathbb{E}_{(\boldsymbol{x},\tilde{u},s)\sim \mathbf{A}^L}[\mathcal{R}(\omega;\boldsymbol{x},s,\tilde{u},\pi)]}_{\text{IPS-weighted classification loss}}$$
$$\underbrace{-\,\mathbb{E}_{(\boldsymbol{x},\tilde{u},s)\sim \mathbf{A}^L, d=1}[\mathcal{L}(\theta,\phi;\boldsymbol{x},s,\tilde{u})] - \mathbb{E}_{\boldsymbol{x},s\sim \mathbf{A}^{UL}, d=0}[\mathcal{U}(\theta,\phi;\boldsymbol{x},s)]}_{\text{ELBO}} \tag{2}$$

where the latter terms corresponds to the ELBO capturing the goodness of fit of the VAE; and the first term measures the accuracy of the classifier model that estimates the utility of labeled data. The hyperparameter α balances the classification loss term relative to ELBO. Notice that defining our model with respect to the completely observed $\tilde{U}$ allows us to compute the ELBO on both labeled data and unlabeled data, removing the need for IPS.[9]

Evidence Lower Bound (ELBO). More specifically, for all accepted applicants ($d = 1$), we compute the ELBO of $\log p_\theta(\mathbf{X}, \tilde{U}|S, D = 1)$ as:

$$\begin{aligned}\mathcal{L}(\theta,\phi;\boldsymbol{x},s,\tilde{u}) &= \mathbb{E}_{z\sim q_\phi(\mathbf{Z}|\boldsymbol{x},s,\tilde{u},D=1)}[\log p_\theta(\boldsymbol{x}|z,s)\\ &\quad + \log p_\theta(\tilde{u}|z,s,D=1)] - KL(q_\phi(\mathbf{Z}|\boldsymbol{x},\tilde{u},s,D=1)||p(\mathbf{Z}))\end{aligned} \tag{3}$$

For all rejected applicants, which received $d = 0$, utility is a constant $\tilde{U} = 0$, such that $\log p(\tilde{U} = 0|\mathbf{Z}, D = 0) = \log 1 = 0$. Thus, the ELBO for unlabeled data is given by:

$$\begin{aligned}\mathcal{U}(\theta,\phi;\boldsymbol{x},s) &= \mathbb{E}_{\tilde{u}\sim q_\omega(\tilde{U}|\boldsymbol{x},d=1)}\mathbb{E}_{z\sim q_\phi(z|\boldsymbol{x},\tilde{u},d=1)}[\log p_\theta(\boldsymbol{x}|z,s)]\\ &\quad - KL[\int q_\phi(z|\boldsymbol{x},\tilde{u},s,d=1)q_\omega(\tilde{u}|\boldsymbol{x},s,d=1)d\tilde{u}||p(\mathbf{Z}))\end{aligned} \tag{4}$$

Classification Loss. We utilize labeled data and cost-sensitive cross-entropy loss to train the classifier $q_\omega(\tilde{U}|\mathbf{X}, S, D = 1)$. Based on decision costs, false negatives are weighed by the lost profit $(1 - c)$, and false positives by c. Note for labeled data, we have a binary prediction task as $\tilde{U} \in \{-c, 1 - c\}$, which in this context can be taken as $\{0, 1\}$. As we learn only on labeled data, we apply IPS weights based on the policy π to correct for selective bias as:

$$\begin{aligned}\mathcal{R}(\omega;\boldsymbol{x},s,\tilde{u},\pi) &= -\Big[c(1-\tilde{u})\log(1-q_\omega(\tilde{u}=1|\boldsymbol{x},s,d=1))\\ &\quad + (1-c)\tilde{u}\log q_\omega(\tilde{u}=1|\boldsymbol{x},s,d=1)\Big]\underbrace{\frac{1}{\pi(d=1|\boldsymbol{x},s)}}_{\text{IPS}}\end{aligned} \tag{5}$$

In our implementation, we learn functions $p_\theta(\mathbf{X}|\mathbf{Z}, S), p_\theta(\tilde{U}|\mathbf{Z}, S, D = 1)$, $q_\omega(\tilde{U}|\mathbf{X}, S, D = 1)$, $q_\phi(\mathbf{Z}|\tilde{U}, S, \mathbf{X}, D = 1)$ with fully-connected (deep) neural networks. See Appendix C and D for practical considerations and training setups.

4.2 Learning a Fair Policy

At each time step after improving our representation learning model, we update the policy π that maps the fair representation $\mathbf{Z}$ into a distribution over decisions D. For each data point in the complete dataset $\mathbf{A} = \mathbf{A}^L \cup \mathbf{A}^{UL}$ we sample z using the *encoder* $q_\phi(\mathbf{Z}|\tilde{u}, s, \boldsymbol{x}, D = 1)$, and use the decoder to get an estimated proxy utility, $\tilde{u} \sim p_\theta(\tilde{U}|\mathbf{Z}, S, D = 1)$. Note, for $D = 1$, the proxy utility is binary. We thus train the policy by minimizing the cross-entropy binary loss: $(1 - \tilde{U}) \log \pi(\mathbf{Z}) + \tilde{U} * \log(1 - \pi(\mathbf{Z}))$. We describe other options for the policy model for our approach in Appendix D.4.

4.3 Exploiting Fully Unlabeled Data

In real-world settings, a decision-maker often has prior access to large unlabeled datasets containing i.i.d. samples from the population of interest. For example, in the case of university admissions, a university may have access to a database of all students who passed their high school diploma to enter university – including those who did not apply at that particular university. Such dataset contains features $\mathbf{X}, S$, but no labels $\tilde{Y}$, i.e., it is fully unlabeled. However, as we show, such unlabeled data can significantly improve the policy learning process. In particular, the fully unlabeled dataset can help learn fair representations $\mathbf{Z}$ by approximating the conditional $p(\mathbf{X} \mid S)$. That is, we can learn a VAE optimized as:

$$\begin{aligned}\mathcal{K}(\theta',\phi';x,s) &= \mathbb{E}_{z\sim q_{\phi'}(z|\boldsymbol{x},s)}[\log p_{\theta'}(\boldsymbol{x} \mid z,s)]\\ &\quad - \mathrm{KL}\left(q_{\phi'}(\mathbf{Z} \mid \boldsymbol{x},s)||p(\mathbf{Z})\right)\end{aligned} \tag{6}$$

The resulting unsupervised VAE model can then be used to initialize the parameters of the semisupervised VAE proposed in Section 4.1 via transfer learning. Transfer learning is typically studied in supervised learning [43, 54], where models are pre-trained on large datasets of one domain and then transferred to a different domain with fewer data. In our proposed *two-phase* approach, we utilize the unlabeled data in a Phase I to initialize our semi-supervised VAE model of the online decision-making Phase II. We initialize parameters ϕ and θ of semi-supervised VAE with the trained parameters ϕ' and θ' of the unsupervised VAE. Note from Eq. 4.1 that the semi-supervised VAE encoder additionally takes $\tilde{U}$ as input and that the decoder also outputs $\tilde{U}$. To account for this, we add *new neural connections*. At the encoder, we add connections from input $\tilde{U}$ to each neuron in the first hidden layer. At the decoder's output, we add a new head to output $\tilde{U}$. The new connections are initialized randomly at the start of Phase II.

4.4 FairAll Overview

Our FairAll learning framework is illustrated in Figure 2 and consists of two phases. In the first step (Phase I), we learn a fair representation using an unsupervised VAE trained in an *offline* manner using only unlabeled data. We then use the resulting model to initialize the parameters of the semisupervised VAE via transfer learning. In a second step (Phase II), we enter the *online* decision-learning process, where at each time step, we first update our semisupervised VAE using both labeled and unlabeled data, and then update the decision policy π.

5 EXPERIMENTAL RESULTS

In this section, we evaluate our fair policy learning framework with regard to i) its convergence to the optimal policy; ii) its training effectiveness until convergence; and iii) its deployment performance after convergence. We can evaluate (i) on synthetic data only, and evaluate (ii) and (iii) on real-world datasets.

[9]Previous work [49] defined a semi-supervised VAE with respect to partially observed label $\tilde{Y}$. In this case, one would need to apply IPS on the ELBO, which may introduce high variance.

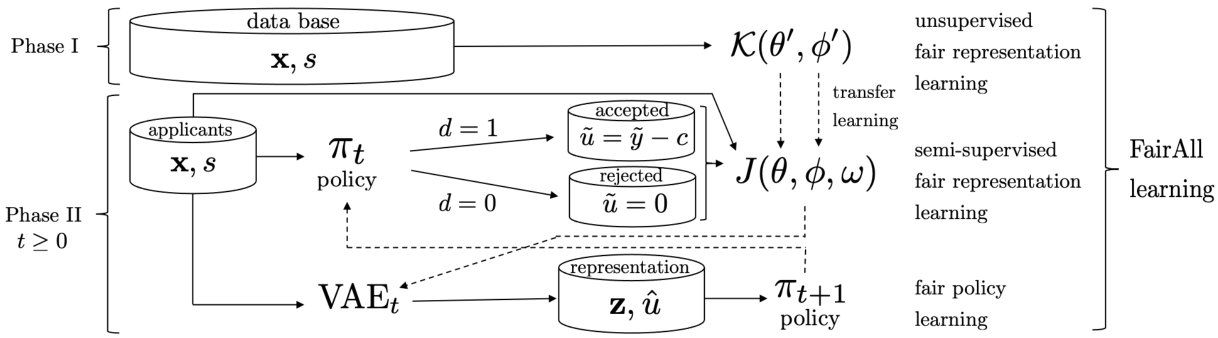

Figure 2: Pipeline of our approach FairAll. In Phase I, we pre-train the VAE using a large pool of unlabeled data with features x, s. We transfer the trained parameters ϕ' and θ' of the VAE at the start of Phase II. Next, at each time step t, decisions d for a new batch of data are drawn from the current policy π_t. In case of acceptance ($d = 1$), labels $\tilde{y}$ are revealed. The VAE is updated via loss J using both labeled and unlabeled data. Subsequently, the updated VAE outputs fair representations z and estimated utility $\hat{u}$ for all applicants. The policy is updated with these fair representations. In this way, FairAll learns to decide by using all available data.

Baseline and Reference Models. We perform rigorous empirical comparisons among the following learning frameworks:

- **FairAll (I+II)**: *Our complete proposed* learning framework including Phase I (i.e., offline unsupervised representation learning) and Phase II (online semisupervised representation and policy learning).
- **FairAll (II)**: *Baseline* approach that make use of only Phase II of the proposed FairAll. This approach allows us to evaluate the impact of Phase I, i.e., fully unlabeled data.
- **FairLab (I+II)**: *Baseline* approach that consist of unsupervised Phase I and a *fully supervised* Phase II using only the IPS-weighted ELBO on labeled data. It allows us to evaluate the importance of unlabeled data in Phase II.
- **FairLog** [31]: *Competing approach* that minimizes the IPS weighted cross entropy (Eq. 5), denoted by $\mathcal{L}^{\text{UnfairLog}}$, with a Lagrange fairness constraint, i.e., $\mathcal{L}^{\text{UnfairLog}} + \lambda * \text{DPU}$ with DPU as defined in Def. 2.2.
- **UnfairLog** [31]: *Unfair reference* model, corresponding to FairLog without fairness constraint, i.e., $\lambda = 0$. It allows us to measure the *cost of fairness.*

We refer to the Appendix for a detailed description of baselines and competing methods (Appendix D.5), details on the hyperparameter selection (Appendix D.2) and other practical considerations (Appendix C).

Metrics. We measure *observed proxy utility* $\widetilde{\text{UT}}$ (Def. 2.1 w.r.t. $\tilde{Y}$) and *demographic parity unfairness* DPU (Def. 2.2) on i.i.d. test data on both synthetic and real-world datasets. Further, on synthetic data with access to the ground truth generative process, we report *counterfactual unfairness* CFU (Def. 2.3)[10] as well as the *unobserved ground truth utility* UT (Def. 2.1 w.r.t. Y). For the real-world settings, we report *effective* $\widetilde{\text{UT}}$ [31], which is the average $\widetilde{\text{UT}}$ accumulated by the decision-maker on the training data up to time t. Similarly, we report *effective* DPU. We also report the *temporal variance* (Def. 2.4) of $\widetilde{\text{UT}}$ and DPU over time interval $t = [125, 200]$.

[10] See Appendix C.4 for further details on generating counterfactuals on the synthetic dataset.

Datasets. We report results on one synthetic and three real-world datasets (more details in Appendix B):

- Synthetic, where $\mathbf{X}$ contains 2 features, with Gender S and *grades after university admission* $\tilde{Y}$ (note that the ground truth *intellectual potential* Y is considered unobserved and only used for evaluation).
- COMPAS [3, 37], where $\mathbf{X}$ contains 3 features, with Race S and *no recidivism* $\tilde{Y}$.
- CREDIT [15], where $\mathbf{X}$ contains 19 non-sensitive features, with Gender S and *credit score* $\tilde{Y}$.
- MEPS [24], where $\mathbf{X}$ contains 39 non-sensitive features, with Race S and *high healthcare utilization* $\tilde{Y}$.

Optimal Policies. In our Synthetic setting, where observed $\mathbf{X} \not\perp S$ and unobserved $K \perp S$, the optimal unfair policy (OPT-UNFAIR) takes decisions $d \sim p(\tilde{Y} \mid X, S)$ and the optimal fair policy (OPT-FAIR) decides $d \sim p(\tilde{Y} \mid K)$. OPT-UNFAIR can be approximated with access to i.i.d. samples from the posterior distribution, while OPT-FAIR additionally requires access to unobserved K. See Appendix C.3 for details.

Setup. We assume access to the proxy labels. Since labeled data is often scarce in practice, we assume an initial HARSH policy which labels around 10-18% of the data that we see in Phase II at $t = 0$. We report results for a lenient policy in Appendix E. For details on initial policies, see Appendix D.1. The decision cost is $c = 0.1$ for MEPS and $c = 0.5$ for all other datasets. See Appendix E.4 for a case study on the impact of the cost value. Wherever applicable, Phase I was trained over a large number of epochs. Phase II was trained for 200 time steps with the same number of candidates in each step. All policy training were done over 10 independent random initializations. For a full description of the experimental setup, see Appendix D. Our code is publicly available[11].

[11] https://github.com/ayanmaj92/fairall

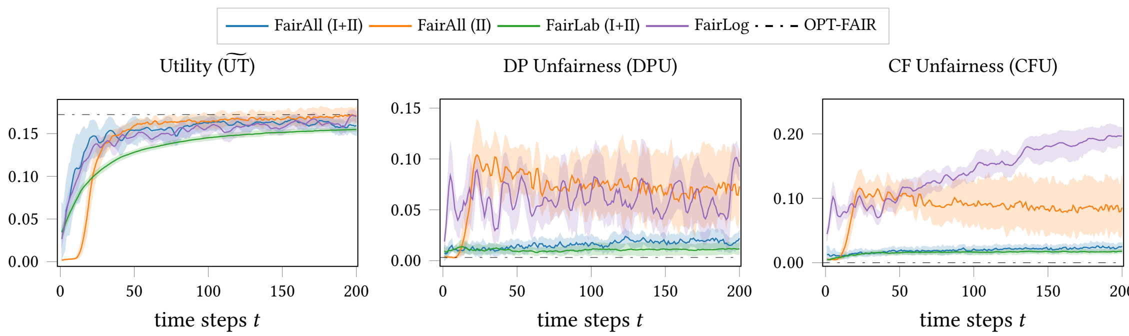

Figure 3: Utility ($\widetilde{\mathrm{UT}}$) with respect to the proxy, demographic parity unfairness (DPU), and counterfactual unfairness (CFU) on the Synthetic dataset. FairAll (I+II) converges to the optimal fair policy in utility and fairness while being more fair than FairLog.

5.1 Can We Reach the Optimal Fair Policy?

In this section, we use synthetic data to evaluate if, given enough time-steps, the different learning methods yield policies that converge to the *optimal (fair) policy* (OPT-FAIR) both in terms of proxy utility and fairness.

Results. Figure 3 reports $\widetilde{\mathrm{UT}}$, DPU and CFU on test data across time. Recall that FairLab (I+II) uses unlabeled data only in Phase I, FairAll (II) only in Phase II, and FairAll (I+II) in both Phase I and Phase II. FairAll (II) starts convergence after approximately 20 steps. FairLab (I+II) starts at higher utility but then exhibits slower convergence behavior and has not fully converged at $t = 200$. FairAll (I+II) instead starts at a higher utility and at $t = 200$ yields utility and fairness (both in demographic parity and counterfactual fairness) values close to OPT-FAIR. Regarding fairness, FairAll (II) does not converge to OPT-FAIR and exhibits high variance in DPU and CFU both across seeds and over time. Instead, FairAll (I+II) and FairLab (I+II) both rely on unsupervised fair representation learning (Phase I) to converge to a value very close to the optimal one. Thus, the fully unsupervised Phase I appears crucial for convergence to the optimal fair policy with low variance. Comparing FairAll to FairLog, we observe that both methods asymptotically converge to the optimal fair utility. However, FairLog suffers from higher and more noisy DPU while CFU *increases over time*. This can be explained by the fact that FairLog enforces DPU constraints, and, as shown empirically, DPU *does not imply* CFU. In contrast, FairAll (I+II) learns a fair representation $\mathbf{Z}$ that achieves CFU and, as a result, also DPU (see Section 3).

This evaluation concludes that i) all fair models approximately converge to the optimal utility; ii) unlabeled data helps in faster convergence; iii) utilizing Phase I leads to significantly lower unfairness; iv) our approach FairAll, compared to FairLog, achieves convergence while satisfying both counterfactual and demographic parity fairness notions.

5.2 Do We Actually Trade-off Utility for Fairness?

In Figure 4, we observe that FairAll yields higher fairness but lower observed *proxy utility* $\widetilde{\mathrm{UT}}$ compared to the unfair reference model UnfairLog. This observation is often referred to as a *fairness-accuracy trade-off* [6], which assumes that fairness comes at the cost of utility (or accuracy in predictive settings). As pointed out in [16, 52], if utility is a function of biased labels, then the utility measurement is also biased. Recall the assumption that a decision-maker aims to take decisions based on Y and aims to maximize UT. For example, potential ($Y = 1$) drives a successful career, not high university grades ($\tilde{Y} = 1$). In the synthetic setting, we can measure unbiased *ground truth utility* UT.

Results. In Figure 4, we observe that both FairAll and UnfairLog achieve a similar level of ground truth utility UT, while FairAll reports significantly less discrimination (DPU). This suggests that a decision-maker *may not* actually trade-off fairness and (true) utility, although we observe lower proxy utility. Note that despite being fair, FairAll (and FairLog) do not reach the UT of the optimal (fair) policy OPT-FAIR. This could be due to the noise $\mathcal{E}$ in the dataset, which is known to the optimal policy, but is naturally not captured by the models. An in-depth discussion of the phenomenon is outside the scope of this paper, and we refer the reader to the literature [11, 16, 52].

5.3 How Effective Is the Learning Process?

We have shown that FairAll asymptotically outputs a policy that approximates the optimal. Now we investigate how much it costs the decision-maker in terms of utility and the society in terms of fairness to learn this policy. We evaluate the *effective* proxy $\widetilde{\mathrm{UT}}$, and DPU that the online learning process accumulates across time on *real-world datasets.*

Results. Table 1 summarizes the results for several real-world datasets. FairAll (I+II) consistently accumulates more utility and less unfairness during the online learning process compared to the other approaches. Note that FairLab (I+II), which uses only labeled data in Phase II, accumulates less utility and more unfairness than FairAll (II), which skips Phase I but uses both labeled and unlabeled data in Phase II. This suggests that joint training on labeled and unlabeled data in Phase II significantly improves learning of both a fair representation and policy. Furthermore, FairAll (I+II) outperforms FairAll (II), suggesting that using unlabeled data in

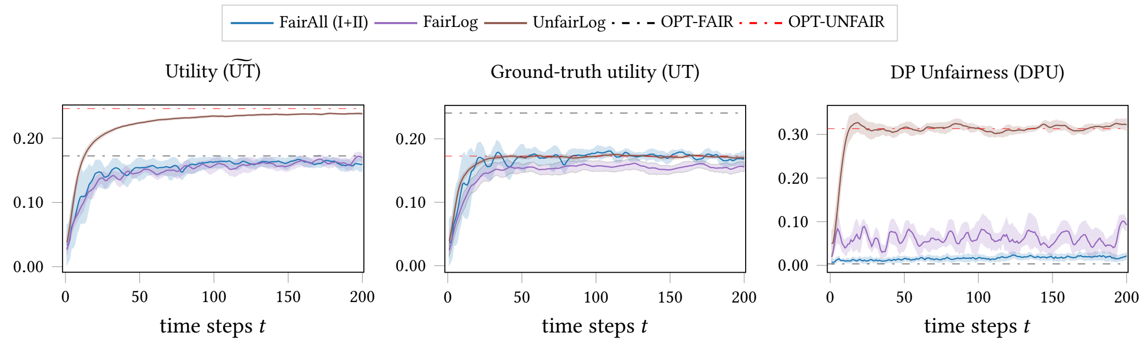

Figure 4: Results on the Synthetic dataset. For FairAll (I+II) and UnfairLog a trade-off between fairness, i.e., demographic parity (DPU), and utility when measured with respect to proxy labels ($\widetilde{\text{UT}}$) – but not when measured with respect to ground truth (UT).

Table 1: Accumulated utility w.r.t. observed proxy (Effect. $\widetilde{\text{UT}}$) and demographic parity unfairness (Effect. DPU) measured during the policy learning process for different real-world datasets over time interval $t = [0, 200]$. We report mean values over the same ten independent seeds. The numbers in the brackets show the standard deviation. All reported values are multiplied by 100 for readability.

Model	COMPAS		CREDIT		MEPS	
	Effect. $\widetilde{\text{UT}}$(↑)	Effect. DPU(↓)	Effect. $\widetilde{\text{UT}}$(↑)	Effect. DPU(↓)	Effect. $\widetilde{\text{UT}}$(↑)	Effect.DPU(↓)
FairAll (I+II)	6.2 (0.8)	10.4 (0.7)	20.3 (0.5)	8.2 (1.6)	8.0 (0.4)	10.1 (1.4)
FairAll (II)	5.1 (0.6)	10.6 (0.7)	19.8 (1.0)	7.8 (1.9)	7.7 (0.2)	9.4 (1.8)
FairLab (I+II)	2.3 (0.9)	10.7 (0.9)	16.3 (0.7)	10.0 (1.9)	5.6 (0.5)	10.3 (0.4)
FairLog	3.6 (0.4)	10.8 (0.9)	19.4 (1.0)	10.2 (1.6)	6.8 (0.5)	11.0 (1.2)
UnfairLog	4.6 (0.7)	14.6 (1.5)	20.4 (0.5)	10.8 (1.9)	7.6 (0.3)	19.1 (2.3)

Phase I also improves the process. In summary, our results show that *unlabeled data available at any learning stage should not be thrown away, but should be used to learn fair and profitable policies.* Importantly, our method also outperforms the competing approach FairLog both in terms of utility and fairness. Moreover, even if compared to the unfair approach UnfairLog, FairAll (I+II) accumulates comparable or higher utility with significantly lower unfairness. This suggests that we *we may not observe a trade-off between observed utility and fairness in the real world, assuming an unbiased ground truth exists.* Note that without access to the ground truth label, we cannot comment on the performance with respect to the ground truth utility.

5.4 How Do the Learned Policies Perform During Deployment?

In this section, we analyze how a given strategy, when applied to the population of interest, is expected to perform in the long run. To this end, we compare the performance of the resulting strategy at each time step using an i.i.d. test set.

Result. Figure 5 shows how utility $\widetilde{\text{UT}}$ and unfairness DPU evolve over time for COMPAS. Results on the other real-world datasets are in Appendix E. FairAll (II) achieves both higher utility and higher unfairness compared to FairLab (I+II) with significantly higher variance. This suggests that unlabeled data in Phase I benefit fairness, while unlabeled data in Phase II benefit utility. FairAll (I+II) provides significantly higher utility and lower unfairness than the other learning models after approximately 50 time steps. Moreover, FairAll (I+II) even provides higher utility than even the unfair reference model UnfairLog while being as fair as FairLog. This empirically confirms the importance of unlabeled data *in both Phase I and Phase II* to achieve high test utility and fairness in real-world scenarios.

5.4.1 Can We Reliably Stop Learning at Any Time? Assume that we want to stop the policy learning process from time t_1. Can we stop the learning process and deploy the policy at any time $t \geq t_1$? We study this by measuring how much utility (fairness) of the output policies vary over time interval $[t_1, t]$ for a *fixed test dataset.* A low temporal variance (TV) indicates stable behavior, such that it is safe to terminate the learning process at any time. However, when the variance is high, the decision-maker must carefully select the best stopping point. High TV leads to unstable policies that may lead to low utility and/or high fairness. In addition, we measure the temporal average μ of utility (unfairness). It is desirable for a learning process to have both low variance and high (low) μ for utility (unfairness).

Results. In Table 2, we report $\text{TV}_{\text{DPU}}, \mu_{\text{DPU}}$ for unfairness and $\text{TV}_{\widetilde{\text{UT}}}, \mu_{\widetilde{\text{UT}}}$ for utility (see Def. 2.4). Compared to FairLab (I+II) and FairAll (II), FairAll (I+II) provides similar TV values. However, FairAll (I+II) exhibits a better temporal average utility and

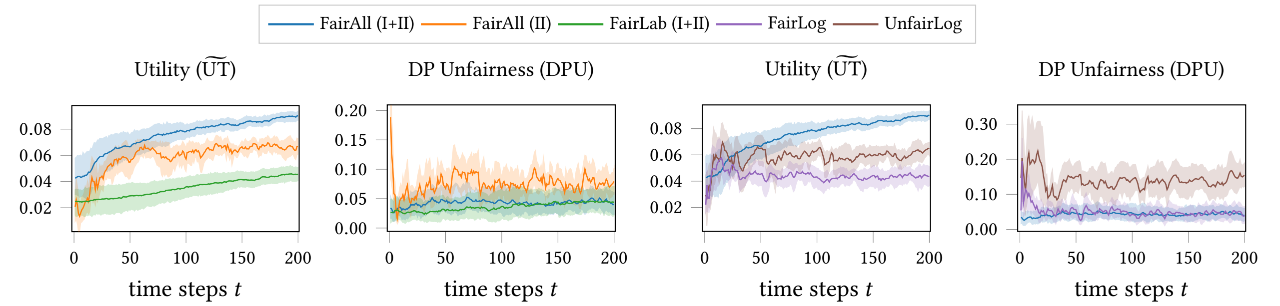

Figure 5: Proxy utility ($\widetilde{\mathrm{UT}}$) and demographic parity unfairness (DPU) of policy models deployed on test-data for COMPAS dataset at each time-step. Left two plots show the benefit of using unlabeled data in both phases, comparing FairAll against FairAll (II) and FairLab (I+II). Right two plots show the better performance of FairAll compared to competing methods FairLog, UnfairLog.

fairness. For CREDIT, for example, note that FairLab (I+II) has a lower $\mathrm{TV}_{\mathrm{DPU}}$ and a higher average unfairness level μ_{DPU} than FairAll (I+II). Compared to FairLog, our method FairAll (I+II) is much more stable. It has lower $\mathrm{TV}_{\mathrm{DPU}}$ and $\mathrm{TV}_{\widetilde{\mathrm{UT}}}$ as well as a higher average utility $\mu_{\widetilde{\mathrm{UT}}}$ and lower unfairness μ_{DPU}.

6 DISCUSSION OF ASSUMPTIONS AND OUTLOOK

In this section, we discuss the main assumptions, limitations, and potential consequences of our proposed framework.

Assumptions on the Data Generation Process. In this work, we assume that the true data generation process follows Figure 1a and that S is a social construct like *gender* or *race*. This follows the understanding that discrimination is based on social constructs and not biological differences [28]. We assume that each individual can be assigned to a social group within a social construct (e.g., *gender*) and that S is a root node in the graphical model. While being common [35], this is a debated modeling assumption [6, 28]. Furthermore, we assume that we observe a biased proxy variable $\tilde{Y}$, and that an unobserved unbiased ground truth Y exists that is independent of S. This means that there are no innate differences between social groups with respect to Y. However, see Appendix F.1 for examples where $Y \not\perp S$. We advise any practitioner using our pipeline to carefully assess whether these assumptions hold in their case.

Assumptions of Our Policy Learning Pipeline. First, we assume access to a large unlabeled dataset of i.i.d. samples from the population of interest for our pipeline (Phase I). For example, a university may have access to the grades of all students who graduated from high school in a given time period. We also assume access to sensitive information, which, in the real world, may conflict with privacy regulations (e.g., the principle of data minimization). In this paper, we show that access to a large unlabeled dataset of sensitive information not only increases the utility of the decision-maker but also fairness. We hope that this contributes to the debate between fair algorithmic decision-making and privacy.

Second, we assume a decision-maker has an unlimited budget at each time step, i.e., it can take as many positive decisions as desired. A related line of work [29, 34], deals with fairness in selection problems, where candidates compete for a limited number of positions, or with pipeline settings [18], where candidates enter the decision process one at a time. This is an interesting direction for future research.

Third, we assume that the underlying data distribution does not change over time and thus is not affected by the decisions. This assumption does not necessarily hold in the real world. While it exceeds the scope of this paper, it would be interesting to extend our pipeline to address *distribution shift* as a consequence of the decision-making process.

Lastly, in Phase II we learn a stochastic policy at each time step and use it to collect new data. We follow Kilbertus et al. [31] in their call for a general discussion about the ethics of non-deterministic decision-making.

Assumptions on Fairness Metrics. While the ethical evaluation of the applicability of a particular fairness notion in a specific context lies outside of our scope, we give an overview of when the use of our pipeline may be helpful in practice. We evaluate fairness based on the demographic parity (DP) notion [17]. Heidari et al. [25] map DP to Rawl's moral understanding, according to which unequal treatment may only be justified on the basis of innate potential or ambition, not gender or race. Within this framework, the underlying assumption of DP is that individuals should receive utility solely based on the factors for which they are accountable. In this paper, such factors are assumed to be captured by the unobserved ground truth label Y. Hertweck et al. [26] show that one should enforce DP not only if socio-demographic groups have similar innate potential at birth, but in some cases even if unjust life biases lead to differences in realized abilities. Wachter et al. [51] similarly argue that DP and counterfactual fairness are bias transforming metrics that acknowledge historical inequalities and assume that certain groups have a worse starting point than others. However, they also warn that, e.g., giving an individual a loan that they cannot repay can exacerbate inequalities.

7 CONCLUSION

In this paper, we considered the problem of learning optimal fair decision policies in the presence of label bias and selection bias.

Table 2: Temporal variance and means of utility ($\mathrm{TV}_{\widetilde{\mathrm{UT}}}, \mu_{\widetilde{\mathrm{UT}}}$) and demographic parity unfairness ($\mathrm{TV}_{\mathrm{DPU}}, \mu_{\mathrm{DPU}}$). Metrics are measured on the time interval $t = [125, 200]$ on real-world datasets. We report the mean over ten runs with the standard deviation in brackets. For TV, lower values are better. For μ higher (lower) is better for $\widetilde{\mathrm{UT}}$ (DPU). Values multiplied by 100 for readability.

Model	COMPAS		CREDIT		MEPS	
	$\mathrm{TV}_{\mathrm{DPU}}(\downarrow)$ | μ_{DPU} ($\downarrow$)	$\mathrm{TV}_{\widetilde{\mathrm{UT}}}$ ($\downarrow$) | $\mu_{\widetilde{\mathrm{UT}}}$ ($\uparrow$)	$\mathrm{TV}_{\mathrm{DPU}}$ ($\downarrow$) | μ_{DPU} ($\downarrow$)	$\mathrm{TV}_{\widetilde{\mathrm{UT}}}$ ($\downarrow$) | $\mu_{\widetilde{\mathrm{UT}}}$ ($\uparrow$)	$\mathrm{TV}_{\mathrm{DPU}}$ ($\downarrow$) | μ_{DPU} ($\downarrow$)	$\mathrm{TV}_{\widetilde{\mathrm{UT}}}$ ($\downarrow$) | $\mu_{\widetilde{\mathrm{UT}}}$ ($\uparrow$)
FairAll (I+II)	1.0 (0.7) | 4.3 (3.1)	0.4 (0.3) | 8.6 (0.7)	3.0 (2.0) | 5.0 (5.0)	1.0 (0.7) | 18.8 (1.3)	1.8 (1.6) | 5.1 (3.1)	0.2 (0.1) | 7.9 (0.5)
FairAll (II)	2.8 (1.7) | 7.8 (4.0)	0.8 (0.5) | 6.6 (1.0)	2.7 (2.4) | 4.6 (3.0)	0.7 (0.6) | 18.2 (1.5)	3.8 (2.9) | 6.4 (4.8)	0.3 (0.2) | 7.4 (0.5)
FairLab (I+II)	0.7 (0.6) | 4.3 (3.1)	0.3 (0.3) | 4.2 (0.9)	1.7 (1.3) | 9.1 (5.7)	0.5 (0.4) | 14.9 (1.4)	0.7 (0.6) | 4.0 (2.3)	0.3 (0.2) | 5.9 (0.7)
FairLog	1.8 (1.4) | 4.5 (3.8)	0.5 (0.4) | 4.3 (1.1)	3.6 (2.0) | 7.6 (4.4)	1.3 (1.1) | 17.1 (2.3)	2.6 (2.8) | 5.4 (3.3)	0.5 (0.4) | 6.4 (0.7)

Prior work that attempted to solve the problem by learning stochastic exploratory decision policies in an online process neglects a large amount of unlabeled data and suffers from high variance in the learned policies over time. In this work, we proposed a novel two-phase framework that leverages both labeled and unlabeled data to learn stable, fair decision policies. We introduced a practical sequential decision-making pipeline FairAll that uses the latent representations of a variational autoencoder to learn, over time, the *optimal fair policy according to the unobserved ground truth.* In line with our assumptions, the decision policies learned by FairAll satisfy both the notion of counterfactual fairness and demographic parity without requiring additional fairness constraints. Through theoretical analysis and experiments with synthetic data, we validate that FairAll converges to the optimal (fair) policy with respect to the unobserved ground truth both in terms of utility and fairness. Compared to the prior work, we show how our modeling approach helps us to be counterfactual and demographic parity fair. On real-world data, we show how FairAll provides not only a significantly more effective learning method, but also higher utility, higher fairness, and a more stable learning process than the existing approach. In comparison to baseline models, we also show the importance of using unlabeled data in both phases to achieve a more accurate, fair, and stable decision learning process.

ACKNOWLEDGMENTS

The authors would like to thank Adrián Javaloy Bornás for his guidance, discussion and for providing an initial codebase for working on VAEs with heterogenous data. The authors also thank Pablo Sánchez-Martín, Batuhan Koyuncu for providing valuable feedback. Special thanks goes to Diego Baptista Theuerkauf for help with formalizing proofs and notation. Finally, the authors thank the anonymous reviewers for their detailed feedback and comments.

Miriam Rateike is supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B. Ayan Majumdar and Krishna P. Gummadi are supported by the ERC Advanced Grant "Foundations for Fair Social Computing" (no. 789373). Olga Mineeva is supported by the Max Planck ETH Center for Learning Systems. The authors appreciate the generous funding support.

REFERENCES

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

[2] Alexander A. Alemi, Ben Poole, Ian Fischer, Joshua V. Dillon, Rif A. Saurous, and Kevin Murphy. 2017. An Information-Theoretic Analysis of Deep Latent-Variable Models. (2017). arXiv:1711.00464

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. *ProPublica* 23 (2016).

[4] Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. 2022. Scarf: Self-Supervised Contrastive Learning using Random Feature Corruption. In *International Conference on Learning Representations.* https://openreview.net/forum?id=CuV_qYkmKb3

[5] Mislav Balunovic, Anian Ruoss, and Martin Vechev. 2021. Fair Normalizing Flows. In *International Conference on Learning Representations.*

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning.* fairmlbook.org. http://www.fairmlbook.org.

[7] Solon Barocas and Andrew D Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104 (2016), 671.

[8] Yahav Bechavod, Katrina Ligett, Aaron Roth, Bo Waggoner, and Steven Z. Wu. 2019. Equal Opportunity in Online Classification with Partial Feedback. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/084afd913ab1e6ea58b8ca73f6cb41a6-Paper.pdf

[9] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. https://arxiv.org/abs/1810.01943

[10] L Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth K Vishnoi. 2018. An algorithmic framework to control bias in bandit-based personalization. *arXiv preprint arXiv:1802.08674* (2018).

[11] A. Feder Cooper, Ellen Abrams, and NA NA. 2021. Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.* Association for Computing Machinery, New York, NY, USA, 46–54. https://doi.org/10.1145/3461702.3462519

[12] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) *(KDD '17).* Association for Computing Machinery, New York, NY, USA, 797–806. https://doi.org/10.1145/3097983.3098095

[13] Andrew Cotter, Heinrich Jiang, and Karthik Sridharan. 2019. Two-Player Games for Efficient Non-Convex Constrained Optimization. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory (Proceedings of Machine Learning Research, Vol. 98)*, Aurélien Garivier and Satyen Kale (Eds.). PMLR, 300–332. https://proceedings.mlr.press/v98/cotter19a.html

[14] Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. 2019. Flexibly Fair Representation Learning by Disentanglement. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 1436–1445. https://proceedings.mlr.press/v97/creager19a.html

[15] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[16] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. 2020. Is There a Trade-Off Between Fairness and Accuracy? A Perspective Using Mismatched Hypothesis Testing. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 2803–2813. https://proceedings.mlr.press/v119/dutta20a.html

[17] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (Cambridge, Massachusetts) *(ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[18] Cynthia Dwork, Christina Ilvento, and Meena Jagadeesan. 2020. Individual Fairness in Pipelines. In *1st Symposium on Foundations of Responsible Computing (FORC 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

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

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (Eds.), Vol. 27. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[21] Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. 2020. Adversarial learning for counterfactual fairness. *arXiv preprint arXiv:2008.13122* (2020).

[22] Swati Gupta and Vijay Kamble. 2021. Individual Fairness in Hindsight. *Journal of Machine Learning Research* 22, 144 (2021), 1–35. http://jmlr.org/papers/v22/19-658.html

[23] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

[24] Agency for Healthcare Research & Quality. 2018. Medical Expenditure Panel Survey (MEPS). https://www.ahrq.gov/data/meps.html

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

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

[27] Daniel G Horvitz and Donovan J Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association* 47, 260 (1952), 663–685.

[28] Lily Hu and Issa Kohler-Hausmann. 2020. What's Sex Got to Do with Machine Learning?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 513. https://doi.org/10.1145/3351095.3375674

[29] Mohammad Mahdi Khalili, Xueru Zhang, and Mahed Abroshan. 2021. Fair Sequential Selection Using Supervised Learning Models. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 28144–28155. https://proceedings.neurips.cc/paper/2021/file/ed277964a8959e72a0d987e598dfbe72-Paper.pdf

[30] Ilyes Khemakhem, Ricardo Monti, Robert Leech, and Aapo Hyvarinen. 2021. Causal Autoregressive Flows. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 130)*, Arindam Banerjee and Kenji Fukumizu (Eds.). PMLR, 3520–3528. https://proceedings.mlr.press/v130/khemakhem21a.html

[31] Niki Kilbertus, Manuel Gomez Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. 2020. Fair Decisions Despite Imperfect Predictions. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 108)*, Silvia Chiappa and Roberto Calandra (Eds.). PMLR, 277–287. https://proceedings.mlr.press/v108/kilbertus20a.html

[32] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised Learning with Deep Generative Models. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (Eds.), Vol. 27. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2014/file/d523773c6b194f37b938d340d5d02232-Paper.pdf

[33] Diederik P Kingma and Max Welling. 2014. Auto-encoding Variational Bayes. In *2nd International Conference on Learning Representations*, Yoshua Bengio and Yann LeCun (Eds.), Vol. 2.

[34] Jon Kleinberg and Manish Raghavan. 2018. Selection Problems in the Presence of Implicit Bias. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[35] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

[36] Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) *(KDD '17)*. Association for Computing Machinery, New York, NY, USA, 275–284. https://doi.org/10.1145/3097983.3098066

[37] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[38] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. 2016. The Variational Fair Autoencoder. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1511.00830

[39] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning Adversarially Fair and Transferable Representations. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 3384–3393. https://proceedings.mlr.press/v80/madras18a.html

[40] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. 2018. Invariant Representations without Adversarial Training. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/415185ea244ea2b2bedeb0449b926802-Paper.pdf

[41] Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. 2022. Fair Contrastive Learning for Facial Attribute Classification. *arXiv preprint arXiv:2203.16209* (2022).

[42] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer.* John Wiley & Sons.

[43] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 512–519. https://doi.org/10.1109/CVPRW.2014.131

[44] Danilo Rezende and Shakir Mohamed. 2015. Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 37)*, Francis Bach and David Blei (Eds.). PMLR, Lille, France, 1530–1538. https://proceedings.mlr.press/v37/rezende15.html

[45] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proceedings of the 31st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 32)*, Eric P. Xing and Tony Jebara (Eds.). PMLR, Bejing, China, 1278–1286. https://proceedings.mlr.press/v32/rezende14.html

[46] Anian Ruoss, Mislav Balunovic, Marc Fischer, and Martin Vechev. 2020. Learning Certified Individually Fair Representations. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 7584–7596. https://proceedings.neurips.cc/paper/2020/file/55d491cf951b1b920900684d71419282-Paper.pdf

[47] Pablo Sanchez-Martin, Miriam Rateike, and Isabel Valera. 2021. VACA: Design of Variational Graph Autoencoders for Interventional and Counterfactual Queries. *arXiv preprint arXiv:2110.14690* (2021).

[48] Moninder Singh and Karthikeyan Natesan Ramamurthy. 2019. Understanding racial bias in health using the Medical Expenditure Panel Survey data. *arXiv preprint arXiv:1911.01509* (2019).

[49] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning Structured Output Representation using Deep Conditional Generative Models. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2015/file/8d55a249e6baa5c06772297520da2051-Paper.pdf

[50] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. 2019. Learning Controllable Fair Representations. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 89)*, Kamalika Chaudhuri and Masashi Sugiyama (Eds.). PMLR, 2164–2173. https://proceedings.mlr.press/v89/song19a.html

[51] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. *West Virginia Law Review* 123, 3 (2021).

[52] Michael Wick, swetasudha panda, and Jean-Baptiste Tristan. 2019. Unlocking Fairness: a Trade-off Revisited. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/373e4c5d8edfa8b74fd4b6791d0cf6dc-Paper.pdf

[53] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2018. FairGAN: Fairness-aware Generative Adversarial Networks. In *2018 IEEE International Conference on Big Data (Big Data)*. 570–575. https://doi.org/10.1109/BigData.2018.8622525

[54] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (Eds.), Vol. 27. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2014/file/375c71349b295fbe2dcdca9206f20a06-Paper.pdf

[55] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) *(WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1171–1180. https://doi.org/10.1145/3038912.3052660

[56] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2019. Fairness Constraints: A Flexible Approach for Fair Classification. *Journal of Machine Learning Research* 20, 75 (2019), 1–42. http://jmlr.org/papers/v20/18-262.html

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

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

[59] Tao Zhang, Tianqing Zhu, Jing Li, Mengde Han, Wanlei Zhou, and Philip S. Yu. 2022. Fairness in Semi-Supervised Learning: Unlabeled Data Help to Reduce Discrimination. *IEEE Transactions on Knowledge and Data Engineering* 34, 4 (2022), 1763–1774. https://doi.org/10.1109/TKDE.2020.3002567

[60] Zhaowei Zhu, Tianyi Luo, and Yang Liu. 2022. The Rich Get Richer: Disparate Impact of Semi-Supervised Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=DXPftn5kjQK