Vigilant Intersectionality: Exposing Unfairness of Ethically Adversarially Trained Models Considering Multiple Protected Attributes

Viney Regunath*
Adam Vandenbussche*
viney.regunath.gr@dartmouth.edu
adam.c.vandenbussche.22@dartmouth.edu
Dartmouth College
Hanover, New Hampshire, USA

Abstract

As machine learning models are increasingly deployed in contexts where they have significant influence over sensitive decisions, such as in criminal justice, lending, and hiring, architects must ensure that these models do not encode and propagate historical biases against vulnerable populations. While existing works have extensively proposed techniques to ensure fairness along single protected attributes, few have explored ways to ensure intersectional fairness, or fairness simultaneously considering multiple protected attributes. Our work builds on the ethical adversary fairness-ensuring training technique proposed by Delobelle et al. After motivating the need for intersectional fairness, we successfully expand their methodology to optimize for fairness along multiple protected attributes simultaneously. Using the COMPAS dataset, we demonstrate that attempting to ensure fairness by solely optimizing models along one protected attribute leaves them susceptible to propagating biases against subgroups holding multiple, intersecting identities.

Keywords: ethical adversary, fairness, intersectional, neural network, protected attribute

ACM Reference Format:
Viney Regunath and Adam Vandenbussche. 2021. Vigilant Intersectionality: Exposing Unfairness of Ethically Adversarially Trained Models Considering Multiple Protected Attributes. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Both authors contributed equally to this work.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

As machine learning models are rapidly deployed in all facets of society, from government and finance to healthcare and education, system designers must pay particular attention to ensure their fairness. Failure to do so risks encoding and propagating the historical biases towards certain demographics due to deeply embedded biases in training data.

Most existing works attempt to quantify or ensure fairness against a single attribute, such as racial identity, sex, gender identity, or nationality [15]. Widely used metrics such as *demographic parity* and *equality of opportunity* help evaluate the effectiveness of fairness-ensuring techniques like *fairness through unawareness*, *individual fairness*, and *counterfactual fairness*.

While ensuring fairness along a single attribute is a starting point to ensuring the ethical and just application of models, such techniques fall short when acknowledging that people cannot be defined by a single identity. As such, there is a need to consider *intersectionality* when examining model fairness: if a model is to truly be deemed fair, one must demonstrate it to be simultaneously fair along each *protected attribute* in the dataset [10]. Works such as that by Buolamwini et al. highlight how ignoring intersectionality when contemplating model fairness can lead to oversight, leaving already marginalized minority groups particularly vulnerable to disparate or disproportional outcomes [3].

One existing fairness-ensuring technique proposed by Delobelle et al. conducts evasion attacks at training time using *ethical adversarial learning* to make models more resilient against discrimination [8]. While a promising approach, the fact that their work only optimizes against a single protected attribute leaves room for improvement in light of the need for intersectionality. In this work, we expand the ethical adversary technique proposed by Delobelle et al. to optimize models for fairness along more than one protected attribute. Our main contributions include:

- a motivation of the need to consider intersectionality when contemplating model fairness; and

- an examination of the discrepancy in protection offered against bias towards one protected attribute compared to two protected attributes following the ethical adversary method proposed by Delobelle et al.

The rest of this paper is organized as follows. Section 2 summarizes core background concepts relevant to the ensuing discussion surrounding intersectional fairness. Section 3 surveys existing literature upon which we base our technique. Section 4 defines our threat model. Section 5 explains our methodology. Section 6 discusses our findings. Sections 7 and 8 evaluate limitations of our work and highlight areas for future exploration, respectively. Finally, Section 9 concludes our work.

2 Background

In this section, we introduce background concepts that are central to any discussion surrounding intersectional fairness.

2.1 Protected Attributes

Protected attributes are qualifiers relating to the personal identity of an individual against which discrimination is prohibited [15]. In the United States, the Fair Housing Act and Equal Credit Opportunity Act codify these attributes to include properties such as age, familial status, gender identity, marital status racial identity, and sex, among others [4, 6, 7, 13, 15]. We define *unprotected attributes* in contrast to protected attributes as attributes against which discrimination is not explicitly prohibited [2]. Conventionally, protected attributes are denoted by A, where a member of the *privileged class* receives $A = 0$ and a member of the *unprivileged class* receives $A = 1$.

2.2 Individuals, Groups, and Subgroups

Following the convention of Yang et al., we define an *individual* as a single person represented in the training data, a *group* as a collection of individuals sharing a common protected attribute, and a *subgroup* as a collection of individuals sharing two or more protected attributes [16].

2.3 The 80% Rule

The *80% rule* is a guideline codified in the Code of Federal Regulations used to detect the violation of anti-discrimination laws such as Title VII of the Civil Rights Act of 1964 [5]. It asserts that legal evidence of discrimination exists if the ratio of the probability of a favorable outcome between a privileged and an unprivileged group is less than 0.8 [10]. In our evaluation, we use the 80% rule as a baseline to detect the presence of unfairness in the studied models.

2.4 Existing Fairness Metrics

Quantifying fairness, being a concept deeply rooted morality, is difficult because morality, too, is subjective. Unfortunately, evaluating the fairness of inherently rigorous applications such as machine learning models demands attempts to do so nonetheless. Several metrics attempting to quantify model fairness have been proposed, most measuring solely along on one protected attribute. Therefore, in practice, evaluating model fairness using several metrics simultaneously ensures a more holistic evaluation. We note several such measures.

Demographic Parity (DP) Demographic parity seeks to minimize the discrepancy between the probability that a privileged individual and an unprivileged individual receive the same classification $\hat{Y} = 1$:

$$DP = \Big| P(\hat{Y} = 1 \mid A = 1) - P(\hat{Y} = 1 \mid A = 0) \Big| \leq \epsilon$$

Fairness improves as $\epsilon \longrightarrow 0$, indicating decreased disparity across demographic lines [9, 13]

Demographic Parity Ratio (DPR) A commonly used variation of DP, the DPR, attempts to quantify this fairness disparity in the form of the ratio of outcomes across privilege levels:

$$DPR = \frac{P(\hat{Y} = 1 \mid A = 1)}{P(\hat{Y} = 1 \mid A = 0)} \geq \tau$$

Fairness improves as $\tau \longrightarrow 1$, indicating the occurrence of increasingly similar outcomes regardless of the value a protected attribute takes on [13]. Cases where $\tau < 0.8$ or $\tau > 1.2$ represent violations of the 80% rule.

Equality of Opportunity (EO) In contrast with DP, EO focuses on the true positive rate of the algorithm under study:

$$EO = \Big| P(\hat{Y} = 1 \mid A = 1, Y = 1) - P(\hat{Y} = 1 \mid A = 0, Y = 1) \Big| \leq \upsilon$$

Fairness improves as $\upsilon \longrightarrow 0$, indicating decreased disparity across demographic lines [11].

3 Related Works

Although fairness evaluation is a relatively new subfield in machine learning research, works proposing techniques to mitigate biases in training data and model outputs have begun to emerge. A recent survey by Mehrabi et al. categorizes these works by their domains and definitions of fairness [15].

In this section, we present existing works relevant to our undertaking. Specifically, we discuss those justifying the need for intersectional fairness, those proposing metrics to measure fairness along multiple protected attributes, and those offering techniques to ensure fairness during training.

3.1 Need for Intersectional Fairness

As part of their survey, Mehrabi et al. distinguish between works attempting to ensure fairness towards groups, individuals, and subgroups. In particular, they highlight a gap in research examining fairness vulnerabilities that affect subgroups. This oversight can have significant consequences for people as these models begin to influence real-world decisions, namely because people are not defined by a singular identity but rather by the intersection of several identities. Existing works have examined the consequences of this oversight. For example, Buolamwini et al. compared the accuracy

of a widely used facial recognition model on visually diverse faces. Their work highlighted how considering gender identity and not race while training a facial recognition model overlooked the model's poor performance on dark-skinned, female faces compared to light-skinned ones [3]. A hasty deployment of such a biased model could lead to the further exclusion of an already marginalized demographic.

3.2 Existing Intersectional Fairness Metrics

Recognizing the need for intersectionality, some works have sought to expand upon existing fairness metrics by measuring model fairness along multiple protected attributes simultaneously. We discuss two such measures.

Statistical Parity Subgroup Fairness (SF) Kearns et al. proposed SF as a metric aimed to reveal discrimination towards certain subgroups [12]. Given a collection $\mathcal{G}$ of protected attributes $g : A \rightarrow \{0, 1\}$, where $g(\mathbf{s}) = 1$ signifies that an individual with protected attributes $\mathbf{s}$ is in group g, and a binary classification mechanism $M(\mathbf{x})$, then $M(\mathbf{x})$ is said to be γ-SF fair with respect to training parameters θ and $\mathcal{G}$ if $\forall\, g \in \mathcal{G}$ we have that

$$| P_{M,\theta}(M(\mathbf{x}) = 1) - P_{M,\theta}(M(\mathbf{x}) = 1 \mid g(\mathbf{s} = 1)) | \times P_\theta(g(\mathbf{s}) = 1) \leq \gamma.$$

Foulds et al. took issue with SF, arguing its scaling of fairness by the prevalence of the subgroup in the training data (represented by the $P_\theta(g(\mathbf{s}) = 1)$ term) would ineffectively protect often-marginalized minority groups holding intersecting identities [10].

Differential Fairness (DF) As an improvement to SF, Foulds et al. proposed DF, which similarly seeks to quantify fairness along intersecting identities while also assuring protection for minorities [10]. Given tuples of all protected attributes $\mathbf{s}_i, \mathbf{s}_j \in A$ and the set Θ of all possible distributions θ which could plausibly generate each instance $\mathbf{x}$, model $M(\mathbf{x})$ is said to be ϵ-DF with respect to (A, Θ) if $\forall\, \theta \in \Theta$ with $\mathbf{x} \sim \theta$ and $y \in \text{Range}(M)$, $\forall\, (\mathbf{s}_i, \mathbf{s}_j) \in A \times A$ where $P(\mathbf{s}_i \mid \theta) > 0$ and $P(\mathbf{s}_j \mid \theta) > 0$ we have that

$$e^{-\epsilon} \leq \frac{P_{M,\theta}(M(\mathbf{x}) = y \mid \mathbf{s}_i, \theta)}{P_{M,\theta}(M(\mathbf{x}) = y \mid \mathbf{s}_j, \theta)} \leq e^{\epsilon}.$$

DF asserts that probabilities of outcomes should be similar regardless of the combination of protected attributes. Another advantage of DF over SF is that the former can be compared to the 80% rule by setting $\epsilon = -\log 0.8 \approx 0.223$ whereas the latter cannot [10]. Henceforth, we refer to the natural exponentiation of negative DF as the differential fairness ratio (DFR):

$$DFR = \exp(-DF)$$

3.3 Existing Fairness-Ensuring Techniques

Many existing works have explored techniques to minimize the propagation of historical biases against groups and subgroups. Mehrabi et al. provide a comprehensive survey of the proposed methods, but we provide a sampling here to better contextualize our work [15].

Fairness Through Unawareness (FTU) Chen et al. proposed FTU as a naive approach to ensure model fairness by removing protected attributes from a model's features altogether. The rationale behind their technique is that it is impossible to discriminate along attributes that are not explicitly included in a decision-making process [4]. However, FTU has been criticized as being ineffective for neglecting to acknowledge how historical biases may be propagated by basing decisions on unprotected attributes that act as proxies for protected attributes. For example, ZIP codes can be proxies for race in the United States given historic housing segregation policies [13].

Individual Fairness (IF) In contrast to FTU, Dwork et al. proposed IF as a way to explicitly ensure similar predictions for similar individuals. They first defined a distance metric $d(\cdot, \cdot)$ to quantify this similarity, then they optimized the model to ensure that the discrepancy between their two predictions are approximately proportional to this metric [9]. This technique requires that $d(\cdot, \cdot)$ be mindfully crafted given the model's intended application so as not to introduce further biases through flawed design.

Counterfactual Fairness (CF) Kusner et al. proposed CF as a rigid, explicitly causal framework to ensure model fairness [13]. The technique seeks to mindfully use knowledge of protected attributes to infer unbiased latent attributes that are actually relevant to an application based on potentially biased data. In other words, the technique can be thought of as one that compensates for historical biases.

Adversarial Learning Delobelle et al. [8] proposed a method using an ethical adversary to train neural networks using evasion attacks in order to ensure fairness. As illustrated in Figure 1, this technique is bipartite: (1) a *Feeder* model uses evasion attacks to produce adversarial examples highlighting unfair representation of a group within the training data, and (2) a *Reader* model attempts to infer the values of a protected attribute. These two models iteratively work together to optimize for both fairness and utility in a two-step process.

First, the pre-trained *Target* model to be optimized, whose goal is to predict a main attribute Y, is connected to the adversarial Reader, which in turn attempts to predict the protected attribute A of an input X while a gradient reversal layer attempts to minimize the confidence of these predictions. The Target is trained with a joint loss of the original classification task and the protected attribute.

Second, the Feeder generates a set of adversarial examples by performing evasion attacks on an approximative *Surrogate* model trained on the same dataset as the Target (in our case, following Delobelle et al., the Surrogate was a support vector machine with a radial basis function kernel). These adversarial examples are ideally similar to training examples but are reliably misclassified (the perturbation allowed by the

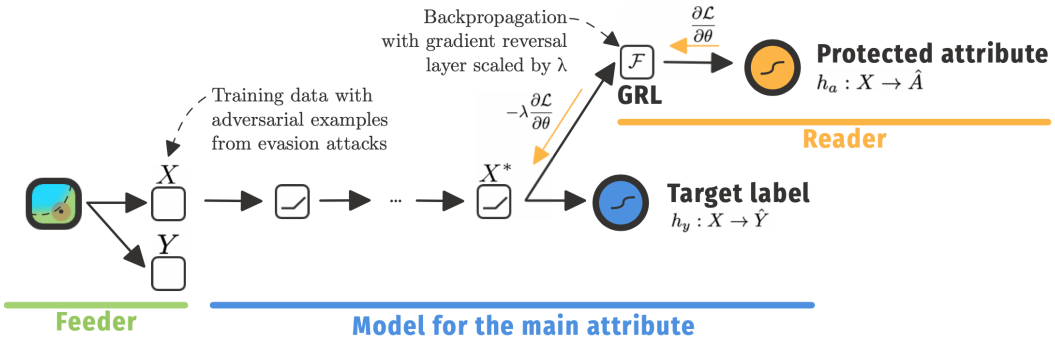


Figure 1. Architecture of the ethical adversary method proposed by Delobelle et al. Diagram taken from Figure 1 of [8].

Feeder is constrained to ensure similarity between original and adversarial examples).

At each iteration, the generated adversarial examples are incorporated into the Target's training set before retraining it. The same label is used for each adversarial example as the original example from which it was derived. Theoretically, by repeating this process several times, utility and fairness can be simultaneously improved.

Delobelle et al. evaluated the effectiveness of their methodology on the widely used COMPAS, German Credit, and Adult Census datasets. Their empirical findings supported their conceptualization: by most metrics across most trials, their technique offered the strongest fairness guarantees with minimal utility loss. A shortcoming of their technique, however, is that it only enables optimization against a single protected attribute.

We based our work on this ethical adversary method proposed by Delobelle et al. Our efforts sought to expand on theirs by enabling simultaneous optimization against multiple protected attributes and by quantifying fairness along their intersection.

4 Threat Model

We identified two primary actors to consider in our threat model, namely a *ethical adversary trainer* and a *model auditor*.

The ethical adversary trainer seeks to improve the fairness of the Target model by carrying out the adversarial attack outlined in Section 3.3, which is a black-box process. Similar to the trainer in Delobelle et al., we assume that this "adversary" has access to both the pre-trained Target model as well as its original training data [8].

The model auditor behaves less like an "adversary," seeking to merely detect unfairness in the Target model. Unlike the trainer, the model auditor can only probe the model for a label given a particular input and use fairness metrics such as those presented in Sections 2.4 and 3.2 to evaluate its behavior.

5 Methodology

Our work seeks to examine and quantify the discrepancy in protection offered by the Delobelle et al. technique against bias towards one binary protected attribute and two binary protected attributes. Fortunately, the authors published their codebase at github.com/iPieter/ethical-adversaries, so we could adapt it as needed for our investigation. In order to facilitate future explorations, we ensured that our modifications were modular and abstracted beyond the scope of this particular work. The updated version of the Delobelle et al. codebase with our modifications is hosted at github.com/avandenbussche/ethical-adversaries.

The original codebase assumed there would only be a single binary one-hot-encoded protected attribute for any dataset to optimize against. As such, we first needed to adapt the original code so it could accept more than one binary protected attribute. To facilitate future expansions, we abstracted the logic behind the number of protected attributes.

Second, we wanted to measure the model's performance according to the DFR metric, which was not included by Delobelle et al. in their original work. Moreover, we needed to modify the code so as to allow the measuring of fairness both along the *optimized* protected attributes, against which the model was trained to ensure fairness, as well as along *unoptimized* attributes; the original code was only capable of the former. Fortunately, Foulds et al. published their code to compute DF at github.com/rashid-islam/Differential_Fairness. As such, we were able to easily integrate their code into our codebase [10]. Once we computed DF, we trivially computed DFR as outlined in Section 2.4 for comparison with the 80% rule.

To more intuitively measure intersectional fairness, we computed DPRs directly comparing outcomes for two distinct subgroups. For example, consider a dataset with *race* (R) and *sex* (S) as protected attributes. Following the convention of setting $A = 0$ for privileged groups (e.g., men, white) and $A = 1$ for unprivileged groups (e.g., women, people of color), we could denote white men as `S0R0` and women of color as

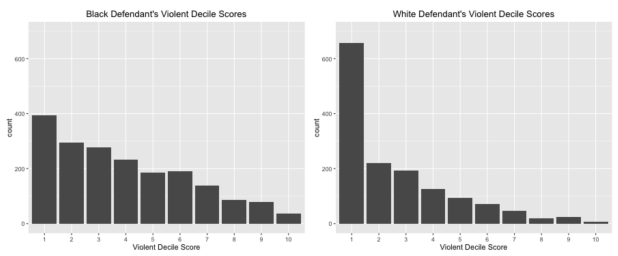


Figure 2. Comparison of COMPAS decile score distributions for African-American (left) and Caucasian (right) defendants. Higher scores indicate higher predicted risk of recidivism. Note that the distribution for African-American defendants is skewed towards higher risk scores. Figure taken from [14].

`S1R1`. As such, measuring `DPR (S1R1/S0R0)` would quantify the discrepancy of outcomes according to DPR between women of color and white men, where a $DPR < 0.8$ would indicate failure to comply with the 80% rule and discrimination towards women of color, being an often marginalized demographic in the West.

Finally, in the spirit of ensuring ease of experimentation and eventual expansion, we sought to render all of our modifications accessible via the original codebase's command line interface. For example, to set the protected attributes to optimize and to measure against, respectively, the user could set the following arguments:

```
-optimize-attribute "race,sex"
-measure-attribute "race"
-measure-attribute "sex"
```

where *race* and *sex* are valid attribute names in the training dataset.

6 Evaluation

In this section, we outline how we verified our expansion of the ethical adversarial method proposed by Delobelle et al.

6.1 Dataset Description

Similar to Delobelle et al., we used the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) dataset to evaluate our methodology. The dataset contains information about criminal defendants in Broward County in Florida, including demographic information such as their race, sex, and age range as well as their criminal history, the amount of time they spent in jail or prison, and whether they had reoffended when released while awaiting sentencing. Using this information, a model trained on the dataset could produce a decile score predicting the likelihood of a defendant reoffending while awaiting trial. Judges would use this score to help decide whether to grant bail to defendants while awaiting trial.

Figure 3. Disparity in COMPAS decile risk score for two defendants with similar criminal histories, Dylan Fugett (left) and Bernard Parker (right), who only meaningfully differ in racial identity. Both were arrested for drug possession. Fugett's prior offense was one instance of attempted burglary while Parkers's prior offense was one instance of resisting arrest without violence. Fugett received a COMPAS score of three and was granted bail while Parker received a score of ten and was not. Fugett was subsequently arrested three more times on drug charges [1]. Figure taken from [1].

The dataset gained notoriety after a 2016 ProPublica investigation found it to be nearly twice as likely to misclassify African-Americans as higher risk than Caucasian defendants [1, 14].[1] Figures 2 and 3 illustrate this racial disparity. The investigation also found that defendants younger than 25 years old were 2.5 times as likely to get a higher score than older offenders, and that female defendants were 1.2 times as likely to get a higher score than men [14].

The prevalence of biases in the dataset led to COMPAS's widespread use in literature to evaluate the efficacy of techniques to mitigate model unfairness. While the dataset contains subjects identifying as African-American (51.4%), Caucasian (34.1%), Hispanic (8.3%), Asian (0.5%), Native American (0.2%), and with other races (5.6%), we only included subjects identifying as either Caucasian or African-American in our analysis due to the relatively low representation of other racial identities. Ultimately, our training dataset contained 5278 individuals represented by 12 features, including

[1]The COMPAS dataset uses the terms *African-American* and *Caucasian* to refer to the racial identities of all Black- and white-identifying subjects, respectively. While many of the Black-identifying subjects may very well have also personally identified as African-American at the time of the dataset's compilation, we take issue with the lumping of all Black-identifying subjects under the African-American identity, particularly given Broward County's large African and African-Caribbean communities. These identities do not appear to have been collected by the dataset's compilers [1, 14]. In this work, we continue using the umbrella terms *African-American* and *Caucasian* for all Black- and white-identifying subjects as these are the terms that are used in the dataset, but we would be remiss to neglect this detail in a paper discussing intersectional identities.

race and *sex* as binary protected attributes, where

$$race \in \{ \text{African-American}, \text{Caucasian} \}$$
$$sex \in \{ \text{Female}, \text{Male} \}.$$

As discussed in Section 8, we focused our efforts on the COMPAS dataset for its intuitive sense of fairness: it would be easier to validate our results given previous analyses preformed on the COMPAS dataset; working with datasets where the extent of the present biases are less understood could have led us astray. We left explorations of our technique on the Adult Census and German Credit datasets, which were also used by Delobelle et al., to future endeavors.

6.2 Experimental Setup

We conducted experiments using a neural network with three hidden layers of 32 neurons each. Each hidden unit used ReLU activation to mitigate vanishing gradients, while we used a linear activation for the output neurons. We trained the system using the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.9999$ as well as a learning rate $l_r = 0.1$ that we adjusted by a factor of 0.1 when the model reached a plateau. This setup is identical to that used by Delobelle et al. [8].

We designed two experiments to validate our methodology. These experiments were largely identical in execution but differed in their hyperparameters so as to better understand their effect on model optimization. Namely, as in Delobelle et al., these hyperparameters include λ, which tunes the tradeoff between utility and attack strength as illustrated in Figure 4 of the original paper, the *batch size*, or the number of adversarial examples that are added to the training dataset at each attack iteration, and the number of *epochs*, or the number of passes used to train the model at each attack iteration.

Experiment 1 copied the parameters used by Delobelle et al. in their paper [8]. Under this experiment, we set $\lambda = 50$ with 100 training epochs, a training batch size of 1024, and an injection of 50 adversarial points per attack iteration. We ran 40 attack iterations to reach an adversarial concentration of over 50%.

Experiment 2 was based on the default parameters left by Delobelle et al. in their GitHub repository [8]. Under this experiment, we maintained $\lambda = 50$, we set the training batch size to 128, and we injected ten adversarial points per attack iteration. We again executed 40 attack iterations. However the model only reached an adversarial concentration of 10% given fewer attack points per iteration. These experimental parameters offered a more granular view into the attack's early behavior, albeit requiring less computational power than Experiment 1 because of the need to generate fewer adversarial examples.

For each experiment, we optimized the models along a different set of protected attributes $A \in \mathcal{A}$, where

$$\mathcal{A} \in \{ \{race\}, \{sex\}, \{race, sex\} \}.$$

Henceforth, we denote trials optimizing solely along *race* ($A = \{race\}$) with *O: R*, trials optimizing solely along *sex* ($A = \{sex\}$) with *O: S*, and trials optimizing along the intersection of *race* and *sex* ($A = \{race, sex\}$) with *O: I*. When merely comparing a metric against a set of unoptimized protected attributes, we denote the measured values with *C: R*, *C: S*, and *C: I*, receptively. We repeated each trial five times, averaging the results of each for analysis.

6.3 Evaluation Metrics

In their original work, Delobelle et al. measured model fairness according to EO, DP, and DPR along the primary protected attribute of each dataset they examine (e.g., *race* for COMPAS). They measured utility by computing the model's accuracy and F1 scores, both across the entire data distribution as well as within a 95% confidence interval in order to ensure that the resulting fair models would still be useful for their intended applications. Finally, they also logged the *adversarial fraction*, or the prevalence of examples in the Target's training data that were generated by the Feeder model.

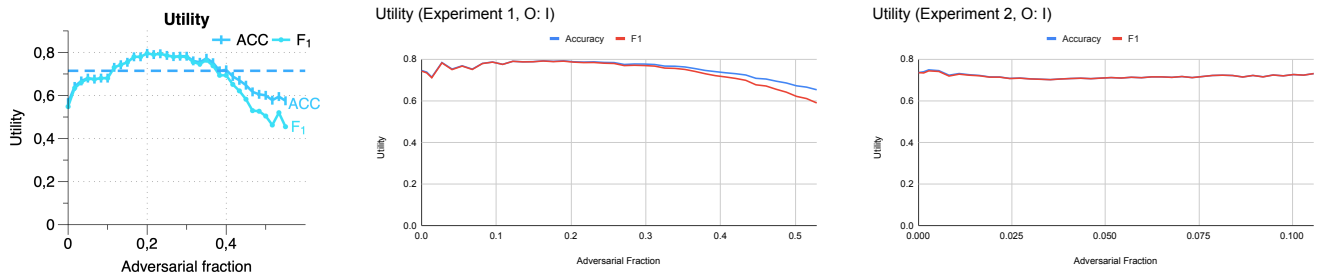
In addition to these metrics, we measured model performance by DF and DFR as well as by subgroup-pairwise DPR, which we introduced in Section 5. Given the context of the COMPAS dataset, this DPR measurement offers a direct measurement of disparate rates of being labelled "high-risk" across intersectional lines. For example, a `DPR (S1R0/S1R1)` of 0.56 would imply that Caucasian women are only 56% as likely to be classified as high-risk as African-American women. Such insight provides a much more intuitive interpretation of model fairness compared to DFR on its own. To avoid redundancy and to facilitate interpretation, we only computed these pairwise DPR measurements with the perceived privileged subgroup in the "numerator" and the perceived unprivileged subgroup in the "denominator" so as to anticipate an initial value of $DPR < 1$.

By tracking model fairness and utility as the adversarial fraction increases, we hoped to gain a clearer picture of any tradeoffs that appear during training.

6.4 Experiment Results

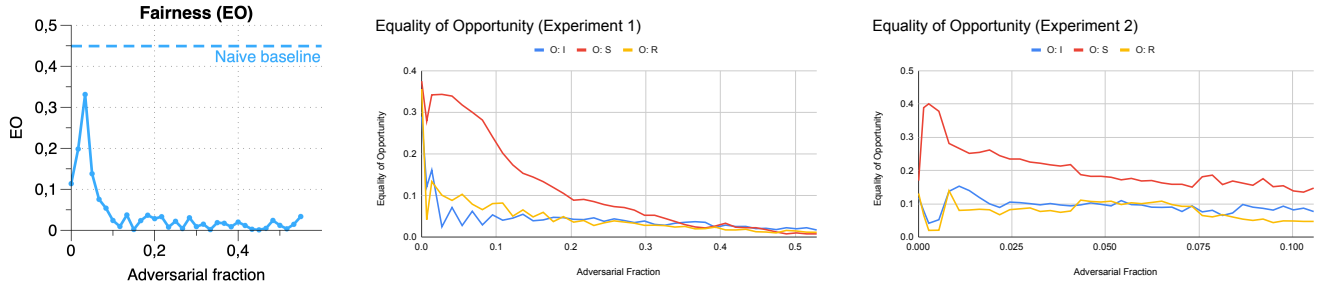
In this section we evaluate the findings of our experiments across each of the studied metrics.

Utility Figure 4 illustrates the utility of the models optimized for fairness along the intersection of *race* and *sex* by measuring accuracy and F1 scores as adversarial points are progressively added to the training dataset. While we only include the intersectional case in the figure, we observed similar situations across both experiments when optimized solely on *race* or *sex*. The figure shows how model utility peaks at approximately 25% concentration of adversarial examples in the training dataset in both our results for Experiment 1 and those of Delobelle et al., but then drops off



(a) Delobelle et al. (b) Experiment 1 (c) Experiment 2

Figure 4. Model utility when optimized for fairness along (a) *race* using the method by Delobelle et al. and (b, c) along the intersection of *race* and *sex* using our method across Experiments 1 and 2, respectively. We observed similar respective trends for each model when optimized solely along *race* and *sex* using our method. Subfigure (a) taken from Figure 3 (a) of [8].



(a) Delobelle et al. (b) Experiment 1 (b) Experiment 2

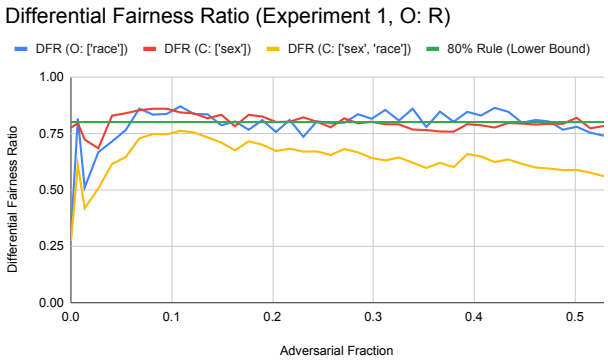
Figure 5. Equality of opportunity found by (a) Delobelle et al. when optimized for fairness along *race*, and (b; c) our method according to Experiment (1; 2) and by optimized attribute (intersection of *race* and *sex*: I; *race*: R; or *sex*: S). Subfigure (a) taken from Figure 3 (c) of [8].

as this concentration increases further. Delobelle et al. explained that utility decreases in this way because after a certain adversarial fraction, marginal adversarial examples merely add noise. Although long-term behavior in the case of Experiment 2 remains unclear, its utility appears consistent with the near-peak levels observed in Experiment 1.

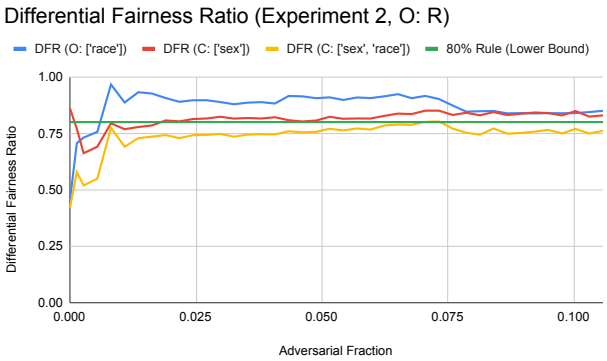
Equality of Opportunity Figure 5 shows the EO of the model as the adversarial example concentration increases. Recall that fairness by EO increases as the metric tends towards zero. Both the results of Delobelle et al. and our results across all trials show that, as expected, fairness by EO improves as the model is optimized, regardless of the optimized protected attributes. The models optimized by solely *sex* appear to optimize along this metric more slowly than the intersectional and solely *race* cases, which appear to converge at similar rates. Convergence between the results of Delobelle et al. and our Experiment 1 appears to occur at similar rates. As found by Delobelle et al., fairness in all cases significantly outperforms the naive baseline for the COMPAS dataset.

Differential Fairness Figure 6 illustrates the DFR metrics across both experiments according to the three optimization cases, measuring fairness along both the attribute(s) being optimized against as well as those not optimized against for comparison. As explained in Section 2.3, we used the 80% rule as the threshold to determine fairness. Generally, we noted similar trends across both experiments.

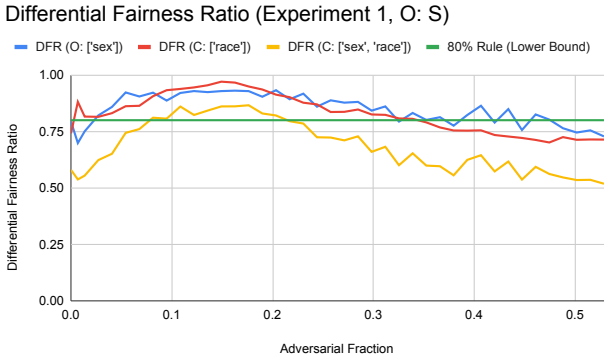
The intersectional measurement was consistently the lowest DFR measurement in absolute terms compared to when measuring optimizing against a sole protected attribute. We found this unsurprising, as we would expect the unfairness present along individual protected attributes to compound when measuring across multiple protected attributes simultaneously. Thus, even when the intersectional DFR measurement failed to achieve the 80% threshold, the individual attributes along which intersectionality is considered may achieve fairness within the 80% threshold. However, this phenomenon is not guaranteed. Foulds et al. asserted that only ensuring fairness at the intersectional case is sufficient to ensure fairness along each individual protected attribute [10]. Of all our trials, only the intersectional case for Experiment 2 achieved this requirement for virtually the entire duration of the attack. However, we acknowledge that this experiment also studies a shorter attack than Experiment 1.



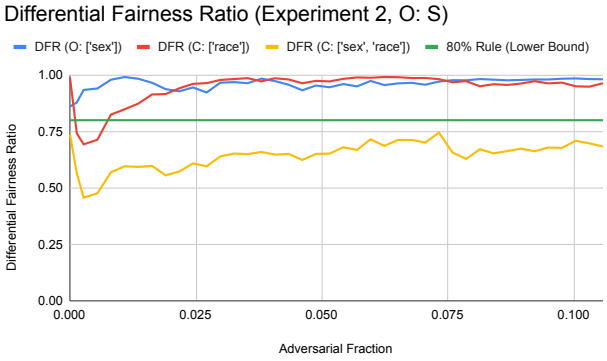
(a) Experiment 1, optimized on *race*



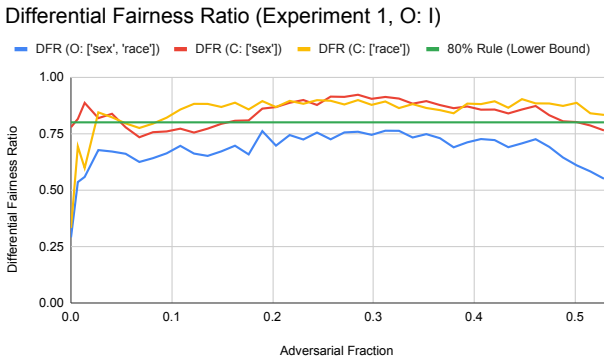
(b) Experiment 2, optimized on *race*



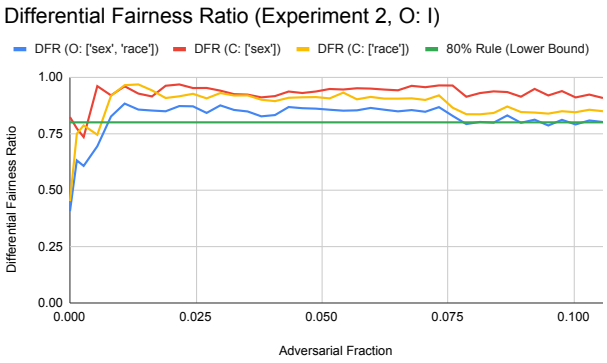
(c) Experiment 1, optimized on *sex*



(d) Experiment 2, optimized on *sex*



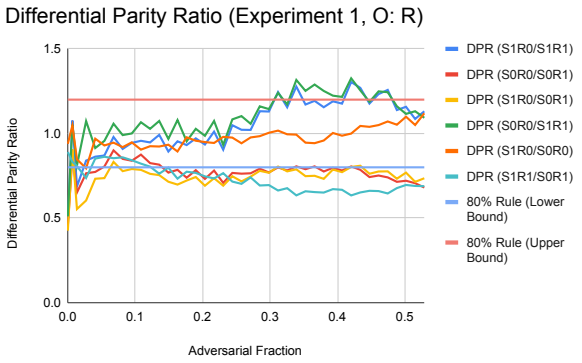
(e) Experiment 1, optimized on *race* and *sex*



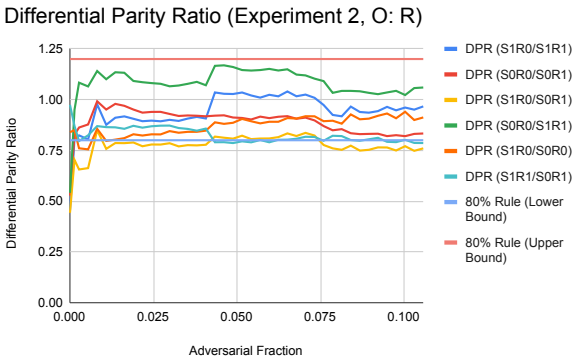
(f) Experiment 2, optimized on *race* and *sex*

Figure 6. Differential fairness ratios of models optimized by our method according to Experiment (1; 2) and by optimized attribute.

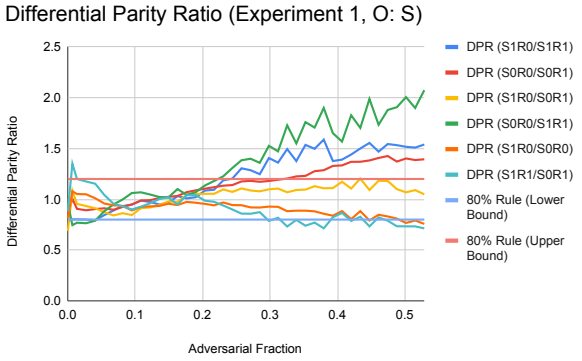
When optimizing solely along *race*, we noted only borderline adherence to the 80% rule for individual attributes across both experiments. When optimizing solely along *sex*, in Experiment 1 we interestingly noted a peak in DFR for all three measurement cases at around 15% concentration of adversarial points followed by a rapid deterioration as the attack continues. In Experiment 2, we noted the strongest individual fairness along *race* and *sex* respectively but also some of the weakest performance along the intersectional measurement. When optimizing along intersectional lines, we observed the most consistent adherence to the 80% rule for the entire attack duration across all three measurements. This adherence confirmed that the optimization of the model



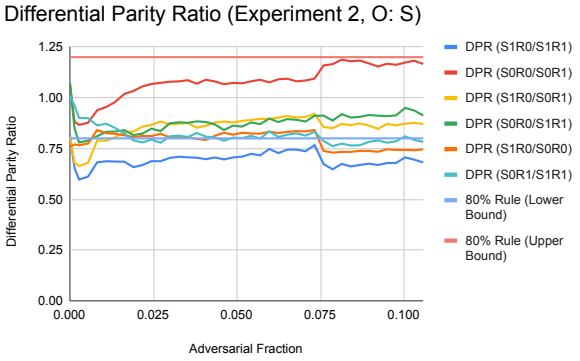
(a) Experiment 1, optimized on *race*



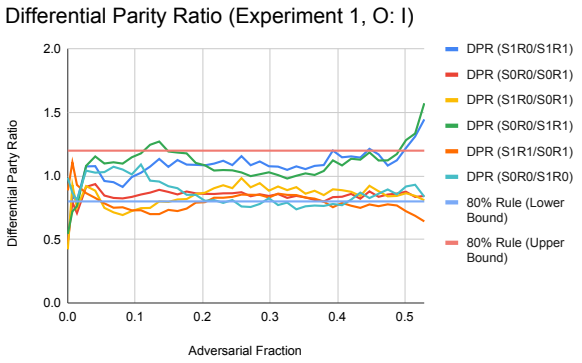
(b) Experiment 2, optimized on *race*



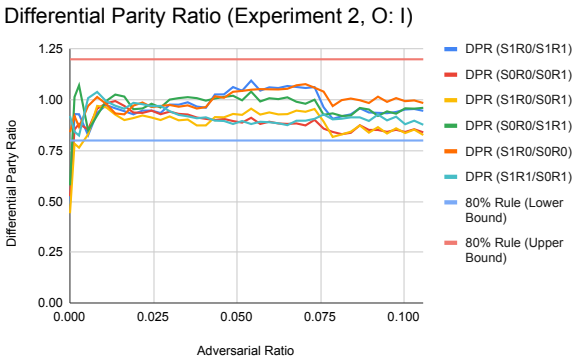
(c) Experiment 1, optimized on *sex*



(d) Experiment 2, optimized on *sex*



(e) Experiment 1, optimized on *race* and *sex*

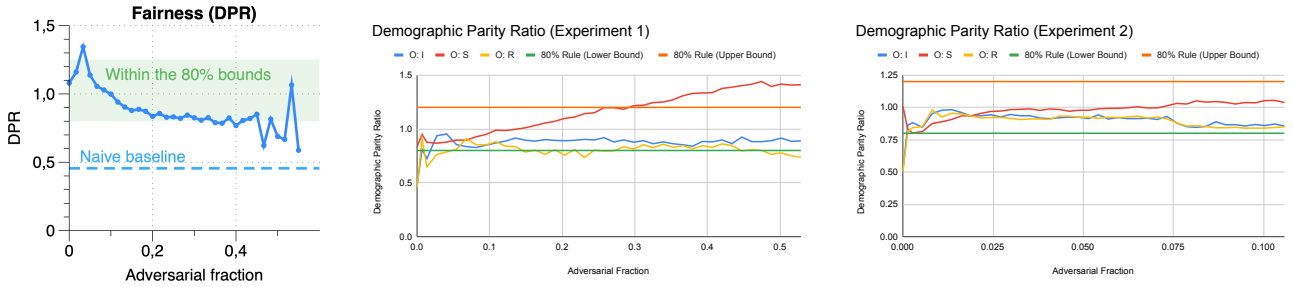


(f) Experiment 2, optimized on *race* and *sex*

Figure 7. Subgroup pairwise demographic parity ratios of models optimized by our method according to Experiment (1; 2) and by optimized attribute (intersection of *race* and *sex*: I; *race*: R; or *sex*: S). As explained in Section 5, following the convention of setting $A = 0$ for privileged groups and $A = 1$ for unprivileged groups, we denote Caucasian men and women as `S0R0` and `S1R0`, respectively, and African-American men and women as `S0R1` and `S1R1`, respectively.

was working as expected. Finally, we observed that the deterioration across all DFR measurements in Experiment 1 correlates with the deterioration in model utility illustrated in Figure 4 (b). This trend could be due to noise from excessive adversarial examples.

Demographic Parity Ratios We validated our observations of performance according to DFR by examining the more intuitive metric of pairwise DPRs. Figure 7 illustrates trends in DPR measurements directly comparing outcomes for pairs of subgroups (e.g., comparing African-American



(a) Delobelle et al. (b) Experiment 1 (b) Experiment 2

Figure 8. Model fairness measured by demographic parity ratio when optimized for fairness along *race* following method by Delobelle et al. Subfigure (a) taken from Figure 3 (b) of [8].

Table 1. Pre-optimization pairwise demographic parity ratios for Experiment 1, when optimized by *race*. Values represent ratio of probability that subgroup X is labelled as "high-risk" compared to the same probability for subgroup Y. Similar values were observed for other trials when optimizing by *sex* and along the intersection of *sex* and *race*, as well as over Experiment 2.

Subgroup X \ **Subgroup Y**	African-American Women	African-American Men	Caucasian Men
Caucasian Men	51%	45%	-
Caucasian Women	48%	43%	93%
African-American Women	-	89%	-

women to Caucasian women) when optimizing along *race*, *sex*, and the intersection of both.

Table 1 provides a snapshot of a sample pre-optimization state of the models, which can also be seen at the initial data points (when adversarial fraction is 0) in Figure 7. We observed that initial disparities across *race* are blatantly obvious and in clear violation of the 80% rule.

Generally, DPR performance under Experiment 1 improved at the beginning of the attack but devolves after a certain adversarial example concentration. Under Experiment 2, performance was more constrained to within the 80% threshold, albeit to varying degrees, but long-term behavior admittedly remained unobserved.

When optimizing solely along *race*, under Experiment 1 we discovered improved fairness towards the beginning of the attack. However these improvements devolved at approximately 40% adversarial concentration, at which point we observed significant disparities. Interestingly, these disparities were in some cases flipped from the initial conditions in that subgroups initially benefiting from unfairness are eventually discriminated against. Specifically, Caucasian men and women became deemed as high-risk at higher rates than African-American women, but African-American men became deemed as high-risk at higher rates than Caucasian men and women. When we averaged over both racial identities, neither was consistently discriminated against over the other; although in an ineffective way, one could argue fairness had been achieved. Under Experiment 2, we saw more stable performance (albeit for a shorter attack), and African-American men became deemed high-risk at higher rates than Caucasian women. Most of these instances of discrimination occured as a result to differences in *sex*, which makes sense when we optimize the model solely to be fair along *race*.

When optimizing solely along *sex*, nearly perfect fairness appeared to be achieved at approximately 15% under Experiment 1. However, this performance rapidly deteriorated as the attack continues. By the end of the attack, Caucasian men and women became deemed high-risk at significantly higher rates than African-American women, and Caucasian men became deemed high-risk at significantly higher rates than African-American men. Under Experiment 2, African-American women became deemed high-risk at higher rates than Caucasian women, in violation of the 80% rule, and Caucasian men became deemed high-risk at higher rates than African-American men. Again, most of these instances of discrimination occured as a result of the differences in *race*. This racial bias made sense when optimizing the model for fairness solely along *sex*.

Finally, when optimizing along the intersection of *race* and *sex*, we saw the most sustained adherence of the 80% rule of all trials across both Experiments. At the end of the attack under Experiment 1, however, we observed a sudden divergence from fairness as the adversarial concentration approaches 50%. As a result of this deviation, above 50% adversarial concentration, Caucasian men and women became deemed high-risk at higher rates than African-American

women, and African-American men became deemed high-risk at higher rates than Caucasian men. This decrease in performance could possibly have been attributed to the drop in utility due to excess adversarial noise.

We then compared our observations with those found by Delobelle et al. In addition to measuring EO, Delobelle et al. evaluated the effusiveness of their methodology by examining the *race*-optimized model's performance per DPR along this protected attribute (specifically, comparing the likelihoods of Caucasians to African-Americans of being labeled as high-risk). For comparison, we computed the DPR by *race* for our models regardless of the protected attributes they are optimized against. Figure 8 illustrates these results. Delobelle et al. observe an early peak in DPR followed by a gradual plateau towards 80% before noisy oscillations as the adversarial fraction continues to increase. Under our Experiment 1, DPR for the intersectional- and *race*-optimized models quickly plateaued between 0.8 and 1.0 once the attack begins, while the *sex*-optimized racial DPR unsurprisingly saw fairness diverge as the attack continues. We observed similar behavior for Experiment 2, albeit with less insight into its long-term behavior.

Key Takeaways Ultimately, our findings underscored the importance of considering intersectionality when evaluating model fairness. When only focusing on fairness along one protected attribute, such as *race*, biases along another, such as *sex*, may be overlooked. A broader-view comparison of the results from Experiments 1 and 2 suggests that, at least in the context of the COMPAS dataset, the parameters of the latter might offer a fair solution with fewer attack iterations, each requiring less computational intensity.

7 Limitations

There are several limitations to our work. First, resource and time constraints hindered the validation of our methodology on more than just the COMPAS dataset. Given more time and processing power, we would have attempted to use the Adult Census and German Credit datasets as Delobelle et al. did for further analysis. Although we successfully adapted the Delobelle et al. codebase for analysis with the Adult Census dataset, the dataset's more complex feature set required a denser model that took too long to attack given the computational resources available to us (Delobelle et al. trained a neural network with three hidden layers of 128 neurons each on this dataset, compared to the three layers of 32 neurons each required by COMPAS as explained in Section 6.2) [8]. Despite attempts to reduce the complexity the feature set from 146 to 108 features (after one-hot encoding) by consolidating redundant attributes, we found the attack far too slow to gather sufficient data for analysis.

Second, Experiment 2 reached a far lower adversarial threshold than Experiment 1 due to a decreased attack size at the same number of attack iterations. Because the experiment offered less insight into long-term behavior for that parameter configuration, it remains unclear whether increased fairness could be sustained with fewer, less computationally intensive attack iterations.

Finally, our attack architecture presupposed a restricted definition of identities. In requiring binary protected attributes, we neglected to acknowledge that (a) there are often more than two values a protected attribute could take on (e.g., there are multiple racial, gender, and national identities), and (b) an individual may simultaneously identify with more than one identity, and perhaps to varying extents. Despite still improving on the work of Delobelle et al., our technique falls short of truly offering meaningful and realistic intersectional fairness.

8 Future Work

Much remains to be studied in the field of intersectional fairness, and our work offers several starting points for future explorations.

First, our methodology should be verified using a dataset featuring more than two protected attributes. In many instances, vulnerable demographics may identify with several protected attributes at the intersection of which discrimination could occur (e.g., age, racial identity, and gender identity). Ideally, model fairness would be demonstrated for all mandated protected attributes such as those defined by the Fair Housing Act and Equal Credit Opportunity Act, assuming these protected attributes are not believed to be relevant to the application. For example, it may be unwise to ensure that a medical diagnosis model be fair along attributes such as ethnicity or sex if it is believed that a certain demographic is more susceptible to a particular disease, but such a case could not be morally made for a lending algorithm.

Second, as explained in Section 7, our technique must be expanded to enable optimization against non-binary protected attributes.

Third, we would like to further investigate the behavior of each hyperparameter offered by Delobelle et al., especially in intersectional cases. Our investigation of the tradeoffs between attack strength and efficacy could be expanded by investigating the long-term behavior of Experiment 2's hyperparameters and by varying λ. Figure 4 of Delobelle et al. offers a Pareto analysis illustrating tradeoffs between accuracy, demographic parity, and λ serves as an example of how this could be explored [8]. Moreover, deriving a method to identify the optimal adversarial fraction would help avoid over-optimization, which we have shown models to be prone to.

Finally, it would be interesting to explore the combination of our methodology with the counterfactual method proposed by Kusner et al. This potential relationship could ensure intersectional fairness backed by a rigid framework [13].

More broadly, there may exist other works whose combination with our efforts would yield interesting results.

9 Conclusion

In this paper, we built on the ethical adversary training technique proposed by Delobelle et al. We began by motivating the need for intersectionality when contemplating model fairness and by highlighting the shortcomings of existing fairness metrics. We then expanded the work of Delobelle et al. by optimizing models for fairness along two protected attributes as opposed to just one. Our efforts were largely successful. We demonstrated how only considering fairness along one protected attribute where multiple exist risks overlooking biases along another protected attribute, and proved that intersectional optimization can be performed without significant loss to utility. Furthermore, we showed that it may be possible to over-optimize models for fairness, which is perhaps not an intuitive conclusion, and offered several opportunities for further research.

As explained in Section 2.4, fairness is an inherently human instinct and is thus difficult to encode in rigid contexts such as machine learning. Our work ultimately underscores the need for further endeavors to meaningfully embody it in models so as to ensure their ethical application.

References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *ProPublica* (May 2016).

[2] Yazeed Awwad, Richard Fletcher, Daniel Frey, Amit Gandhi, Maryam Najafian, and Mike Teodorescu. 2020. *Exploring fairness in Machine Learning for international development.* Technical Report. CITE MIT D-Lab.

[3] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[4] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness Under Unawareness. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Jan 2019). https://doi.org/10.1145/3287560.3287594

[5] Equal Employment Opportunity Commission. 1978. 29 CFR 1607 - Uniform Guidelines on Employee Selection Procedures.

[6] United States Congress. 1968. Fair Housing Act. *42 U.S.C. 3601* (1968). https://www.justice.gov/crt/fair-housing-act-1

[7] United States Congress. 1974. Equal Credit Opportunity Act. *15 U.S.C. 1691* (1974). https://www.justice.gov/crt/equal-credit-opportunity-act-3

[8] Pieter Delobelle, Paul Temple, Gilles Perrouin, Benoît Frénay, Patrick Heymans, and Bettina Berendt. 2020. Ethical Adversaries: Towards Mitigating Unfairness with Adversarial Machine Learning. arXiv:2005.06852 [cs.LG]

[9] 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

[10] James Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2019. An Intersectional Definition of Fairness. arXiv:1807.08362 [cs.LG]

[11] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. arXiv:1610.02413 [cs.LG]

[12] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 2564–2572. https://proceedings.mlr.press/v80/kearns18a.html

[13] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *arXiv preprint arXiv:1703.06856* (2017).

[14] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)* 9, 1 (2016).

[15] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6, Article 115 (July 2021), 35 pages. https://doi.org/10.1145/3457607

[16] Forest Yang, Moustapha Cisse, and Sanmi Koyejo. 2020. Fairness with Overlapping Groups. arXiv:2006.13485 [cs.LG]