Ethical Adversaries

Delobelle, Pieter; Temple, Paul; Perrouin, Gilles; Frénay, Benoît; Heymans, Patrick; Berendt, Bettina

Published in:
SIGKDD Explorations

DOI:
10.1145/3468507.3468513

Publication date:
2021

Document Version
Peer reviewed version

Ethical Adversaries: Towards Mitigating Unfairness with Adversarial Machine Learning

No Author Given

Abstract. Machine learning is being integrated into a growing number of critical systems with far-reaching impacts on society. Unexpected behaviour and unfair decision processes are coming under increasing scrutiny due to this widespread use and also due to theoretical considerations. Individuals, as well as organisations, notice, test, and criticize unfair results to hold model designers and deployers accountable. This requires transparency and the possibility to describe, measure and, ideally, prove the 'fairness' of a system. While this involves concepts such as fairness, transparency and accountability that have been contested for a long time, progress has been made on the way towards (partial) formalisations and proofs that will hopefully make machine learning more amenable to criticism and improvement proposals towards the fulfilment of societal goals. We concentrate on fairness, taking into account that both the transparency of the neural networks and accountability of actors and systems will require further methods.

We offer a new framework that assists in mitigating unfair representations in the dataset used for training. Our framework relies on adversaries to improve fairness. First, it evaluates a model for unfairness w.r.t. protected attributes and ensures that an adversary cannot guess such attributes for a given outcome, by optimizing the model's parameters for fairness while limiting utility losses. Second, the framework leverages evasion attacks from adversarial machine learning to perform adversarial retraining with new examples unseen by the model. These two steps are iteratively applied until a significant improvement in fairness is obtained. We evaluated our framework on well-studied datasets in the fairness literature—including COMPAS—where it can surpass other approaches concerning demographic parity, equality of opportunity and also the model's utility. We also illustrate our findings on the subtle difficulties when mitigating unfairness and highlight how our framework can help model designers.

Keywords: Adversarial machine learning, fairness, neural networks

1 Introduction

Machine learning eases the deployment of systems that tackles various tasks: spam filtering, image recognition, gesture recognition, etc. One of the most trendy applications is decision support. After collecting data on people and their context, these systems give recommendations on who should get a loan, predict

who may commit subsequent offences, etc. However, this support can have detrimental consequences. Well-studied examples include the COMPAS system that predicts the recidivism of pre-trial inmates [2, 9] or accepting credit applications, or more recently the issues with Apple's credit card that resulted in vastly lower spending limits for women. Such systems may amplify the prevalent situation by imposing more expensive loans to African-American people, who then fail to repay them more often [17, 27]. These "positive" feedback loops should be detected and mitigated.

Training a machine learning model can be costly, is sensitive to the data quality, and may result in a complex model. Hence, the decision process may fail to be transparent, which ushers in discrimination or unfair treatment for protected groups. But how to perform this assessment when decisions are often neither interpretable nor intuitive? Researchers have focused on providing quantitative assessments (*e.g.* demographic parity [14], equalized odds [22], statistical parity [18, 36], disparate impact [9, 18], Darlington criterion [11], threshold testing [31]) all covering a specific fairness aspect.

To tackle this problem, researchers assume that a protected attribute A (*e.g.*, race or gender) exists while it should not be predictable, despite existing dependencies between this attribute and others (ZIP code, ...). Thus, only removing the protected attribute, sometimes referred to as *fairness through unawareness*, is known to be insufficient [7, 30].

Because of the far-reaching consequences—being encoded in legal obligations—of these machine learning systems, state-of-the-art methods employ more advanced approaches to mitigate unfairness issues with these models inspired by other machine learning domains, like domain adaptation. One technique is not enough to harness this complex problem. Here, we propose to use two kinds of adversarial machine learning techniques, which we motivate through the following scenarios.

1.1 Motivating Scenarios

Our goal is to develop a notion of *ethical adversaries* based on adversarial machine learning techniques, initially designed to fool machine learning classifiers, to improve their fairness. Our scenarios take place in the context of an ethics assessment activity while designing a new machine learning-based system for a fictional company called "Fancy-Fair AI".

The Feeder: Black-box External Attacks. Alice is a specialist in adversarial machine learning (advML) attacks. She is hired for a mission at Fancy-Fair AI to assess and improve the dataset to increase the performance of the trained machine learning model, later on. However, she is only given the dataset and not the inner details of the already trained black-box machine learning model. Therefore, she has to train a surrogate classifier on the dataset and will feed the machine learning model with new instances. This process, known as *evasion attacks* [4], starts from an existing instance and create a new one by modifying feature values along the gradient in yet unseen zones of the feature space (*i.e.*, where the prediction confidence is low) through successive displacements. Alice

ensures the validity of modified features and tunes the attacks' parameters to improve the system via retraining. Alice crafts examples that are also black-box for fairness evaluation as they are not tuned to optimise a particular metric. Yet, such examples can help the Reader to alleviate misrepresentations as they will provide new feature value combinations.

The Reader: White-box Adversarial Fairness. Bob is an ethics assessment officer at Fancy-Fair AI. He leans on a set of carefully selected fairness metrics assuming that fair decisions should not depend on some protected attributes (race, gender, etc.). Therefore Bob wants to ensure that an insider adversary, striving to predict the value of a protected attribute (e.g., gender) given an outcome (*e.g.*, credit limit), fails to do so. This approach, called *adversarial fairness*, has been applied for autoencoders [15, 25, 26] and for both classification and regression networks [1, 32, 37]. It relies on a *gradient reversal* [19] to update the weights of the adversary so that the chance of predicting the protected attribute is no better than random. This relies on diminishing the dependencies between the protected attribute and the other attributes by backpropagating the gradients with gradient ascent.

Feed, Read and Fix: Grey-box Fairness. This last scenario, which illustrates the main contribution of this paper, reunites Alice and Bob approaches as depicted in Figure 1. This integrated architecture thus works both at the data level (by providing new instances) and at the model level (by preventing it from guessing protected attributes). Fancy-Fair AI monitors the effectiveness of the integrated solution by monitoring demographic parity and equal opportunity and the impact on the utility (accuracy, F_1) of the decision system.

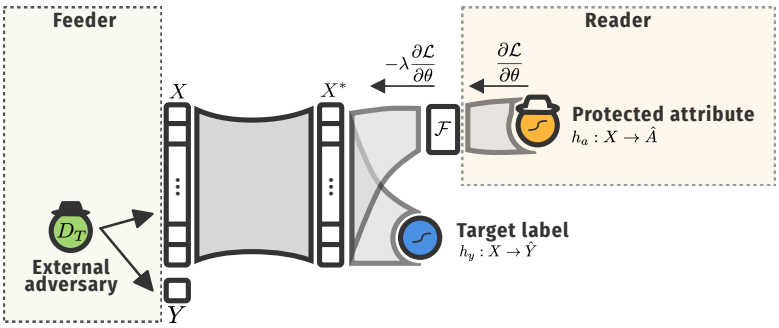


Fig. 1: Our ethical adversaries framework: the feeder (Alice) improves the target Y by creating new instances X, while the reader (Bob) prevents from guessing A. The modified prediction due to model adaptation is shown bottom-right.

1.2 Contributions and Organisation

In this paper, we propose a new framework implementing the grey-box fairness scenario. We make the following contributions:

- The definition of the behaviour of our ethical adversaries in terms of evasion attacks and gradient reversal analyses;
- Demonstration of an undesired side-effect of gradient-based fairness models;
- An implementation of our framework in Python[1];
- Evaluation on three datasets: COMPAS, German Credit and Adult; showing state-of-the-art level results in demographic parity and equal opportunity metrics while globally improving the model's utility.

This paper is organised as follows: Section 2 discusses related work on adversarial machine learning techniques as well as on measuring and mitigating unfairness. Section 3 investigates a problem of gradient reversal methods. Section 4 presents our new framework, followed by its evaluation on the COMPAS, German Credit and Adult datasets in Section 5. Section 6 concludes and gives an outlook on future work.

2 Background and related work

2.1 Adversarial machine learning

Adversarial machine learning aims at finding or creating examples that are problematic for a machine learning model, *e.g.*, [4, 28, 29]. Biggio and Roli [5] synthesised a decade of research in adversarial machine learning. These techniques follow the same process: probe an existing target machine learning based system to gain information about it, copy an existing example, apply an adversarial technique that will modify the example depending on the desired goal. Modified examples show an interesting behavior: they remain similar to the original ones while being misclassified by the trained model. Various models can be attacked including support vector machines (SVMs), linear models or even (deep) neural networks (NNs). While adversarial machine learning begins with a model of attackers' possibilities, it also enables the design of defenses against attacks. In particular, adversarial retraining has been very popular with the emergence of generative adversarial nets (GANs) proposed by Goodfellow et al. [20]. Other approaches to robustify neural networks exist and use existing examples: *e.g.*, Edwards and Storkey [15] used an adversary to force an encoder-decoder network to learn domain-independent representations [15, 26, 37].

2.2 Fair Representations with Neural Networks

Several works [1, 19, 26, 32] aim at training models to obtain internal representations that are fair. The embeddings produced by these models cannot be used to predict the protected attribute A. Such works integrate an adversary with a new goal: trying to predict the protected attribute A (and not degrading the model's performance anymore).

[1] Available at `http://<url removed for double blind review>`

A new model is created but with two goals: (i) predicting the main attribute Y (which we will refer to as the utility of the model); (ii) not being able to predict the protected attribute A. They can be formally defined using minimax [15]:

$$\min_{\theta} \max_{\phi} L\left(\theta, \phi\right), \tag{1}$$

with an adversary ϕ and an encoder with parameters θ. We use this representation to predict both Y and A via an adversarial network. Adel et al. [1], Ganin et al. [19], Raff and Sylvester [32] all proposed to optimize a variant of the following loss function:

$$L(\theta, \phi) = E_{\theta,\phi}(X, Y) - \lambda D_{\theta,\phi}(X, A), \tag{2}$$

with $D_{\theta,\phi}$ the loss for predicting A from X, and $E_{\theta,\phi}$ the loss for the target prediction Y also from X and λ a hyper-parameter.

Zemel et al. [36] learn *fair representations* X^* of the original input features X. The idea is to remove existing dependencies between the representation X and the protected attribute A, making its prediction impossible for adversaries. This would make the practice of *red-lining* also impossible, as these dependencies can no longer be correlated with A. We consider this goal as a good proxy for fairness and this approach has been further investigated [1, 26, 32].

2.3 Fairness through a Gradient Reversal Layer (GRL)

Ganin et al. [19] introduced a *gradient reversal layer* (GRL) originally for domain adaptation. Both Raff and Sylvester [32] and Adel et al. [1] treated the protected attribute A as a domain label. The gradient reversal strategy assumes that multiplying by a negative sign will increase the loss of the branch $h_a : X \to \hat{A}$ and yields a representation X^* that is maximally invariant to changes in A [1, 32]. For a model with two target outputs and a hidden internal representation, Equation 2 applies. In our framework, the Reader (see Figure 1) reuses this approach to mitigate the ability to predict A.

2.4 Adversarial attacks on model inputs

Our framework uses a second kind of adversarial machine learning, known as *evasion attacks* [4], to diversify the training set. The goal of the evasion attacks is to generate new examples that do not follow the same distribution as the original set. Generated examples combine different characteristics, initially under-represented. These examples can be added to the training set to perform adversarial retraining.

Evasion attacks are a gradient-based method and use a step size parameter t to converge towards a local optimum. An attack: (i) chooses a starting example for which the classifier's decision is known; (ii) computes the gradient directed towards the separating functions; (iii) applies this direction to the example's position scaled by t; (iv) repeats until a stopping criterion is met (number of

iterations or a plateau is reached). This algorithm has been implemented and made publicly available in the Python secML package [2] that we will use in our experiments.

Demontis et al. [12] showed the transferability potential of such attacks. In particular, from the attacker's point of view, building the exact, same model is not necessary. Data distributions to train both models should be similar. Hence, one can approximate any complex or non-derivable ML models with simpler ones and still generate relevant examples to influence the original model while retraining. The Feeder of our framework aims at providing adversarial examples for retraining that will mitigate unfairness.

2.5 Discrimination-Aware Data Mining

In works on *discrimination-aware data mining* (DADM) and fairness in machine learning, modifications to the data, the learning algorithms, or the resulting patterns and models [21] have been developed and applied. Pedreschi et al. [30] introduced an approach to tackle discrimination by extracting classification rules and ranking them based on a measure. DADM focuses on *discovering discrimination* as well as *preventing discrimination*, both direct or indirect discrimination, the latter is the reason why simply removing protected attributes is not effective (see Section 1). Our framework performs both steps in an integrated manner by generating new examples and tuning the model to prevent discrimination.

3 Why gradient reversal is not a silver bullet

As described in Section 2.3, GRL is a currently popular approach, also known as 'adversarial fairness'. We also use this technique, and like the authors who used it to learn 'fair(er) representations' [1, 32], we find that it can mitigate unfairness in classification/prediction tasks.

In the remainder of this section, we formulate and prove this problem and illustrate it in Figure 2.

The introduction of a gradient reversal layer by Ganin et al. [19] targeted domain adaptation. Adel et al. [1], Raff and Sylvester [32] continued on this by viewing the protected attribute A as a domain label. However, Ganin et al. [19] offered no guarantees as to how the domain was represented internally. In this section, we argue that the adversarial branch achieves its goal by learning specifically to predict the protected attribute, rather than obfuscating it.

The gradient reversal strategy assumes that multiplying by a negative sign will increase the loss of the branch $h_a : X \to \hat{A}$ that then yields a representation X^* that is maximally invariant to changes in A [1]. This is intuitive, but there is no guarantee that gradient descent with flipped gradients does guarantee this maximal invariance.

[2] https://secml.gitlab.io/

Lemma 1. *Gradient reversal equates to perform gradient ascent on the shared layers with respect to the protected attribute A, whilst simultaneously performing gradient descent on the dedicated branch for the attribute A.*

Proof. Consider the final layer L with two independent branches, governed by parameters θ and ϕ respectively, and the shared penultimate layer $L-1$. The shared loss function for both is stated earlier in Equation 2. For the branch that predicts the protected attribute A, the loss $D'_{\theta,\phi}(X,A)$ gives rise to the weight updates $\Delta w_{\phi}^{(L)}$ for the weights of the branch $h_a : X \to \hat{A}$, following

$$\Delta w_{\phi}^{(L)} = -\alpha\lambda \frac{\partial D_{\theta,\phi}}{\partial w_{\phi}^{(L)}}. \tag{3}$$

The branch giving target label Y is updated similarly. The shared penultimate layer's weights rely on the shared loss $L(\theta,\phi)$ and are updated following

$$\Delta w^{(L-1)} = -\alpha \frac{\partial E_{\theta,\phi}}{\partial w^{(L-1)}} + \alpha\lambda \frac{\partial D_{\theta,\phi}}{\partial w^{(L-1)}}. \tag{4}$$

The weights $w_{\phi}^{(L)}$ are updated following gradient descent with respect to the loss $D_{\theta,\phi}(X,A)$, thus minimizing the loss for this branch. However, gradient reversal simultaneously performs gradient ascent with respect to the same loss $D_{\theta,\phi}(X,A)$ on all weights of layers $1,\ldots,L-1$. The shared layers still perform gradient descent with regard to the loss $E_{\theta,\phi}(X,Y)$ for the target label Y.

□

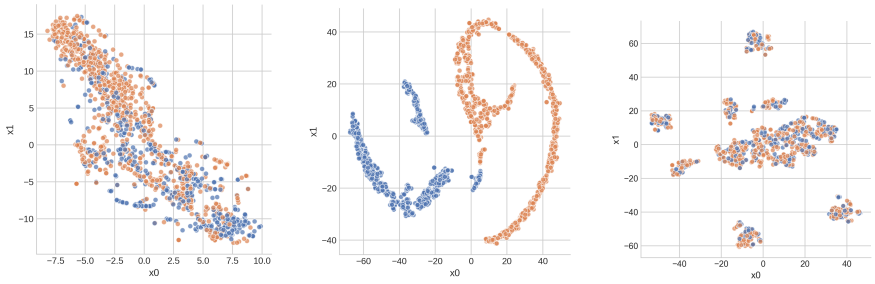
We have shown that the shared penultimate layer does not perform gradient descent, but gradient ascent. This is in accordance with the implicit definition for maximal variance [1, 19, 32] following

$$\max_{\phi} D_{\theta,\phi}(X,A). \tag{5}$$

This fits in the larger minimax problem from Equation 1 and results in a saddle point [19]. However, the end result is not guaranteed to be a maximal invariant representation. In the worst case, maximizing this loss $D_{\theta,\phi}(X,A)$ can even result in the opposite optimum for the shared trunk with regard to A. This means that the model is not necessarily maximally invariant on $(L-1)$. We need to emphasize that this is a *theoretical* result, but it calls for caution when adversarial fairness is to be used to, for example, publish or re-use a supposedly 'fair' data representation. We illustrate this issue on the COMPAS dataset in Figure 2.

For each individual for the COMPAS test set, all three models derive a representation in the last hidden layer, on which we applied a t-SNE dimensionality reduction for a two-dimensional visualisation.

The model without fairness constraints (Figure 2a) has slight separation with regard to the protected attribute, but it is clearly separable in the representation from the model trained with a GRL (Figure 2b). This is also shown by retraining



(a) Naive model (b) Model trained with a GRL ($\lambda = 50$) (c) Model trained with our framework

Fig. 2: T-SNE dimensionality reduction of the activations in the last hidden layer on the held-out COMPAS test set. Distinct colors are used for the reported race of individuals in the dataset: either African-American ● or Caucasian ● .

a one-layer perceptron on these representation. The model that was originally trained to predict only recidivism could be used to classify the protected attribute race with $AUC = 0.71$. And although the original GRL reported an $AUC = 0.44$, Theorem 1 tells us that this adversary cannot be trusted. Which is the case here, as an independent perceptron has $AUC = 0.92$. Elazar and Goldberg [16] made an empirical observation on *leakage* of protected attributes specifically for text-based classifiers that can also be traced back to this.

Here, we demonstrated that the hidden representation obtained by gradient reversal, not only still contains information about the protected attribute, but contains a stronger signal. Our architecture that joins 'adversarial fairness' and 'adversarial learning' (see Section 1.1 and Fig. 1) leverages utility- *and* fairness-focused methods in a better way than the modification of the model alone. By injecting noise with the adversarial Feeder, our framework makes the protected attribute A a useless predictor, as shown in Figure 2c. Our results, discussed in Section 5.3, confirm this expectation.

4 Ethical Adversaries Framework

In this section, we present how the two adversarial attacks interact in our framework. The first attacks the inputs of the model whereas the second tries to predict the protected attribute A as part of the model. We join both adversaries in a single system to address issues discussed in Section 2.2 and Section 2.3, ultimately resulting in a fairer model.

Figure 1 shows how these two adversaries are incorporated. Our network follows the architecture with a GRL (discussed in Section 2.3 and used by the Reader on the right part of the figure). The external adversary (the Feeder on the left part of the figure) performs evasion attacks as discussed in Section 2.4. We

discuss both parts in this section, including the hyperparameters and complexity they introduce to our architecture.

4.1 Adversarial reader

We augment the original model by adding a second branch with reversed gradients that will predict the protected attribute A. We follow the training setup from Raff and Sylvester [32], discussed in Section 2.3. The model will thus be trained with the joint loss of the original prediction target and the protected attribute. During the backward pass, the signs of the gradients from the adversarial branch are flipped and scaled by a hyperparameter λ.

4.2 Adversarial feeder

As presented in Section 2.4, the feeder needs a starting point that is an approximation of the target model, i.e., a surrogate model (see Section 2).

The evasion attack runs as presented in Section 2.4, and newly generated examples can be included in the training set for adversarial retraining. Note that adversarial retraining may drastically increase convergence time to compute a separating function since included adversarial examples make the separation more difficult to find. Generally, defining the ideal size of batches for training remains an open issue [24].

4.3 Complexity analysis

Our architecture consists of three elements: the model under attack, the Reader and the Feeder. The adversarial reader is trained in conjunction with the model under attack. The time complexity of the attacked model is in part dependent on the chosen model. For neural networks, this is architecture-dependent. After training the model with the adversarial reader, a surrogate is trained, in our case, an SVM with time complexity $\mathcal{O}(n^3)$ for $n \gg d$ with n the number of data points and d the number of features [8]. The time complexity for our entire system becomes $\mathcal{O}(n^3)$ and scales linearly with the number of adversarial attacks. While not the focus of this paper, there are ways to learn SVMs faster and integrating them is subject to future work.

5 Evaluation

We evaluate our model on three popular datasets: COMPAS [2], German Credit and, the Adult Census [23]. The COMPAS dataset was originally a sample of outcomes from the COMPAS system that predicted the risk of recidivism. This caused a debate about whether or not this score was disadvantaging African Americans [2, 9, 10, 13]. The dataset, therefore, includes the race of individuals.

In line with other research [1, 2, 35], we will only use individuals from *Caucasian* or *African-American* descent. As there is much less data on other groups

(e.g., only 31 instances for people of Asian descent), this poses issues during training and evaluation. This implies that there are minorities that are excluded from many studies; more datasets would be needed to study whether patterns of unfairness are similar and mitigation measures can be transferred, or whether these affect different demographics differently.

COMPAS is composed of 5,278 instances and represented by 12 features. The target variable is whether a person has recidivated within two years. The race is used as a protected attribute. The Adult dataset gathers 32,000 instances represented by 9 features. We use gender as a protected attribute and the binary target variable is income, whether someone earns more than 50,000 USD. German Credit is the smallest dataset, with only 1,000 instances and 20 features. There is a class imbalance, with 70% of all samples good credits and only 30% bad credits. The protected attribute is age, with a threshold at 25 years.

For reproducibility purposes, we have publicly released our code and provided users with a template that they can incorporate in their projects. It is compatible with all PyTorch models with only minor modifications, i.e., adding an adversarial branch and replacing the training loop. We recall that we have used the secML package[3] (v0.11) for running evasion attacks.

5.1 Training setup

The model under attack. We start from a neural network of 3 hidden layers with 32 hidden units for COMPAS and German Credit and 128 for Adult, due to its larger encoded input. Each of the hidden units has a ReLU activation. This activation function is computationally efficient and mitigates the issue of vanishing gradients since the function never saturates, which makes it one of the most popular activation functions. For the output units, a softmax activation was used to get the classification and a linear activation for COMPAS. The network—as well as the adversarial reader—are trained with the Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.9999$ and an initial learning rate $l_r = 0.01$, which is adjusted by a factor of 0.1 when reaching a plateau.

The adversarial reader. The adversarial reader is part of the model under attack and therefore follows the same training regime. The joint loss follows Equation 2 by including the GRL. The individual losses for both h_A and h_y are binary cross-entropy loss, except for COMPAS. In that case, the risk score is predicted as a regression problem with the MSE loss and then thresholded at > 4 (low *vs* medium and high risk).

The adversarial feeder. In our setting, we can use the same training set for both the feeder and reader since they are part of the same, unique architecture. We also approximate—relying on the earlier discussed transferability of attacks—the attacked model by an SVM with a radial basis function kernel. We set the hyperparameters C and γ with a grid search with a reduced number of values: respectively $\{0.0001; 0.001; 0.01; 0.1; 1.0\}$ and $\{0.01; 0.1; 1; 10; 100; 1000\}$. We performed 10-fold cross-validation.

[3] https://secml.gitlab.io/

5.2 Evaluating fairness

Since the architecture we proposed in Section 4 aims at mitigating unfairness, we will have to evaluate this aspect in our experiments. There exist several measures of fairness in the literature. In this subsection, we discuss some of the most popular ones for different aspects of fairness.

We define all measures via the predicted values of the classifier $\hat{Y}$ and the protected attribute A. We identify the disadvantaged group with $A = 1$ and the privileged group with $A = 0$. The similarities of predictions are described for $\hat{Y} = 1$. Since the focus of most fairness measures is on the disadvantaged group having fewer (desired) opportunities, $\hat{Y} = 1$ is generally the desired outcome.

One set of measures expresses the requirement that the predicted values of the classifier $\hat{Y}$ conditioned on the protected attribute be equal [6] or the difference to be within an acceptable range.

Definition 1. *Demographic parity (DP). DP is the equality or similarity of prediction outcomes as an absolute difference [14, 32]:*

$$DP = \left| P(\hat{Y} = 1 \mid A = 0) - P(\hat{Y} = 1 \mid A = 1) \right| \leq \epsilon. \tag{6}$$

Definition 2. *Demographic parity ratio (DPR). DPR is the equality or similarity of prediction outcomes as a ratio:*

$$\frac{P(\hat{Y} = 1 \mid A = 1)}{P(\hat{Y} = 1 \mid A = 0)} \geq \tau. \tag{7}$$

Requiring $\epsilon = 0$ or $DPR = 1$ would require exactly equal predicted outcomes for both groups. This is unrealistic for most data, such that real-world usage of such measures is less restrictive. For instance, in a legal setting, the US Equal Employment Opportunity Commission (EEOC) uses the DP ratio with $\tau = 0.8$ ("*80% rule*" [18]), stating that disparate impact caused by employment-related decisions or structures can only be ascertained if $DPR \leq 0.8$.

Demographic parity has received some criticisms, since (i) it can meaninglessly reduce the utility of the classifier and—more worrying— (ii) does not necessarily measure what many would define as fairness [14]. The first issue is due to possible correlations between the protected attribute A and the true outcome Y. Since we expect equality of the classifier concerning the protected attribute, it cannot operate as a perfect classifier.

The second issue stems from ignoring both the true outcome and individual merits. For instance, consider a selection procedure with two subgroups with different values for the protected attribute A. One subgroup can be composed of qualified individuals (*i.e.*, with high chances for a positive true outcome $Y = 1$), but another subgroup can consist of random individuals. This still satisfies demographic parity, but these *token* individuals are not guaranteeing fairness since qualified individuals from the protected subgroup are still mistreated.

Addressing the criticisms of demographic parity, Hardt et al. [22] presented two other metrics that extend the aforementioned ones. By including the true

outcome Y, the authors show that this variable can serve as a *justification* for the predicted outcome. For example, in the case of COMPAS, this is the recidivism rate as measured by violent crimes in a two-year window. Conditioning by the true outcome is a justification that the authors consider to be a suitable interpretation of the *task-specific similarity measure* from Dwork et al. [14], which can otherwise be difficult to come up with. This is also very similar to *disparate mistreatment* [3, 34] used as an evaluation metric by Adel et al. [1].

Definition 3. *Equal opportunity (EO). EO requires an independence* $\hat{Y} \perp\!\!\!\perp A \mid Y$ *of* $\hat{Y}$ *and* A *conditioned on the true outcome* Y*. Expressed as a difference, this yields:*

$$\left| P(\hat{Y} = 1 \mid A = 0, Y = 1) - P(\hat{Y} = 1 \mid A = 1, Y = 1) \right| \leq \nu. \tag{8}$$

"Equality of opportunity" is satisfied if $\nu = 0$, and larger values are indicative of unfairness in the model or data.

5.3 Results

Table 1: Results on the three datasets. An obelisk (†) show results reported by original papers. Results of classifiers without fairness constraints are reported as a baseline. Best results are in bold typeface. An asterisk (∗) indicates a division by zero.

Model	ACC	F1	DP	DPR	EO
Adult					
Baseline without fairness constraints	**0.839** ± 0.009	**0.763**	0.173	0.296	0.096
GRL	0.612 ± 0.012	0.518	0.059	1.931	**0.061**
NBF (NB) [6]	0.773†	—	**0.000**†	—	—
NBF (EM) [6]	0.801†	—	0.001†	—	—
Grad-Pred [32]	0.754†	—	**0.000**†	—	—
FF [33]	0.753†	—	**0.000**†	—	—
LFR [36]	0.702†	—	0.001†	—	—
Ours	0.814 ± 0.009	0.689	0.031	**0.784**	0.179
German Credit					
Baseline without fairness constraints	0.705 ± 0.063	0.624	0.018	0.929	0.198
GRL	0.710 ± 0.063	0.415	**0.000**	∗	**0.000**
Grad-Pred [32]	0.675†	—	0.001†	—	—
FF [33]	0.700†	—	**0.000**†	—	—
LFR [36]	0.591†	—	0.004†	—	—
Ours	**0.730** ± 0.062	**0.640**	0.006	**0.971**	0.175
COMPAS					
Baseline without fairness constraints	0.715	0.709	0.466	2.192	0.449
GRL	0.567	0.549	0.057	**0.926**	0.114
COMPAS	0.655 ± 0.029	0.654	0.289	1.829	**0.000**
Preference-based fairness [35]	0.675†	—	0.380†	—	—
Ours	**0.794**	**0.793**	**0.026**	0.840	0.008

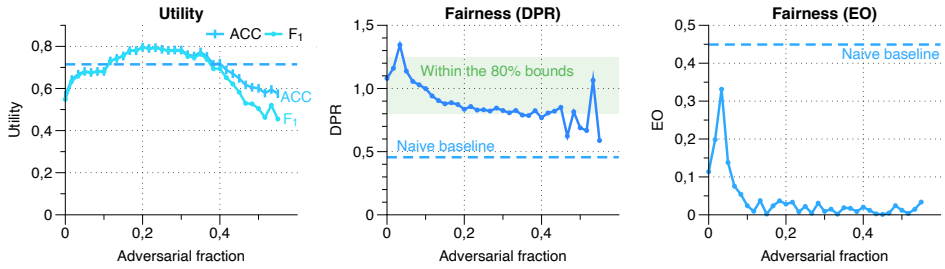


Fig. 3: Fairness and utility measures after each attack iteration on COMPAS (Batch size of 1024, $\lambda = 100$, epochs=100, 50 adversarial points per iteration)

Table 1 presents our results on the three datasets. We compare them with (i) a naive baseline, *i.e.*, the same architecture without any particular control on fairness aspects, (ii) a re-implementation of the GRL [1, 19, 32] and (iii) the reported results from other works that incorporate fairness and cover a wide range of learning algorithms: Naive Bayes [6], random forests [33], SVMs [35] and neural networks [32, 36]. The models' utility was evaluated by binary classification accuracy and macro-averaged F_1 score; the latter highlights some issues when dealing with class imbalances, as is the case for German Credit. Fairness is evaluated with demographic parity, both as an absolute difference (DP) and as a ratio (DPR), and equal opportunity (EO).

Adel et al. [1] also report results on both COMPAS and Adult but use a different setup for the Adult dataset. For COMPAS, the reported results (as well as their unfair baseline) are significantly higher than in our experiments, which we could replicate only when classifying high-risk individuals. To make a meaningful comparison, we also include our replication of *FAD* [1] as *GRL*.

The utility of our framework is the highest on the German Credit and COMPAS datasets, even surpassing the baseline model. On Adult, we achieve the highest utility of any model with fairness constraints. These results show that our model has only a very limited impact on the utility of the classifier, and it can even contribute to the training as is also visualised in Figure 3. Note that on German Credit, a majority classifier would achieve 70% accuracy already, hence the inclusion of the F_1 score.

Regarding fairness evaluation, our framework gives the best results for COMPAS when considering DP. It also increases fairness as measured by DPR, which is the only one of the considered measures that indicates the "direction" of unfairness. More fairness is sometimes given by an *increase* towards parity (DPR=1) for the disadvantaged group: for the German Credit dataset, their chances of getting a loan increase. In COMPAS, the "bias against blacks" [2] *decreases* from a probability of recidivism prediction that is more than twice as high as for white people. Here, the near-equality of 0.926 appears fairer than the "opposite unfairness" of our, further reduced, DPR value.

Figure 3 also highlights the effect of the adversarial fraction in the training dataset on COMPAS. When adversarial examples (equivalent to 25% of the training set size) are added to the training set, the utility is maximal. With higher fractions, the utility decreases and the development of the DP ratio fluctuates. This could stem from the minimax formulation, where a small fraction (i.e., 25%) helps optimize better for this saddle point, but higher fractions only add noise.

6 Summary, conclusions and future work

In this paper, we presented a novel architecture for integrating fairness constraints in machine learning models. Our architecture consists of two adversaries: (i) an adversarial reader that evaluates fairness constraints during model training and attempts to enforce them, and (ii) an adversarial feeder that performs iterative evasion attacks to discover previously uncovered regions in the input space. We evaluated our architecture on three well-studied datasets and showed that it can deliver high utility to models while satisfying fairness constraints. On COMPAS, we illustrated that our architecture yields a model that surpasses an unfair baseline regarding the utility (accuracy and F_1 score), whilst giving better fairness guarantees. We provide evidence that gradient reversal alone is not sufficient (it might even be detrimental) but that our combination of adversaries leads to intrinsically fairer models.

There is room for future work. First, we may optimize the runtime execution of the technique via faster learning of surrogate models. Second, we could use the target model directly instead of a surrogate classifier to support adversarial attacks and assess if transferability properties hold for fairness constraints. This requires heavyweight modification of the secML framework to allow multiple output values in neural networks. Third, while we do not generate invalid instances, one could define constraints involving multiple features: *e.g.*, a 4-year-old child cannot have a Ph.D. Enforcing these *domain-specific* constraints during attack generation raises questions on the representation of the feature space and optimal convergence of the algorithms. Finally, we would like to generate the most dissimilar examples possible to ensure good coverage of the unseen feature space with a minimal number of attacks.

References

1. Adel, T., Valera, I., Ghahramani, Z., Weller, A.: One-Network Adversarial Fairness. In: AAAI Conference on Artificial Intelligence (2019)
2. Angwin, J., Larson, J.: Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. https://www.propublica.org/article/ machine-bias-risk-assessments-in-criminal-sentencing (2016)
3. Barocas, S., Hardt, M., Narayanan, A.: Fairness and Machine Learning. fairmlbook.org (2019)

4. Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., Roli, F.: Evasion attacks against machine learning at test time. In: ECML/PKDD. pp. 387–402 (2013)
5. Biggio, B., Roli, F.: Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition Journal 84, 317–331 (2018)
6. Calders, T., Verwer, S.: Three naive Bayes approaches for discrimination-free classification. Data Min. Knowl. Discov. 21(2), 277–292 (2010)
7. Calders, T., Žliobaitė, I.: Why unbiased computational processes can lead to discriminative decision procedures. In: Discrimination and privacy in the information society, pp. 43–57. Springer (2013)
8. Chapelle, O.: Training a support vector machine in the primal. Neural computation 19(5), 1155–1178 (2007)
9. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. arXiv:1703.00056 (2017)
10. Corbett-Davies, S., Goel, S.: The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. arXiv:1808.00023 (2018)
11. Darlington, R.B.: Another Look at "Cultural Fairness". Journal of Educational Measurement 8(2), 71–82 (1971)
12. Demontis, A., Melis, M., Pintor, M., Jagielski, M., Biggio, B., Oprea, A., Nita-Rotaru, C., Roli, F.: Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In: 28th USENIX Security Symposium. pp. 321–338 (2019)
13. Dieterich, W., Mendoza, C., Brennan, T.: COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity (2016)
14. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness Through Awareness. In: 3rd Innovations in Theoretical Computer Science Conference. pp. 214–226. ACM (2012)
15. Edwards, H., Storkey, A.: Censoring Representations with an Adversary. arXiv:1511.05897 (2015)
16. Elazar, Y., Goldberg, Y.: Adversarial removal of demographic attributes from text data. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 11–21. ACL, Brussels, Belgium (Oct-Nov 2018)
17. Ensign, D., Friedler, S.A., Neville, S., Scheidegger, C., Venkatasubramanian, S.: Runaway Feedback Loops in Predictive Policing. arXiv:1706.09847 (2017)
18. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: 21th ACM SIGKDD International Conference. pp. 259–268 (2015)
19. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The Journal of Machine Learning Research 17(1), 2096–2030 (2016)
20. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. In: NIPS, pp. 2672–2680. Curran Associates (2014)
21. Hajian, S., Domingo-Ferrer, J.: A methodology for direct and indirect discrimination prevention in data mining. IEEE Trans. Knowl. Data Eng. 25(7), 1445–1459 (2013), https://doi.org/10.1109/TKDE.2012.72

22. Hardt, M., Price, E., ecprice, Srebro, N.: Equality of Opportunity in Supervised Learning. In: NIPS, pp. 3315–3323. Curran Associates (2016)
23. Kohavi, R.: Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In: Kdd. vol. 96, pp. 202–207 (1996)
24. Li, M., Zhang, T., Chen, Y., Smola, A.J.: Efficient mini-batch training for stochastic optimization. In: 20th ACM SIGKDD. pp. 661–670 (2014)
25. Louizos, C., Swersky, K., Li, Y., Welling, M., Zemel, R.: The Variational Fair Autoencoder. arXiv:1511.00830 (2015)
26. Madras, D., Creager, E., Pitassi, T., Zemel, R.S.: Learning Adversarially Fair and Transferable Representations. arXiv abs/1802.06309 (2018)
27. Overdorf, R., Kulynych, B., Balsa, E., Troncoso, C., Gürses, S.: Questioning the assumptions behind fairness solutions. arXiv:1811.11293 (2018)
28. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: IEEE European Symposium on Security and Privacy. pp. 372–387 (2016)
29. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. In: Asian Conference on Computer and Communications Security. p. 506–519. ACM (2017)
30. Pedreschi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: 14th ACM SIGKDD. pp. 560–568 (2008)
31. Pierson, E., Corbett-Davies, S., Goel, S.: Fast threshold tests for detecting discrimination. In: Storkey, A., Perez-Cruz, F. (eds.) 21st International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 84, pp. 96–105. PMLR, Playa Blanca, Lanzarote, Canary Islands (2018)
32. Raff, E., Sylvester, J.: Gradient Reversal against Discrimination: A Fair Neural Network Learning Approach. In: IEEE 5th International Conference on Data Science and Advanced Analytics. pp. 189–198 (2018)
33. Raff, E., Sylvester, J., Mills, S.: Fair forests: Regularized tree induction to minimize model bias. In: Conference on AI, Ethics, and Society. pp. 243–250 (2018)
34. Zafar, M.B., Valera, I., Rodriguez, M.G., Gummadi, K.P.: Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. 26th International Conference on World Wide Web pp. 1171–1180 (2017)
35. Zafar, M.B., Valera, I., Rodriguez, M.G., Gummadi, K.P., Weller, A.: From Parity to Preference-based Notions of Fairness in Classification. arXiv:1707.00010 (2017)
36. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International Conference on Machine Learning. pp. 325–333 (2013)
37. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating Unwanted Biases with Adversarial Learning. In: Proceedings of Conference on AI, Ethics, and Society. pp. 335–340. ACM Press (2018)