Ethical Adversaries

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

Published in:
1st workshop on Bias and Fairness in AI, co-located with ECMLPKDD 2020

Publication date:
2020

Document Version
Peer reviewed version

Link to publication

Citation for pulished version (HARVARD):
Delobelle, P, Temple, P, Perrouin, G, Frénay, B, Heymans, P & Berendt, B 2020, Ethical Adversaries: Towards Mitigating Unfairness with Adversarial Machine Learning. in *1st workshop on Bias and Fairness in AI, co-located with ECMLPKDD 2020.* Bias and Fairness in AI (BIAS 2020), 18/09/20.

Ethical Adversaries: Towards Mitigating Unfairness with Adversarial Machine Learning

Delobelle, Pieter; Temple, Paul; Perrouin, Gilles; FRENAY, BENOIT; Heymans, Patrick; Berendt, Bettina

Published in:
1st workshop on Bias and Fairness in AI, co-located with ECMLPKDD 2020

Publication date:
2020

Document Version
Peer reviewed version

Ethical Adversaries: Towards Mitigating Unfairness with Adversarial Machine Learning

Pieter Delobelle[1], Paul Temple[2], Gilles Perrouin[2], Benoît Frénay[2], Patrick Heymans[2], and Bettina Berendt[1,3]

[1] Department of Computer Science, KU Leuven and Leuven.ai,
firstname.lastname@kuleuven.be
[2] PReCISE, NaDi, Université de Namur,
firstname.lastname@unamur.be
[3] Faculty of Electrical Engineering and Computer Science, TU Berlin

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 its theoretical considerations. Individuals, as well as organisations, notice, test, and criticize unfair results to hold model designers and deployers accountable. We offer a framework that assists these groups in mitigating unfair representations stemming from the training datasets. Our framework relies on two inter-operating adversaries to improve fairness. First, a model is trained with the goal of preventing the guessing of protected attributes' values while limiting utility losses. This first step optimizes the model's parameters for fairness. Second, the framework leverages evasion attacks from adversarial machine learning to generate new examples that will be misclassified. These new examples are then used to retrain and improve the model in the first step. 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 assist 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, etc. One of the most trendy applications is decision support. These systems give recommendations on who should get a loan, predict who could commit subsequent offences, etc. based on data describing the individuals affected. Such systems have a desirable property: they provide objective, supposedly consistent decisions based on a collection of data. At first glance, this could counteract unfair decisions made by humans.

However, these support systems still exhibit unfair behaviour. Such behaviours can possibly impact certain individuals, belonging to social or even protected groups. Well-studied examples include the COMPAS system that predicts the recidivism of pre-trial inmates [2, 10] and keep taking decisions in favor of Caucasian people compared with African-Americans. We consider fairness where the impact on individuals can be categorized as either *allocational harm* or *representational harm* [8]. With allocational harm, the favorable outcome (e.g. bail being granted) differs between social groups. Representational harm is more subtle, and include *differences in performance* between social groups, and *stereotyping*. We focus on allocational harm in this work, as decision support systems with different outcomes can affect social groups far beyond the outcome itself.

If an allocational harm exists when advising for a favorable outcome, the decision could, in turn, affect the social group(s) who did not receive that outcome and, ultimately, risking the creation of a feedback loop where unfair behaviour is amplified [17, 26]. For example, consider a system that imposes more expensive loans to African-American people, who then fail to repay them, that will lead them to ask for another loan, etc.

Because of these consequences, researchers increasingly focus on incorporating fairness objectives in their systems. In *discrimination-aware data mining* (DADM), modifications were developed and applied to data, learning algorithms, or resulting patterns and models [21]. More recently, adversarial fairness is continuing in this field with, for instance, research on learning representations [25, 36] and task-specific fair models [1, 30, 32].

Adversaries are also used when assessing the security of machine learning based systems. Biggio and Roli [7] synthesised a decade of research in adversarial machine learning. This domain of research aims at finding or creating examples that are problematic for a machine learning model, *e.g.* Biggio et al. [5], Papernot et al. [27, 28]. These examples can be injected directly into the training phase in order to perturb the training of the model, known as *poisoning attacks*, or they can simply be used to bypass the model that is supposed to act as a filter, in this case, they are called *evasion attacks*.

In this paper, we propose a new framework implementing a gray-box fairness scenario coupling evasion attacks and fair machine learning using gradient reversal. We evaluate our framework on three datasets: (i) COMPAS, (ii) German Credit, and (iii) Adult. We show demographic parity and equal opportunity improved when comparing to the state of the art while globally improving the model's utility. Our framework thus reconciles fairness and model performance.

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

2 Background and Related Work

2.1 Poisoning and Evasion Attacks

Adversarial machine learning assesses the required effort to make a classifier unusable by forcing it to perform so many errors that users will not trust its predictions anymore [7]. The generation of adversarial attacks follows this black-box process: (i) probe an existing target model to gain information about it, (ii) copy an existing example, (iii) apply an adversarial technique that will modify the example depending on the desired goal.

Various models can be attacked including support vector machines (SVMs), linear models and even neural networks (NNs) [5, 27, 28]. Since all machine learning models are based on a similar set of assumptions, including the fact that they statistically approximate data distributions, adversarial machine learning leverage on these assumptions to train a surrogate classifier to start the attack on and results are transferred to the target model [12]. Only one restriction remains on the surrogate classifier, attacks are gradient-based techniques requiring the discriminant function to be differentiable. We distinguish between *poisoning attacks* and *evasion attacks*. In the former, malicious examples are introduced in the training set in order to significantly and permanently affect the model to be trained [4, 6].

In concurrent work by Solans et al. [33], poisoning attacks have been used to influence the fairness of machine learning models in a black-box manner. The authors have also linked their poisoning attack to demographic parity, an evaluation metric that will be introduced in Section 2.2.

Kulynych et al. [24] also used poisoning attacks, specifically for countering effects of credit scoring systems. In addition, they provide an outline of how users can affect optimization systems to mitigate negative externalities, called Personal Optimization Technologies (POTs). This framework could also be used to ground the adversarial attacks generated by the *Feeder* from our framework.

In this paper, we consider evasion attacks that are performed on *an already trained model*. We craft adversarial examples that are supposed to belong to a class while the model will assign them with a different one because of specific characteristics, highlighting an unfair behavior regarding a certain population. By carefully reintroducing these examples during retraining, we hypothesize that the retrained model will be fairer. While we rely on a similar example generation technique, we have a distinct exploitation goal.

2.2 Evaluating Fairness

There exist several measures of fairness in the literature, *e.g.* demographic parity [14], equalized odds and equalized opportunity [22], statistical parity [18, 36], disparate impact [10, 18], and threshold testing [29]. In the following, we focus on the most popular and representative measures: (i) demographic parity and (ii) equalized opportunity. 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 [9] 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, 30]:*

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

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

$$DPR = \frac{P(\hat{Y} = 1 \mid A = 1)}{P(\hat{Y} = 1 \mid A = 0)} \geq \tau. \tag{2}$$

Requiring $DP = 0$ or $DPR = 1$ would require exact equality in the outcome predictions 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 the measure does not necessarily report on what many would define as fairness [14]. This issue stems from ignoring both the true outcome and individual merits. For instance, consider a selection procedure where we consider two individuals belonging to the protected group. Let say that one individual is qualified (*i.e.*, with high chances to get a positive true outcome $Y = 1$) and the other one is not. Not selecting the qualified individual could be considered unfair, but the selection would satisfy demographic parity even when selecting the not-qualified individual. So these *token* individuals are not guaranteeing fairness since qualified individuals from the protected group 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:*

$$\left| P(\hat{Y} = 1 \mid A = 0, Y = 1) - P(\hat{Y} = 1 \mid A = 1, Y = 1) \right| \leq \nu. \tag{3}$$

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

2.3 Fair Neural Networks

Fair models have been studied for a variety of learning algorithms, such as Naive Bayes classifiers [9] or SVMs [35]. Nowadays, the focus is also on neural networks due to their prediction performance [1, 25, 31].

Several work have tried to mitigate unfairness in neural networks with white-box adversaries [1, 15, 25, 31, 37]. In all these instances, a new model architecture is proposed with two goals: (i) predicting the main attribute Y (which we will refer to as the *utility of the model*; with $Y = 1$ being the positive outcome); (ii) not being able to predict the protected attribute A (with $A = 1$ considered as belonging to the protected group). The joint goals can be formally defined as a min max optimization problem [15] over the loss function L, i.e., $\min_\theta \max_\phi L(\theta, \phi)$, with an adversary ϕ and an encoder with parameters θ. We use this representation to predict both Y and A via a white-box adversary and a neural network. Adel et al. [1], Ganin et al. [19], Raff and Sylvester [30] all proposed to optimize a variant of the following loss function following

$$L(\theta, \phi) = E_{\theta,\phi}(X, Y) - \lambda D_{\theta,\phi}(X, A), \tag{4}$$

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.

Gradient reversal was introduced by Ganin et al. [19] for domain adaptation, and later adapted by Raff and Sylvester [30] and Adel et al. [1] who treated the protected attribute A as a domain label. The gradient reversal strategy assumes that multiplying by a negative sign will increase the loss $D_{\theta,\phi}(X, A)$ of the branch $h_a : X \to \hat{A}$ and yields a representation X^* that is maximally invariant to changes in A [1, 30].

Using gradient reversal for fairness is based on the intuition that the inability to predict A is a suitable fairness goal. This differs slightly from the fairness evaluations presented in Section 2.2, but a similar loss function from Equation 4 based on demographic parity led to the architecture of FAD [1], which leverages gradient reversal specifically for fairness.

However, there is no guarantee that gradient descent with flipped gradients does guarantee the maximal invariance required for fairness. In the worst case, maximizing the loss $D_{\theta,\phi}(X, A)$ can even result in the opposite optimum for the shared layers with regard to A, because flipping the gradients with regard to A makes it perform gradient ascent for A. With the shared layers performing gradient ascent w.r.t. A followed by gradient descent in the adversarial branch, this creates a discrepancy between the parameters defining both components for predicting A. This means that the model is not only not maximally invariant on the last shared layer, but that the shared layers are still explicitly learning to predict the protected attribute A.

This is one of the major limitations of using GRL for fair models, as predictions of main attribute Y are not made on 'fair' representations. 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. In Section 3 we clarify how our ethical adversaries framework mitigates this issue, thus allowing GRL to be used for training fair models.

3 Ethical Adversaries Framework

Our main contribution is a framework that joins evasion attacks (see Section 2.1) and fair neural networks (see Section 2.3) to improve the overall fairness of the system. Thus, it relies on two types of ethical adversaries: (i) a *Feeder* that uses evasion attacks to create examples highlighting unfair representation of a certain population and (ii) an *adversarial Reader* that is trying to predict the protected attributes of interest (age, gender, race, etc.). In addition of exhibiting fairness issues in the data and in the trained model, our framework leverages gradient reversal to minimise the ability of the reader to guess protected attributes ultimately yielding a fairer ML model without sacrificing utility.

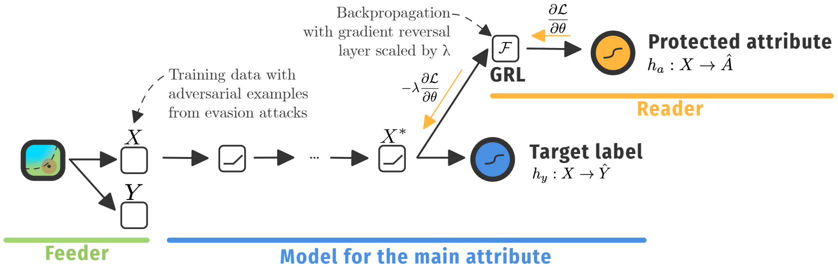

Backpropagation with gradient reversal layer scaled by λ
$\frac{\partial \mathcal{L}}{\partial \theta}$
$\mathcal{F}$
GRL
Protected attribute
$h_a : X \to \hat{A}$
Reader
Training data with adversarial examples from evasion attacks
$-\lambda \frac{\partial \mathcal{L}}{\partial \theta}$
X
X^*
...
Target label
$h_y : X \to \hat{Y}$
Y
Feeder
Model for the main attribute

Fig. 1: Ethical adversaries architecture: adversarial feeder on the left, and integrated adversarial reader on the right.

Figure 1 presents the global architecture. Our network follows a typical architecture with a GRL (discussed in Section 2.3 and is represented by the Reader). The Feeder, on the left part, performs evasion attacks as discussed in Section 2.1. Both adversaries interact with each other in an iterative manner—which is the main difference between our framework and GANs [20]. To achieve better fairness and utility outcomes, the process, that consists of two steps, can be performed multiple times.

The first step starts with a trained neural network (target label in Figure 1) predicting a main attribute Y. In this network, the adversarial Reader adds a second branch that tries to predict a protected attribute A while the gradient reversal layer strives to minimise the confidence of the Reader to predict A. Additionally, as we discussed in Section 2.3, during the backward pass, a hyperparameter λ contributes to prioritize the utility versus the adversarial branch of the network. The model is trained with the joint loss of the original prediction target and the protected attribute.

In a second step, the Feeder, on the left part, performs evasion attacks as discussed in Section 2.1. The Feeder creates a set of adversarial points from an approximation of the target model, a.k.a a surrogate model, that is constructed on the same dataset as the model under attack. Our surrogate model is an SVM which it uses a radial basis (RBF) kernel function to cope with different level of model complexity. We selected this kernel since preliminary results on COMPAS showed that it is expressive enough and Biggio et al. [5] detailed how evasion attacks can be directly applied to SVMs with RBF kernels. The Feeder performs multiple evasion attacks on the surrogate function to generate adversarial examples that are similar to the training examples, but are wrongly classified.

For each iteration of this two-step process, adversarial examples are generated and included in the training set for adversarial retraining. Each adversarial example is added to the training set with the same label as the original example from which it was generated. The effect of the ratio of adversarial points in the dataset—the adversarial fraction—is further analyzed empirically in Section 4.3.

In terms of performance, constructing a surrogate classifier is the limiting factor. Using SVMs implies that the time complexity of the entire framework is $\mathcal{O}(n^3)$ with n the number of data points. The impact of adversarial attacks is linear on the overall complexity. But note that adversarial retraining may drastically increase time to compute a separating function since included adversarial examples make the separation more difficult to find, or on the contrary, may not affect the function at all, if few adversarial examples are included.

Both reading and feeding steps are run successively until we achieve better fairness and utility outcomes, which we demonstrate in Section 4.4. A key benefit of this process is that we prevent the Reader from learning biased representations, since these features cannot be used as proxies for the protected attribute anymore.

4 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, 10, 11, 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 other groups are clearly less represented (e.g., only 31 instances for people of Asian descent), this poses issues during training and evaluation. It 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[4] (v0.11) for running evasion attacks.

4.1 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, including the adversarial reader, is 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 4 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).

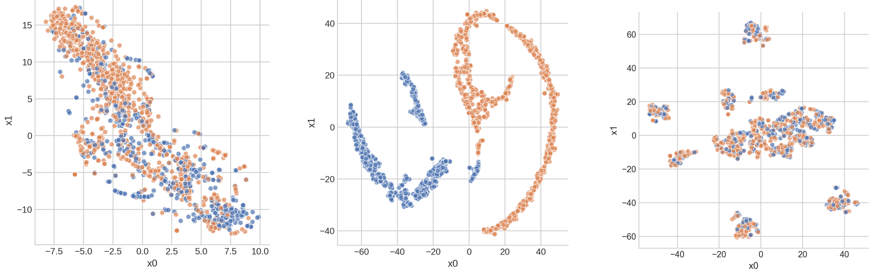
The adversarial feeder. In our setting, we can use the same training set for both the feeder and reader since they are part of the same, unique architecture.

We also approximate—relying on the earlier discussed transferability of attacks [12]—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

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

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.

4.2 Mitigating unfair representations



(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 ● .

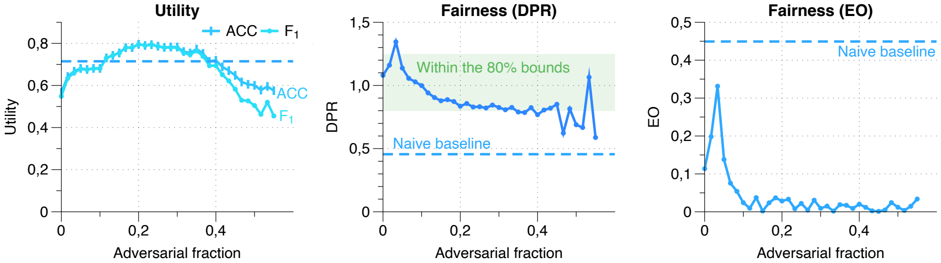
For each individual for the COMPAS test set, all three models derive a representation in the last hidden layer, on which we applied a t-SNE dimensionality reduction for a two-dimensional visualisation.

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 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 $AUC = 0.71$. The adversarial branch h_a that was trained simultaneously has an $AUC = 0.44$ As we mentioned earlier, this branch can be limited in predicting the protected attribute A. Which is the case here, as an independent perceptron has $AUC = 0.92$.

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', also called the Reader, and 'adversarial learning', or the Feeder, (see Figure 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 successfully mitigated this unfair representation, as shown in Figure 2c.

4.3 Effect of adversarial fraction

Figure 3 displays 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. We use this fraction for all further experiments, in future work this could be automated with a custom stopping criterion.



(a) Utility (b) DPR (c) EO

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)

4.4 Benchmark results

Table 1 presents our results on the three datasets. We compare them with (i) a baseline without fairness goals, *i.e.*, a neural network without any particular control on fairness aspects, (ii) a re-implementation of the GRL [1, 19, 30] and (iii) the reported results from other works that incorporate fairness and cover a wide range of learning algorithms: Naive Bayes [9], random forests [31], SVMs [35] and neural networks [30, 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

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) [9]	0.773†	—	**0.000**†	—	—
NBF (EM) [9]	0.801†	—	0.001†	—	—
Grad-Pred [30]	0.754†	—	**0.000**†	—	—
FF [31]	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 [30]	0.675†	—	0.001†	—	—
FF [31]	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 risk predictions [2]	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

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 shown 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 baseline has a EO of 2.192, the "bias against blacks" [2] *decreases* substantially with our model. For GRL, the near-equality of DPR (0.926) appears fairer, but this is not the case for DP and EO, where we observe an EO of 0.449 for GRL versus 0.008 for our model.

5 Code

We release an open source implementation— under the MIT licence—of our framework at `https://github.com/iPieter/ethical-adversaries`.

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) and fairness. 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 modifications of the secML framework to allow multiple output values in neural networks. Third, one could define constraints involving multiple features. Enforcing these *domain-specific* constraints during attack generation raises questions on the representation of the feature space and optimal convergence of the algorithms. Fourth, our framework is evaluated against allocational harms. More subtle differences— like a difference in the model's performance—are also affecting social groups. With some minor modifications, we suspect that these types of unfairness can be addressed with our framework. 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.

Acknowledgements

Pieter Delobelle was supported by the Research Foundation - Flanders under EOS No. 30992574 and received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. Gilles Perrouin is an FNRS Research Associate. This research was partly supported by EOS Verilearn project grant no. O05518F-RG03. We also want to thank the secML developers from the PRALab (Pattern Recognition and Applications Laboratory, University of Cagliari, Sardegna, Italy) for having answered our numerous questions and helping us in using their newly developed library.

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. ProPublica (2016)
3. Barocas, S., Hardt, M., Narayanan, A.: Fairness and Machine Learning. fairmlbook.org (2019)
4. Biggio, B., Didaci, L., Fumera, G., Roli, F.: Poisoning attacks to compromise face templates. In: 2013 International Conference on Biometrics (ICB). pp. 1–7 (2013)
5. 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)
6. Biggio, B., Nelson, B., Laskov, P.: Poisoning attacks against support vector machines. In: Proceedings of the 29th International Conference on Machine Learning. pp. 1467–1474. ICML'12 (2012)
7. Biggio, B., Roli, F.: Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition Journal 84, 317–331 (2018)
8. Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (Technology) is Power: A Critical Survey of "Bias" in NLP. arXiv:2005.14050 [cs] (May 2020)
9. Calders, T., Verwer, S.: Three naive Bayes approaches for discrimination-free classification. Data Min. Knowl. Discov. 21(2), 277–292 (2010)
10. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. arXiv:1703.00056 (2017)
11. Corbett-Davies, S., Goel, S.: The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. arXiv:1808.00023 (2018)
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 (10)
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. Kulynych, B., Overdorf, R., Troncoso, C., Gürses, S.: POTs: Protective Optimization Technologies. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency pp. 177–188 (Jan 2020)
25. Madras, D., Creager, E., Pitassi, T., Zemel, R.S.: Learning Adversarially Fair and Transferable Representations. arXiv abs/1802.06309 (2018)
26. Overdorf, R., Kulynych, B., Balsa, E., Troncoso, C., Gürses, S.: Questioning the assumptions behind fairness solutions. arXiv:1811.11293 (2018)
27. 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)
28. 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. pp. 506–519. ACM (2017)
29. 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)
30. 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)
31. 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)
32. Sattigeri, P., Hoffman, S.C., Chenthamarakshan, V., Varshney, K.R.: Fairness GAN: Generating Datasets With Fairness Properties Using a Generative Adversarial Network. IBM Journal of Res. and Dev. p. 12 (2019)
33. Solans, D., Biggio, B., Castillo, C.: Poisoning attacks on algorithmic fairness. arXiv preprint arXiv:2004.07401 (2020)
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)