Ethical Adversaries: Towards Mitigating Unfairness with Adversarial Machine Learning

Pieter Delobelle
Department of Computer Science, KU Leuven, Leuven.AI
Leuven, Belgium
pieter.delobelle@kuleuven.be

Paul Temple,
Gilles Perrouin,
Benoit Fréney,
Patrick Heymans
PReCISE, NaDi, University of Namur
Namur, Belgium

Bettina Berendt
Department of Computer Science, KU Leuven, Leuven.AI
Leuven, Belgium
Faculty of Electrical Engineering and Computer Science, TU Berlin
Berlin, Germany
bettina.berendt@kuleuven.be

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 investigated the trade-offs between these targets in terms of model hyperparameters and also illustrated 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 individuals. 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, they still exhibit unfair behaviour. Such behaviours can impact individuals belonging to a specific social group. 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 for a group, the decision towards another group to not receive the same outcome can, ultimately, create a feedback loop where unfair behaviours are amplified [19, 32]. 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.

To avoid such consequences, researchers increasingly focused on incorporating fairness as objectives in their systems. In discrimination-aware data mining (DADM), modifications were developed and applied to data, learning algorithms, or resulting patterns and models [25]. Recently, adversarial fairness continues in this direction with, for instance, research on learning representations [31, 13] and task-specific fair models [11, 30, 38]. Adversaries are also used when assessing the security of machine learning based systems. Biggio and Roli [5] synthesized a decade of research in adversarial machine learning. This domain aims at finding or creating examples that are problematic for a machine learning model, e.g. Biggio et al. [7], Papernot et al. [43, 44]. These examples can be injected directly into the training phase 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. Our framework improves demographic parity and equal opportunity when comparing to the state of the art while globally improving the model’s utility. We thus reconcile fairness and model performance.

This paper is an extension of our previous work [12], which was presented at the first workshop on Bias and Fairness in AI (BIAS 2020) held jointly with the ECML-PKDD conference. With respect to the original contribution, we provide: (i) a detailed analysis of tradeoffs between fairness and utility by computing and charting a Pareto front; (ii) a more nuanced interpretation of our results based on 30 random hyperparameter trials, where we calculated the expected validation accuracy, and (iii) a discussion on the role of prior domain knowledge to create both valid and realistic examples to feed to the model.

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 7 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. 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) [7, 33, 34]. 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. After the attack finishes, because of these assumptions, newly generated data examples can be transferred back to the original target model [13]. Only one restriction remains on the surrogate classifier, attacks are gradient-based techniques requiring the model to be trained [3]. In the latter, malicious examples are provided at test time to harm the model without changing it [7].

In related work by Solans et al. [30], 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. [29] 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 external impacts, 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, our exploitation goal differs from [39].

2.2 Evaluating Fairness

There exist several measures of fairness in the literature, e.g. demographic parity [16], equalized odds and equalized opportunity [20], statistical parity [21, 43], disparate impact [13, 24], and threshold testing [35]. They are categorized into different family of measures, presented in the FairML book [3], depending on their mathematical expression. In the following, we focus on two different measures: (i) demographic parity which is probably the most popular but also controversial and (ii) equalized opportunity also quite popular but from a different family of measure. 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. Measures usually come in two forms. The first one expresses the requirement that the predicted values of the classifier \( \hat{Y} \) conditioned on the protected attribute be equal [9]. Most of the time, this strict equality in the outcome predictions between both groups is unrealistic. The relaxed form accepts the previous equality not to be strict within a range that is to be determined. For instance, in a legal setting, the US Equal Employment Opportunity Commission (EEOC) uses the Demographic Parity ratio (as defined by Def. 2.1) and 2.2 with \( \tau = 0.8 \) (“80% rule”) [21], stating that disparate impact caused by employment-related decisions or structures can only be ascertained if \( DPR \leq 0.8 \).

**Definition 2.1.** Demographic parity (DP). DP is the equality or similarity of prediction outcomes as an absolute difference [16, 36]:

\[
DP = \left| P(\hat{Y} = 1 \mid A = 0) - P(\hat{Y} = 1 \mid A = 1) \right| \leq \varepsilon. \tag{1}
\]

**Definition 2.2.** Demographic parity ratio (DPR). DPR is the equality or similarity of prediction outcomes as a ratio:

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

Demographic parity has received some criticisms, since the measure does not necessarily report on what many would define as fairness [16]. This issue stems from ignoring both the true outcome and individual merits. For instance, consider a selection procedure where two individuals apply and both belong to the same protected group. The entity that has to make the selection needs to select one individual from this
protected group, regardless of their qualification, in order to achieve demographic parity. 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. Deliberately selecting the unqualified individual would not be considered as fair regarding the other (qualified) applicant; however, the entity would still satisfy demographic parity. So these tokens individuals are not guaranteeing fairness since qualified individuals from the protected group are still mistreated.

Addressing the criticisms of demographic parity, Hardt et al. 20 presented two 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. 16, which can otherwise be difficult to come up with. This is also very similar to disparate mistreatment 44, 41 used as an evaluation metric by Adel et al. 1.

**Definition 2.3.** Equal opportunity (EO). EO requires an independence $\hat{Y} \perp A \mid Y \hat{Y}$ and $A$ conditioned on the true outcome $Y$. Expressed as a difference, this yields:

$$P(\hat{Y} = 1 \mid A = 0, Y = 1) - P(\hat{Y} = 1 \mid A = 1, Y = 1) \leq \nu,$$

"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

Several works tried to reduce the impact of data imbalance 20, 36 and data representations over ML models 21, 27. In addition, fair models have been studied for a variety of learning algorithms, such as Naive Bayes classifiers 10 or SVMs 42. Nowadays, the focus is also on neural networks due to their prediction performance 11, 31, 37. Some researchers mitigate unfairness in neural networks using white-box adversaries 17, 21, 31, 37, 14. 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 17 over the loss function $L$, i.e., $\min_\theta \max_\phi L(\theta, \phi)$, with an adversary $\phi$ and an encoder with parameters $\theta$. 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. 22, Raff and Sylvester 36 all proposed to optimize a variant of the following loss function following

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

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 $\lambda$ a hyperparameter.

Gradient reversal was introduced by Ganin et al. 22 for domain adaptation, and later adapted by Raff and Sylvester 36 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 $h_a : X \rightarrow A$ and yields a representation $X'$ that is maximally invariant to changes in $A$ 14, 26.

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 3 based on demographic parity led to the architecture of FAD 11, 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 18 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

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 particular population and (ii) an adversarial Reader that tries 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.

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, forming the main difference between our framework and GANs 23. To achieve better fairness and utility outcomes, our 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
The Feeder is constrained to a maximum perturbation, so adversarial points in the dataset—the adversarial fraction—is limited minorities. However, we recognize that this unconstrained approach can be at odds with business logic and future work could focus on including constraints.

In terms of performance, constructing a surrogate classifier is the limiting factor. Using SVMs implies that the time complexity of the entire framework is $O(n^3)$ with $n$ the number of data points. The impact of adversarial attacks is linear on the overall complexity. But we should notice 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 too 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 4.3. 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 [28]. 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] [12]. The dataset, therefore, includes the race of individuals. In line with previous research [1] [2] [12], 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 raises 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 [7] (v0.11) for running evasion attacks.

\[\text{https://secml-gitlab.io/}\]
Model trained with a GRL or Caucasian.

In our setting, we can use the same adversarial feeder.

The adversarial reader. The adversarial reader is part of the model under attack.

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 re-training 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 $A$.

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.

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][2] [3][4] and (iii) the reported results from other works that incorporate fairness and cover a wide range of learning algorithms: Naive Bayes [5], random forests [6][7], SVMs [8] and neural networks [9][10][11]. 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
Within the 80% bounds

![Utility](a) Utility

![DPR](b) DPR

![EO](c) EO

Figure 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: 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>ACC</th>
<th>F1</th>
<th>DP</th>
<th>DPR</th>
<th>EO</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Adult</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline without fairness constraints</td>
<td><strong>0.639 ± 0.009</strong></td>
<td><strong>0.763</strong></td>
<td>0.173</td>
<td>0.196</td>
<td>0.196</td>
</tr>
<tr>
<td>GRL</td>
<td>0.612 ± 0.012</td>
<td>0.518</td>
<td>0.059</td>
<td>1.931</td>
<td>0.061</td>
</tr>
<tr>
<td>NBF (NB) [9]</td>
<td>0.773↑</td>
<td>—</td>
<td>0.001↑</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>NBF (EM) [9]</td>
<td>0.801↑</td>
<td>—</td>
<td>0.001↑</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Grad-Pred [96]</td>
<td>0.754↑</td>
<td>—</td>
<td>0.000↑</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>FF [37]</td>
<td>0.753↑</td>
<td>—</td>
<td>0.000↑</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>LFR [43]</td>
<td>0.702↑</td>
<td>—</td>
<td>0.001↑</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>0.814 ± 0.009</strong></td>
<td>0.689</td>
<td>0.031</td>
<td><strong>0.784</strong></td>
<td>0.179</td>
</tr>
<tr>
<td><strong>German Credit</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline without fairness constraints</td>
<td>0.705 ± 0.063</td>
<td>0.624</td>
<td>0.018</td>
<td>0.929</td>
<td>0.198</td>
</tr>
<tr>
<td>GRL</td>
<td>0.710 ± 0.063</td>
<td>0.415</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Grad-Pred [96]</td>
<td>0.675↑</td>
<td>—</td>
<td>0.001↑</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>FF [37]</td>
<td>0.700↑</td>
<td>—</td>
<td>0.000↑</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>LFR [43]</td>
<td>0.591↑</td>
<td>—</td>
<td>0.004↑</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>0.730 ± 0.062</strong></td>
<td><strong>0.640</strong></td>
<td>0.006</td>
<td><strong>0.971</strong></td>
<td>0.175</td>
</tr>
<tr>
<td><strong>COMPAS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline without fairness constraints</td>
<td>0.715</td>
<td>0.709</td>
<td>0.466</td>
<td>2.192</td>
<td>0.449</td>
</tr>
<tr>
<td>GRL</td>
<td>0.567</td>
<td>0.549</td>
<td>0.057</td>
<td><strong>0.926</strong></td>
<td>0.114</td>
</tr>
<tr>
<td>COMPAS risk predictions [2]</td>
<td>0.655 ± 0.029</td>
<td>0.654</td>
<td>0.289</td>
<td>1.829</td>
<td><strong>0.000</strong></td>
</tr>
<tr>
<td>Preference-based fairness [12]</td>
<td>0.675↑</td>
<td>—</td>
<td>0.380↑</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>0.794</strong></td>
<td><strong>0.793</strong></td>
<td><strong>0.026</strong></td>
<td>0.840</td>
<td>0.008</td>
</tr>
</tbody>
</table>

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 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” 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. Figure 5 also reports the accuracy on COMPAS, however in function of the number of hyperparameter trials to illustrate how randomized hyperparameter assignments will affect the results [15]. The hyperparameter $\lambda$ was varied between the interval [0.1,200] and the batch size selected from the set {256, 512, 1024, 2048}. Based on these results in relation to both (i) a model without any fairness constraints and (ii) a model with a GRL, we can conclude that our framework does perform better on this task compared to GRL. In ad-
dation, given a sufficiently large compute budget it will even surpass the naive model without fairness constraints, since our model likely benefits from additional training examples. Nevertheless, it remains sensitive to optimal or suboptimal hyperparameter assignments, as is indicated by the large confidence interval. This sensitivity is also confirmed by the Pareto front in Figure 4, where many trials with large (> 50) values for \( \lambda \) performed quite poor on both utility and fairness.

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

6. TOWARDS DOMAIN KNOWLEDGE INTEGRATION
While adversarial machine learning has been applied to the world of images and videos, we applied it in a different context. Our experiments aim at using evasion attacks to attempt generating new instances that represent personas, profiles or persons’ representations. This context therefore forms a new domain space bringing its own set of constraints and value boundaries. In this section, we discuss the challenges this context brings.

The first challenge relates to constraints’ heterogeneity. Indeed, in the image domain, the representation of pixels is homogeneous and thus easy to constrain while running evasion attacks. Yet, when it comes to different features representing categorical information, the problem is more complex. Consider age category or highest obtained diploma, this information is part of the global description of the profile of a person and are usually categorical as only one choice is important/allowed. Some machine learning algorithms (e.g. SVMs) do not support easily such kind of categorical feature. Still, to be able to use such algorithms with this kind of features, categories can be decoupled into different features so that are mutually exclusive. For instance, regarding the highest diploma, one would create the following features: “has a Bachelor’s Degree”, “has a Master’s Degree”, “has a PhD Degree”, etc. While evasion attacks will bring perturbations to features independently, one needs to check the validity of the attacks [40].

There are different strategies: check constraints after each iteration and reject the latest modifications if any of these constraints is violated, wait until the attack has finished to reject the example in case of constraint violation, or even tradeoffs between these two extreme strategies. In this paper we did not applied a constraint enforcement strategy (see Section 3), but rather focused on limited perturbations that would generate similar examples to instances in the datasets, therefore likely to be valid. A stricter constraint enforcement strategy is left for future work.

The second challenge is linked to the prior domain knowledge that should also be taken into account in order to make a generated example realistic. For instance, in the context of the German credit dataset, it would be unlikely (and most probably illegal) that a kid (let’s say under 10 years old) surfs to an online banking website to ask for a credit. We can easily transform this prior legal knowledge into a hard constraint and therefore consider that case invalid. As previously discussed, such constraints differ from the image domain. We would need the help of solvers which can take time to run and check all of them. Furthermore, some engineering may be needed to transform features’ value into interpretable constraints for solvers and then back to the feature space. We were able to check some boundary constraints for features’ value, but incorporating richer domain-knowledge is left to future work.

Our kid would be even more unusual if they additional stated that they had a PhD. While the fictional Sheldon Cooper from the Big Bang Theory show got his PhD at the age of sixteen and there exist actual cases of people obtaining their PhDs at very young age[1], these are the kind of examples we may still want to avoid since there not representative. Because we relied on machine learning and statistical approaches, we did not see the case of one of our attack generating a kid holding a PhD. This can be explained by the fact that no examples were given looking alike it, thus the data distribution of the examples are not showing that this

https://en.wikipedia.org/wiki/Kim_Ung-yong
is a possibility (the stochastic gradient descent procedure has no interest in trying to generating examples toward this direction of the feature space).

7. CONCLUSION 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 $P_1$ 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 (FWO) under EOS No. 30992574 (VeriLearn). Pieter Delobelle also received funding from the Flemish Government under the ‘Onderzoeksprogramma Artificiële Intelligenie (AI) Vlaanderen’ programme. Paul Temple is also supported by the EOS VeriLearn project (Fonds de la Recherche Scientifique, FNRS). Gilles Perrouin is an FNRS Research Associate. We also want to thank the secML developers from the PRA Lab (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


40. Paul Temple, Mathieu Acher, Gilles Perrouin, Battista Biggio, Jean-Marc Jézéquel, and Fabio Roli. Towards quality assurance of software product lines with adversarial configurations. In Thorsten Berger, Philippe Collet,

