Exacerbating Algorithmic Bias through Fairness Attacks

Ninareh Mehrabi[1,2], **Muhammad Naveed**[1], **Fred Morstatter**[1,2], **Aram Galstyan**[1,2]
[1]University of Southern California - [2]Information Sciences Institute
{ninarehm, mnaveed}@usc.edu, {fredmors, galstyan}@isi.edu

Abstract

Algorithmic fairness has attracted significant attention in recent years, with many quantitative measures suggested for characterizing the fairness of different machine learning algorithms. Despite this interest, the robustness of those fairness measures with respect to an intentional adversarial attack has not been properly addressed. Indeed, most adversarial machine learning has focused on the impact of malicious attacks on the accuracy of the system, without any regard to the system's fairness. We propose new types of data poisoning attacks where an adversary intentionally targets the fairness of a system. Specifically, we propose two families of attacks that target fairness measures. In the *anchoring attack*, we skew the decision boundary by placing poisoned points near specific target points to bias the outcome. In the *influence attack on fairness*, we aim to maximize the covariance between the sensitive attributes and the decision outcome and affect the fairness of the model. We conduct extensive experiments that indicate the effectiveness of our proposed attacks.

Introduction

With proliferation of machine learning (ML) applications in everyday life, it is imperative that ML algorithms underlying those applications do not discriminate, especially when it comes to potentially sensitive and consequential decisions, such as bail decisions (Dressel and Farid 2018). Thus, recent research has looked into possible biases present in ML algorithms, and proposed different measures and definitions for characterizing fairness (Dwork et al. 2012; Hardt, Price, and Srebro 2016; Kusner et al. 2017; Verma and Rubin 2018; Mehrabi et al. 2019).

Despite this interest, not much is known about the robustness of various fairness measures with respect to random, or perhaps malicious, perturbations. Indeed, it is known that machine learning models can be susceptible to various types of adversarial attacks targeted to degrade the performance of machine learning models. However, research in adversarial machine learning has mostly focused on targeting accuracy (Chakraborty et al. 2018; Li et al. 2018). We argue that, like accuracy, fairness measures can be targeted by malicious adversaries as well. For instance, adversaries can attack models used by a government agency with the goal of making them appear unfair in order to depreciate their value and credibility. Some adversaries can even profit from such attacks by biasing decisions for their benefit, e.g., in credit or loan applications. Thus, one should consider fairness when assessing the robustness of ML systems.

Our contributions. In this work, we propose data poisoning attacks that target fairness. We propose two families of poisoning attacks: *anchoring* and *influence*[1]. In anchoring attacks the goal is to place poisoned points to affect fairness without modifying the attacker loss. On the other hand, our influence attack on fairness can affect both fairness and accuracy by injecting poisoned points during train time via a specific adversarial loss that regularizes between fairness and accuracy losses. Some adversaries may want to harm systems with regard to fairness and accuracy at the same time, while others might only consider one that can be achieved by this regularization. In the anchoring attack, we place poisoned points to bias the decision boundary; in the influence attack, we target fairness measures by incorporating a loss function maximizing and attacking which can degrade fairness by maximizing the covariance between the decision outcome and sensitive attributes.

Through experimentation on three different datasets with different fairness measures and definitions, we show the effectiveness of our attacks in achieving the desired goal of affecting fairness. In addition, we incorporate different baseline models to evaluate different aspects of our attacks. We demonstrate that original data poisoning attacks designed to attack accuracy are not suitable for fairness attacks, thus highlighting the importance of attacks designed for fairness. We also compare our methods against concurrent work on adversarial attacks on fairness and show the effectiveness of our methods in comparison.

Background on Poisoning Attacks

Consider a supervised learning problem characterized by a loss function $\mathcal{L}(\theta; \mathcal{D})$ and an adversarial loss $L_{adv}(\hat{\theta}; \mathcal{D})$, where $\hat{\theta}$ is the set of learnable parameters and $\mathcal{D}$ is a labeled dataset. Let $\mathcal{D}_{train}$ be the training dataset. We assume that the adversary can poison a fraction of those data points, so that $\mathcal{D}_{train} = \mathcal{D}_c \cup \mathcal{D}_p$, where $\mathcal{D}_c$ and $\mathcal{D}_p$ are the set of clean and poisoned data points, respectively. We assume

[1]https://github.com/Ninarehm/attack

that $|\mathcal{D}_p| = \epsilon|\mathcal{D}_c|$. Furthermore, $\mathcal{D}_p \subseteq \mathcal{F}_\beta$ where $\mathcal{F}_\beta$ is the feasible set, which is a set selected by a defense mechanism based on anomaly detection techniques, containing elements that the defender considers as sanitized data to train its model. The existence of the feasible set in the objective helps the poisoned points to blend with the natural data and make it more difficult for anomaly detector techniques to detect them (Koh, Steinhardt, and Liang 2018).

A data poisoning attack can be written as the following optimization problem (over the set of poisoned data points):

$$\begin{aligned} &\max_{\mathcal{D}_p} L_{adv}(\hat{\theta}; \mathcal{D}_{test}) \\ &s.t. \;\; |\mathcal{D}_p| = \epsilon|\mathcal{D}_c| \\ &\quad\;\; \mathcal{D}_p \subseteq \mathcal{F}_\beta \\ &\text{where} \;\; \hat{\theta} = \arg\min_{\theta} \; \mathcal{L}(\theta; \mathcal{D}_c \cup \mathcal{D}_p). \end{aligned} \tag{1}$$

In essence, the adversary attempts to maximize its test loss L_{adv} by carefully selecting poisoned data points. These types of attacks are shown to be powerful against defenders that are trying to minimize their own loss $\mathcal{L}$, while the attacker is trying to harm the defense (Koh, Steinhardt, and Liang 2018). In (Koh, Steinhardt, and Liang 2018), authors propose to sample a positive $(\tilde{x}_+, +1)$ and a negative $(\tilde{x}_-, -1)$ instance and make $\epsilon|\mathcal{D}_c|$ copies from these sampled instances to serve as poisoned data points inversely proportional to the class balance such that there are $(|\mathcal{D}_c^+|\epsilon)$ copies from the negative poison instance $(\tilde{x}_-, -1)$ and $(|\mathcal{D}_c^-|\epsilon)$ copies from the positive poison instance $(\tilde{x}_+, +1)$ in which $|\mathcal{D}_c^+|$ and $|\mathcal{D}_c^-|$ represent the number of positive and negative points in the clean data respectively.

Poisoning Attacks against Fairness

Now that we have discussed poisoning attacks, we will discuss how these attacks can be extended to fairness. We follow a common fairness setup where there are two groups: advantaged and disadvantaged. An example of advantaged and disadvantaged groups can be male and female in the job market where males could have advantage over females in getting hired in certain jobs. We assume that all poisoned points belong to either the advantaged or disadvantaged group, $\mathcal{D}_p \subseteq (\mathcal{D}_a \cup \mathcal{D}_d)$, in which $\mathcal{D}_a$ represents data points from the advantaged demographic group and $\mathcal{D}_d$ represents data points from the disadvantaged demographic group.

Influence Attack on Fairness

For the influence attack on fairness, we use the influence attack introduced in (Koh, Steinhardt, and Liang 2018; Koh and Liang 2017), with a modification that includes the demographic information, in which the attack tries to maximize a given loss. We then incorporate a loss function maximizing which using the influence attack can harm fairness. In (Zafar et al. 2015), authors propose a loss function for fair classification with a constraint involving the covariance between the sensitive features (z) and the signed distance from feature vectors to the decision boundary ($d_\theta(x)$) formalized as:

$$Cov(z, d_\theta(x)) \approx \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_\theta(x_i).$$

Algorithm 1: Influence Attack on Fairness

Input: clean data set $\mathcal{D}_c = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, poison fraction ϵ, and step size η.
Output: poisoned data set $\mathcal{D}_p = \{(\tilde{x}_1, \tilde{y}_1), (\tilde{x}_2, \tilde{y}_2), ..., (\tilde{x}_{\epsilon n}, \tilde{y}_{\epsilon n})\}$.
From $\mathcal{D}_a$ randomly sample the positive poisoned instance $\mathcal{I}_+ \leftarrow (\tilde{x}_1, \tilde{y}_1)$.
From $\mathcal{D}_d$ randomly sample the negative poisoned instance $\mathcal{I}_- \leftarrow (\tilde{x}_2, \tilde{y}_2)$.
Make copies from $\mathcal{I}_+$ and $\mathcal{I}_-$ until having $\epsilon|\mathcal{D}_c|$ poisoned copies $\mathcal{C}_p$.
Load poisoned data set $\mathcal{D}_p \leftarrow \{\mathcal{C}_p\}$.
Load feasible set by applying anomaly detector B $\mathcal{F}_\beta \leftarrow B(\mathcal{D}_c \cup \mathcal{D}_p)$.
for *t= 1,2,...* **do**
  $\hat{\theta} \leftarrow argmin_\theta \; \mathcal{L}(\theta; (\mathcal{D}_c \cup \mathcal{D}_p))$.
  Pre-compute $g_{\hat{\theta}, \mathcal{D}_{test}}^\top H_{\hat{\theta}}^{-1}$ from L_{adv} for details refer to (Koh, Steinhardt, and Liang 2018).
  for *i= 1,2* **do**
    Set $\tilde{x}_i^0 \leftarrow \tilde{x}_i - \eta g_{\hat{\theta}, \mathcal{D}_{test}}^\top H_{\hat{\theta}}^{-1} \frac{\partial^2 \ell(\hat{\theta}; \tilde{x}_i, \tilde{y}_i)}{\partial \hat{\theta} \partial \tilde{x}_i}$.
    Set $\tilde{x}_i \leftarrow argmin_{x \in \mathcal{F}_\beta} ||x - \tilde{x}_i^0||_2$. (To project $\mathcal{D}_p$ back to $\mathcal{F}_\beta$).
  end
  Update copies $\mathcal{C}_p$ based on updates on $\mathcal{I}_+$ and $\mathcal{I}_-$.
  Update feasible set $\mathcal{F}_\beta \leftarrow B(\mathcal{D}_c \cup \mathcal{D}_p)$.
end

By combining the above constraint with the original classification loss and maximizing it, the attacker can harm both fairness and accuracy at the same time via a regularization term, λ, that controls the trade-off between these two terms. Thus, the loss in our influence attack on fairness contains two parts: ℓ_{acc} and $\ell_{fairness}$ in which ℓ_{acc} controls for accuracy and $\ell_{fairness}$ controls for fairness constraints.

$$\begin{aligned} L_{adv}(\hat{\theta}; \mathcal{D}_{test}) &= \ell_{acc} + \lambda \ell_{fairness} \\ where \;\; \ell_{fairness} &= \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_{\hat{\theta}}(x_i). \end{aligned} \tag{2}$$

In other words, the influence attack on fairness would try to harm the fairness constraint and affect a model with respect to disparate impact (Zafar et al. 2017). This loss can affect a model in terms of both fairness and accuracy with the regularization term λ that controls the trade-off. In order to maximize the loss in (2), we use the influence attack strategy (Koh, Steinhardt, and Liang 2018; Koh and Liang 2017) with changes that would incorporate demographic information as shown in Algorithm 1. Similar to the convention in (Koh, Steinhardt, and Liang 2018), we sample one positive and one negative instance uniformly at random and make copies of the sampled instances that serve as our poisoned points. However, since we now have to take demographics into consideration for maximizing the bias and harming fairness, we sample the positive instance from $\mathcal{D}_a$ and the negative instance from $\mathcal{D}_d$.

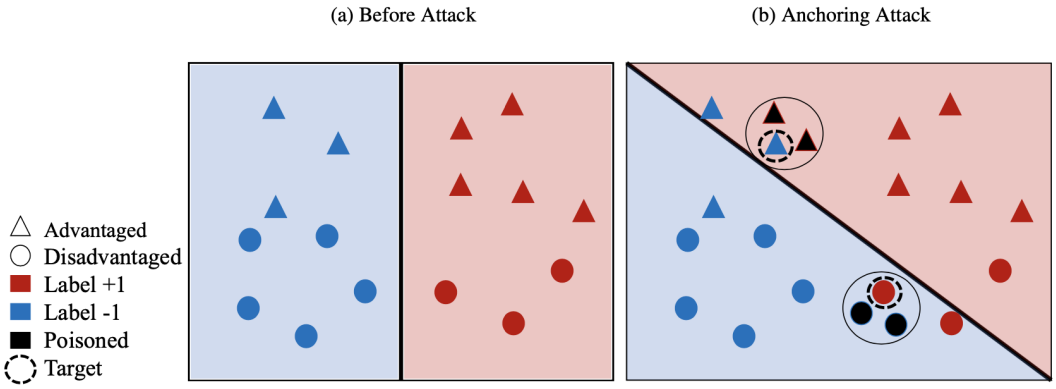

Figure 1: Anchoring attack representation. The figure on the left represents the before attack, while the right figure represents the anchoring attack in which poisoned points are located in close vicinity (depicted as the large solid circle) of target points.

Notice that the opposite is also possible if an adversary wants to skew the disadvantaged group into being advantageous; however, for the goals of this paper and showing how our methods can increase the bias and harm fairness, we follow the aforementioned sampling procedure.

Anchoring Attack

We now describe a simple generic anchoring attack that can work with any loss function. Our results indicate that the proposed attack harms the model with regard to fairness. The anchoring attack works as follows (details in Algorithm 2). First, the attacker samples a target x_{target} that belongs to the clean data, $x_{target} \in \mathcal{D}_c$. Next, the attacker generates poisoned data point $\tilde{x}$ in the vicinity of x_{target}, so that this new point has the same demographic but the opposite label, $demographic(x_{target}) = demographic(\tilde{x})$ and $y_{target} \neq \tilde{y}$. The general idea of the attack is to target some points (x_{target}) and cloud their labels through poisoned points that have opposite labels, which would lead to a skewed decision boundary, change in predictive labels of clean target points, and more biased outcomes. The right plot in Figure 1 depicts an anchoring attack in which the poisoned points colored in black are placed to lie close to the target points that have the same demographic group but opposite label to bias the predictive outcome (black advantaged poisoned points with label +1 are targeting advantaged point with label -1, and black disadvantaged poisoned points with label -1 are targeting disadvantaged point with label +1). This placement of poisoned points in the space during the learning procedure will lead the decision boundary to change and, as a result, will cause more advantaged points to have a predictive outcome of +1 and more disadvantaged points to have a predictive outcome of -1, which is biasing the model's prediction. $\mathbf{x_{target}}$ can be sampled in several ways. We introduce two ways, *random* and *non-random*, for sampling $\mathbf{x_{target}}$.

Random Anchoring Attack. In random anchoring attack, $\mathbf{x_{target}}$ is sampled uniformly at random for each demographic group.

Non-random Anchoring Attack. In the non-random anchoring attack, we choose popular $\mathbf{x_{target}}$ as our target for each demographic group. Here, popular $\mathbf{x_{target}}$ means the point that is close to more similar instances x_i, eligible to serve as targets, such that $demographic(x_i) = demographic(x_{target})$ and $y_i = y_{target}$. By doing this, we can ensure to affect as much as points similar to $\mathbf{x_{target}}$ as possible to maximize our biasing goal. Pick x with $max(c)$ as x_{target} where c is calculated for each x as follows: $\forall x_i$ if $demographic(x_i) = demographic(x)$ and $y_i = y$ and $||x_i - x|| < \sigma$ then increase c for x.

Algorithm 2: Anchoring Attack

Input: clean data set $\mathcal{D}_c = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, poison fraction ϵ, and vicinity distance τ.
Output: poisoned data set $\mathcal{D}_p = \{(\tilde{x}_1, \tilde{y}_1), (\tilde{x}_2, \tilde{y}_2), ..., (\tilde{x}_{\epsilon n}, \tilde{y}_{\epsilon n})\}$.
for *t= 1,2,...* **do**
  Sample negative x_{target^-} from $\mathcal{D}_a$ and positive x_{target^+} from $\mathcal{D}_d$ with random or non-random technique.
  $\mathcal{G}_+$: Generate $(|\mathcal{D}_c^-|\epsilon)$ positive poisoned points $(\tilde{x}_+, +1)$ with $\mathcal{D}_a$ in the close vicinity of x_{target^-} s.t. $||\tilde{x}_+ - x_{target^-}||_2 \leq \tau$.
  $\mathcal{G}_-$: Generate $(|\mathcal{D}_c^+|\epsilon)$ negative poisoned points $(\tilde{x}_-, -1)$ with $\mathcal{D}_d$ in the close vicinity of x_{target^+} s.t. $||\tilde{x}_- - x_{target^+}||_2 \leq \tau$.
  Load $\mathcal{D}_p$ from the generated data above $\mathcal{D}_p \leftarrow \mathcal{G}_+ \cup \mathcal{G}_-$.
  Load the feasible set $\mathcal{F}_\beta \leftarrow B(\mathcal{D}_c \cup \mathcal{D}_p)$.
  for *i=1,2,...,*ϵn **do**
    Set $\tilde{x}_i \leftarrow argmin_{x \in \mathcal{F}_\beta} ||x - \tilde{x}_i||_2$. (To project $\mathcal{D}_p$ back to $\mathcal{F}_\beta$).
  end
  $argmin_\theta \mathcal{L}(\theta; (\mathcal{D}_c \cup \mathcal{D}_p))$.
end

Evaluation

In our experiments, we evaluate our attacks with regards to different measures, such as accuracy and foundational fair-

ness measures: statistical parity, and equality of opportunity differences. We also utilize three real world datasets in our experiments, introduced below. We compare against a suite of baselines that test our attacks' performance with regards to accuracy and fairness. Our results indicate that our attacks, the anchoring attack and influence attack on fairness, are effective in terms of affecting fairness aspects of the model.

Datasets

We use three different real world datasets in our experiments with gender as the sensitive attribute. The data was split into an 80-20 train and test split.

German Credit Dataset. This dataset comes from UCI machine learning repository (Dua and Graff 2017). It contains the credit profile about individuals with 20 attributes associated to each data person. In our experiments, we utilized all the 20 attributes from this dataset. The classification goal is to predict whether an individual has good or bad credit.

COMPAS Dataset. Propublica's COMPAS dataset contains information about defendants from Broward County [2]. We utilized the features in Table 1 as our prediction features. The classification goal is to predict whether an individual will recommit a crime within two years.

Drug Consumption Dataset. This dataset comes from the UCI machine learning repository (Dua and Graff 2017). It contains information about individuals (Fehrman et al. 2017). We utilized the features listed in Table 1 as our prediction features. The classification goal is to predict whether an individual has consumed cocaine or not in their lifetime.

COMPAS	
sex	age_cat
juv_fel_count	juv_misd_count
priors_count	c_charge_degree
race	juv_other_count

Drug			
ID	Age	Gender	SS
Education	Country	Ethnicity	
Nscore	Escore	Oscore	
Ascore	Cscore	Impulsive	

Table 1: Features used from the COMPAS and Drug Consumption datasets.

Measures

In addition to accuracy, we have utilized two well-known fairness measures to analyze the performance of different attacks with regard to fairness, detailed below.

Statistical Parity Difference Statistical parity is a well-known measure (definition) introduced in (Dwork et al. 2012). We utilize this measure as one of our metrics for fairness. It captures the predictive outcome differences between different (advantaged and disadvantaged) demographic groups. The measure is defined below and is referred to as statistical parity throughout our paper.

$$SPD = |p(\hat{Y} = +1|x \in \mathcal{D}_a) - p(\hat{Y} = +1|x \in \mathcal{D}_d)|$$

Equality of Opportunity Difference Equality of opportunity is another well-known fairness definition introduced in (Hardt, Price, and Srebro 2016). We utilized the equality of opportunity difference as another fairness metric. It captures differences in the true positive rate between different (advantaged and disadvantaged) demographic groups. The measure is defined below and is addressed as equality of opportunity throughout this paper.

$$EOD = |p(\hat{Y} = +1|x \in \mathcal{D}_a, Y = +1) - p(\hat{Y} = +1|x \in \mathcal{D}_d, Y = +1)|$$

Methods

To evaluate our attacks, we compared them against an attack that does not consider fairness and only considers accuracy to show that such attacks are not necessarily effective for fairness, motivating the need for fairness attacks. Also, we compared our attacks in terms of how they attack accuracy as a measure versus attacks that are specifically designed to target accuracy. We also compared our attacks to an attack that is optimized for fairness.

The evaluated methods are listed below. In our experiments, the poisoned points are inversely proportional to class balance as also suggested in (Koh, Steinhardt, and Liang 2018), so we made $(|\mathcal{D}_c^+|\epsilon)$ copies from the negative poison instance ($\mathcal{I}_-$) and $(|\mathcal{D}_c^-|\epsilon)$ copies from the positive poison instance ($\mathcal{I}_+$) in which $|\mathcal{D}_c^+|$ and $|\mathcal{D}_c^-|$ denote the number of positive and negative points in the clean data respectively. Hinge loss was used to control for accuracy for all the methods in our experiments as in (Koh, Steinhardt, and Liang 2018).

Influence Attack on Fairness (IAF) In this paper, our influence attack on fairness is where the attack tries to maximize the covariance between the signed distance of feature vectors from the decision boundary to the sensitive features, which would then cause the attack to target and degrade fairness. In our experiments we set $\lambda = 1$.

Random Anchoring Attack (RAA) The anchoring attack where a target point is picked at random. In this new set of attacks, the goal is to place poisoned points in the vicinity of the target points in which the poisoned and target points have the same demographic group but different labels. In our experiments we set $\tau = 0$ indicating the closest vicinity.

Non-random Anchoring Attack (NRAA) This attack builds upon the random anchoring attack; however, in this attack, the target point is not chosen randomly. In this attack, the point with the most neighbors similar to it (with the same demographic group and label) is chosen as the target point so that we can infect as many similar points to the target point as possible. This can be effective because we are infecting more targeted points; however, in some cases it might be less effective since more poisoned points may be needed in order to achieve the goal of infecting many points and shifting the decision boundary. In our experiments we set $\tau = 0$.

Influence Attack (Koh et al.) This is a type of attack that is targeted only toward affecting accuracy (Koh, Steinhardt,

[2] https://github.com/propublica/compas-analysis

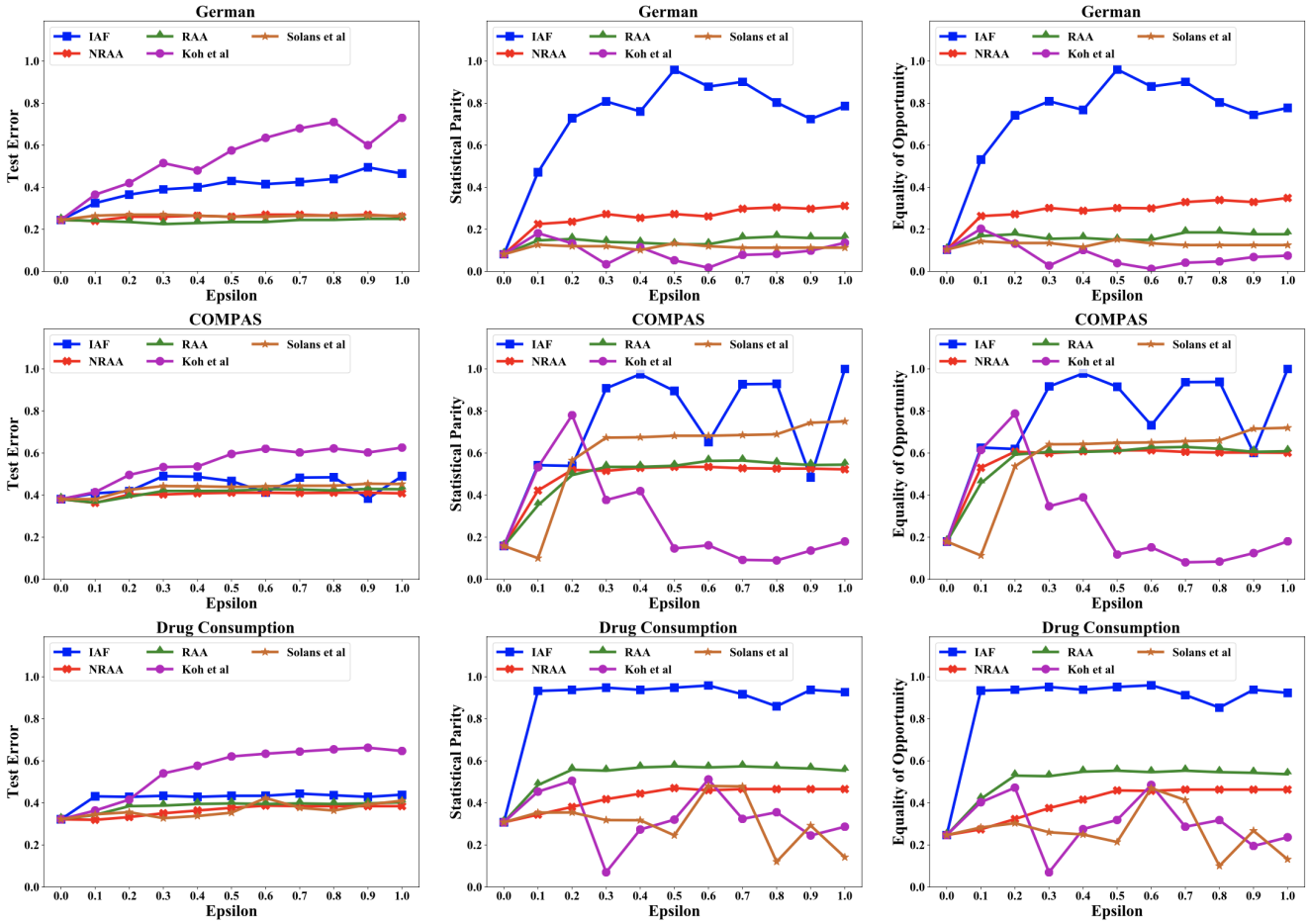

Figure 2: Results obtained for different attacks with regards to different fairness (SPD and EOD) and accuracy (test error) measures on three different datasets (German Credit, COMPAS, and Drug Consumption) with different ϵ values.

and Liang 2018; Koh and Liang 2017). The reason we include this type of attack along with attacks targeted toward fairness is that it can help us understand how attacks targeting only accuracy affect fairness measures. Attacks of this nature can also serve as a good comparison because they show the effect of attacks on accuracy; because this attack is specifically designed to target accuracy, it can be a strong method to compare against.

Poisoning Attack Against Algorithmic Fairness (Solans et al.) In (Solans, Biggio, and Castillo 2020), the authors propose a loss function that claims to target fairness measures. We utilized the loss introduced in this paper as depicted below in equation (3) in the influence attack from (Koh, Steinhardt, and Liang 2018; Koh and Liang 2017) and compared it to our proposed attacks. The goal of (Solans, Biggio, and Castillo 2020) was to incorporate the loss in (3) into an attack strategy that would maximize the loss; thus, we incorporated this loss into the influence attack (Koh, Steinhardt, and Liang 2018; Koh and Liang 2017), which we found to be a strong attack strategy in maximizing the loss and also the same attack strategy used in our influence attack on fairness. In our experiments, we utilized the same λ value as proposed in (Solans, Biggio, and Castillo 2020) to balance the class priors.

$$L_{adv}(\hat{\theta}; \mathcal{D}_{test}) = \underbrace{\sum_{k=1}^{p} \ell(\hat{\theta}; x_k, y_k)}_{\text{disadvantaged}} + \lambda \underbrace{\sum_{j=1}^{m} \ell(\hat{\theta}; x_j, y_j)}_{\text{advantaged}}$$
$$where \;\; \lambda = \frac{p}{m}. \tag{3}$$

Results

The results in Figure 2 demonstrate that the influence attack (Koh et al.), although performing remarkably well in attacking accuracy, does not attack fairness well. The results also confirm that our influence attack on fairness method outperforms (Solans et al.) (Solans, Biggio, and Castillo 2020) in affecting fairness measures, and anchoring attack outperforms (Solans et al.) (Solans, Biggio, and Castillo 2020) in affecting fairness measures in most of the cases. One can observe that influence attack on fairness is the most effective amongst all the attacks in attacking fairness measures.

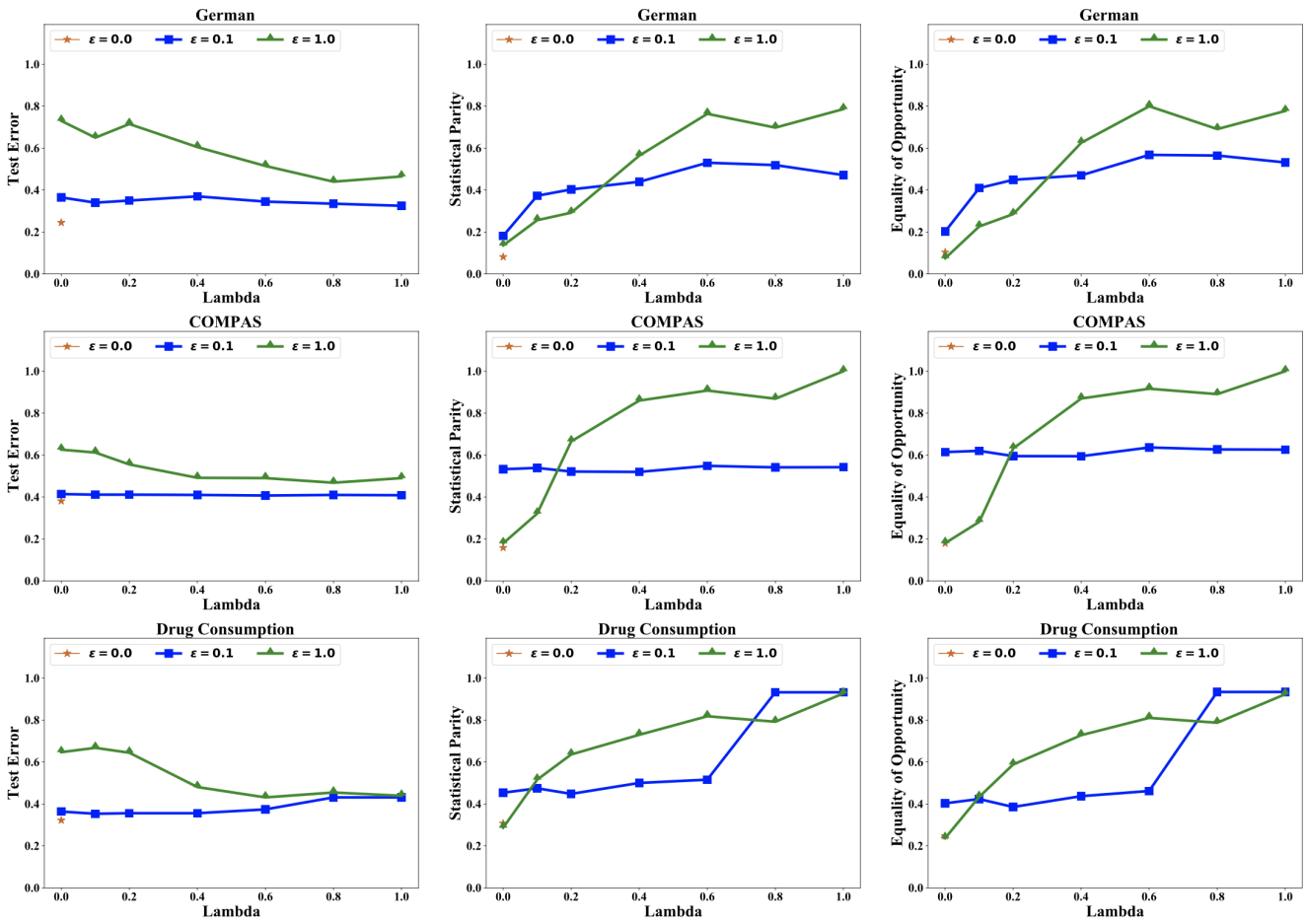

Figure 3: Results obtained for different lambda values for the IAF attack with regards to different fairness (SPD and EOD) and accuracy (test error) measures on three different datasets (German Credit, COMPAS, and Drug Consumption) with different ϵ.

Due to the nature of our influence attack on fairness loss function and its controlling parameter on accuracy and fairness, it can be utilized in scenarios where the adversary wants to maliciously harm the system in terms of accuracy, or fairness, or both. On the other hand, anchoring attacks can be utilized in places where the adversary wants to subtly harm accuracy with an effective harm on fairness. These types of attacks can be used by, e.g., adversaries who would want to gain profit off of biasing decisions for their benefit; thus, to remain less detectable they do not harm accuracy. Although it is possible that anchoring attack can harm accuracy to a higher degree, as shown empirically in our results, it is less likely that anchoring attack is able to degrade accuracy by a large amount in practice for real world datasets.

In addition, in Figure 3 we demonstrate the effect of our regularized loss in the influence attack on fairness. The results show that with the increase of lambda the attack affects fairness measures more as expected from the loss; however, for the lower lambda values the attack acts similar to the original influence attack targeted towards accuracy. The results also show that higher epsilon values highlight the behavior of the loss more as expected such that for high epsilon value of 1 the changes are more significant with modifications to the lambda value in the loss function, while less subtle for lower epsilon values such as 0.1.

Related Work

Here, we cover related work from both fair machine learning as well as adversarial machine learning research.

Adversarial Machine Learning

Research in adversarial machine learning is mostly focused on designing defenses and attacks against machine learning models (Steinhardt, Koh, and Liang 2017; Chakraborty et al. 2018; Li et al. 2018). Ultimately, the goal is for machine learning models to be robust toward malicious activities designed by adversaries. Thus, it is important to consider both sides of the spectrum in terms of designing the attacks and defenses that can overcome the attacks. In adversarial machine learning, different types of attacks, such as data poisoning and evasion attacks, exist. In evasion attacks, the goal is to

come up with adversarial examples that are imperceptible to human eye but can deceive benign machine learning models during test time (Biggio et al. 2013; Moosavi-Dezfooli, Fawzi, and Frossard 2016; Goodfellow, Shlens, and Szegedy 2015). On the other hand, in data poisoning attacks, the goal is to manipulate the training data–via adding, removing, or changing instances–so that the learned model is malicious (Biggio, Nelson, and Laskov 2012; Shafahi et al. 2018). Different algorithms and approaches have been proposed for poisoning attacks focusing on accuracy as the performance measure (Biggio, Nelson, and Laskov 2012; Shafahi et al. 2018). In this paper, we also focused on data poisoning attack while considering fairness as a performance measure in addition to accuracy.

Fair Machine Learning

Research in fair machine learning has gained attention recently, with many active research areas. For instance, some work introduces new definitions and measures for fairness (Dwork et al. 2012; Hardt, Price, and Srebro 2016; Kusner et al. 2017; Mehrabi, Huang, and Morstatter 2020). (Verma and Rubin 2018) has a complete list of the definitions on fairness. Other work utilizes these definitions and tries to design and learn fair classification (Zafar et al. 2015; Ustun, Liu, and Parkes 2019), regression (Agarwal, Dudik, and Wu 2019), and representations (Moyer et al. 2018). The battle to mitigate unfairness can happen in different phases. Some target making the data more fair (Zhang, Wu, and Wu 2017), while others target the algorithms (Zafar et al. 2015). These mitigation techniques can also vary in when and how they are applied. For instance, some approaches are pre-processing techniques (Kamiran and Calders 2012) in which the focus is to remove discrimination from the data before the learning phase. Others try to impose fairness during training via incorporation of fair loss functions or other approaches during the training phase, known as in-processing (Kamishima et al. 2012), while some are post-processing approaches (Pleiss et al. 2017) in which the model is treated as a black box system and discrimination removal is performed on the output of the model. Mehrabi et al. (2019) performs a literature review of fair machine learning research in different subject domains, which can be referenced for more detail. In our work, we utilize some of the definitions and measures widely used in fair machine learning research (Dwork et al. 2012; Hardt, Price, and Srebro 2016) in measuring the performance of our attacks with regard to fairness. We were also inspired by some loss functions introduced in fair classification tasks in one of our attacks (Zafar et al. 2015).

Adversarial Fair Machine Learning

The rapid and significant growth of research in algorithmic fairness highlights the importance of machine learning models being fair and robust toward any unfair behavior. To this end, it is important to think about attacks that can make models unfair in order to strengthen models against such attacks. This recent and interesting line of work combines the two fields of fair and adversarial machine learning. The only work we are aware of that proposes poisoning attacks against algorithmic fairness is (Solans, Biggio, and Castillo 2020). In (Solans, Biggio, and Castillo 2020), the authors propose an attack that targets fairness. We compared this attack with our two newly proposed attacks using three real world datasets. Our anchoring attack does not rely on any loss function making it different in nature with the previous work. In our influence attack on fairness we introduce a new loss function different than the previous work which is more in line with fairness literature and work done in fairness domain making our attack more intuitive. In addition, our influence attack on fairness is able to control a fairness-accuracy trade-off with the hyper-parameter involved in its loss function which is also shown in Figure 3 as an additional experimental result. This line of work can bring researchers from both fields closer and inspire new and interesting research problems. Another interdisciplinary research field combining concepts from fairness and privacy includes the differential privacy line of work (Dwork 2008; Bagdasaryan, Poursaeed, and Shmatikov 2019; Jagielski et al. 2019; Pujol et al. 2020).

Conclusion and Future Work

In this work, we introduced two families of poisoning attacks that can target fairness. We showed the effectiveness of these attacks through experimentation on different real world datasets with different measures. Our influence attack on fairness (IAF) used the attack strategy as in influence attack (Koh, Steinhardt, and Liang 2018; Koh and Liang 2017). As an extension, we modified the loss function so that it can harm fairness as well as accuracy. Furthermore, we explore an attack strategy called the "anchoring attack" that harms fairness by placing poisoned points near target points in order to bias the outcome. Our paper also introduced two ways of sampling these target points. A direct extension of this approach is to explore other methods of sampling points to increase the effectiveness of this attack. The introduced attacks each have their own advantages and disadvantages. The goal was to design attacks that can complement each other. For instance, influence attack on fairness which is gradient-based can be slow. Anchoring attack, however, does not use gradients and is considerably faster. Further, while influence attack on fairness targets fairness harshly, anchoring attack is more subtle. And if anchoring attack can not explicitly control for accuracy fairness trade-off, influence attack on fairness can control this trade-off.

This work points out several important angles for future research. Some important extensions are as follows: what other ways can machine learning systems be harmed by data poisoning attacks? How can we design and adapt defenses that can be effective against malicious attacks targeting fairness? Another question worth pursuing is from the perspective of the defender. Can current defenses against accuracy attacks be useful against the types of attacks that target fairness. If not, how do we adapt defenders so that they can prevent fairness attacks? These questions can help us design more fair, and accurate models that are robust to poisoning attacks. By extension, one can also think about stronger attacks against fairness than ours. We anticipate that continuing to blend the fields of adversarial and fair machine learning can create interdisciplinary ideas that can help us develop more robust and fair machine learning models.

Acknowledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR0011890019. We thank the anonymous reviewers and Mozhdeh Gheini for their constructive feedback.

Ethics Statement

This paper furthers ethics in the machine learning community in two major ways:

- Despite extensive research in adversarial machine learning, not much attention has been given to scenarios where fairness is a possible target of deliberate attacks. We suggest that fairness metrics are as important as accuracy, because they can be manipulated in sensitive environments to achieve malicious goals. Our work points out potential vulnerabilities of machine learning models against fairness-targeting attacks. This line of research can raise awareness, and motivate researchers to introduce methods to mitigate harmful effects of adversarial attacks on fairness. The attacks proposed in this paper are meant to ensure the robustness of fairness in machine learning applications. Nevertheless, we acknowledge that in the wrong hands these type of tools could enable an attacker to harm fairness in extant machine learning systems.
- Fairness and adversarial machine learning are both very important research areas with major safety, security, and ethics implications both within and beyond on the AI/machine learning community. This work combines ideas from both adversarial and fair machine learning, and will hopefully facilitate collaboration among researchers from both communities, eventually leading to more robust and fair machine learning models.

References

Agarwal, A.; Dudik, M.; and Wu, Z. S. 2019. Fair Regression: Quantitative Definitions and Reduction-Based Algorithms. In *International Conference on Machine Learning*, 120–129.

Bagdasaryan, E.; Poursaeed, O.; and Shmatikov, V. 2019. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems*, 15453–15462.

Biggio, B.; Corona, I.; Maiorca, D.; Nelson, B.; Šrndić, N.; Laskov, P.; Giacinto, G.; and Roli, F. 2013. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, 387–402. Springer.

Biggio, B.; Nelson, B.; and Laskov, P. 2012. Poisoning Attacks against Support Vector Machines. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, 1467–1474. Madison, WI, USA: Omnipress. ISBN 9781450312851.

Chakraborty, A.; Alam, M.; Dey, V.; Chattopadhyay, A.; and Mukhopadhyay, D. 2018. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069* .

Dressel, J.; and Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4(1). doi: 10.1126/sciadv.aao5580. URL https://advances.sciencemag.org/content/4/1/eaao5580.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository. URL http://archive.ics.uci.edu/ml.

Dwork, C. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, 1–19. Springer.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Fehrman, E.; Muhammad, A. K.; Mirkes, E. M.; Egan, V.; and Gorban, A. N. 2017. The five factor model of personality and evaluation of drug consumption risk. In *Data Science*, 231–242. Springer.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. URL http://arxiv.org/abs/1412.6572.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 3315–3323.

Jagielski, M.; Kearns, M.; Mao, J.; Oprea, A.; Roth, A.; Malvajerdi, S. S.; and Ullman, J. 2019. Differentially Private Fair Learning. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 3000–3008. Long Beach, California, USA: PMLR. URL http://proceedings.mlr.press/v97/jagielski19a.html.

Kamiran, F.; and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33(1): 1–33. ISSN 0219-3116. doi: 10.1007/s10115-011-0463-8. URL https://doi.org/10.1007/s10115-011-0463-8.

Kamishima, T.; Akaho, S.; Asoh, H.; and Sakuma, J. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 35–50. Springer.

Koh, P. W.; and Liang, P. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1885–1894. JMLR. org.

Koh, P. W.; Steinhardt, J.; and Liang, P. 2018. Stronger data poisoning attacks break data sanitization defenses. *arXiv preprint arXiv:1811.00741* .

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual Fairness. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*, 4066–4076. Curran Associates, Inc. URL http://papers.nips.cc/paper/6995-counterfactual-fairness.pdf.

Li, G.; Zhu, P.; Li, J.; Yang, Z.; Cao, N.; and Chen, Z. 2018. Security matters: A survey on adversarial machine learning. *arXiv preprint arXiv:1810.07339* .

Mehrabi, N.; Huang, Y.; and Morstatter, F. 2020. Statistical Equity: A Fairness Classification Objective. *arXiv preprint arXiv:2005.07293* .

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* .

Moosavi-Dezfooli, S.-M.; Fawzi, A.; and Frossard, P. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2574–2582.

Moyer, D.; Gao, S.; Brekelmans, R.; Galstyan, A.; and Ver Steeg, G. 2018. Invariant representations without adversarial training. In *Advances in Neural Information Processing Systems*, 9084–9093.

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On Fairness and Calibration. In *Advances in Neural Information Processing Systems 30*, 5680–5689. Curran Associates, Inc. URL http://papers.nips.cc/paper/7151-on-fairness-and-calibration.pdf.

Pujol, D.; McKenna, R.; Kuppam, S.; Hay, M.; Machanavajjhala, A.; and Miklau, G. 2020. Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 189–199.

Shafahi, A.; Huang, W. R.; Najibi, M.; Suciu, O.; Studer, C.; Dumitras, T.; and Goldstein, T. 2018. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems*, 6103–6113.

Solans, D.; Biggio, B.; and Castillo, C. 2020. Poisoning Attacks on Algorithmic Fairness. *arXiv preprint arXiv:2004.07401* .

Steinhardt, J.; Koh, P. W. W.; and Liang, P. S. 2017. Certified Defenses for Data Poisoning Attacks. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*, 3517–3529. Curran Associates, Inc. URL http://papers.nips.cc/paper/6943-certified-defenses-for-data-poisoning-attacks.pdf.

Ustun, B.; Liu, Y.; and Parkes, D. 2019. Fairness without Harm: Decoupled Classifiers with Preference Guarantees. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 6373–6382. Long Beach, California, USA: PMLR. URL http://proceedings.mlr.press/v97/ustun19a.html.

Verma, S.; and Rubin, J. 2018. Fairness Definitions Explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, 1–7. New York, NY, USA: Association for Computing Machinery. ISBN 9781450357463. doi:10.1145/3194770.3194776. URL https://doi.org/10.1145/3194770.3194776.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, 1171–1180.

Zafar, M. B.; Valera, I.; Rodriguez, M. G.; and Gummadi, K. P. 2015. Learning fair classifiers. *stat* 1050: 29.

Zhang, L.; Wu, Y.; and Wu, X. 2017. Achieving non-discrimination in data release. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1335–1344. ACM.