Exploiting Fairness to Enhance Sensitive Attributes Reconstruction

Anonymous Author(s)

***Abstract*—In recent years, a growing body of work has emerged on how to learn machine learning models under fairness constraints, often expressed with respect to some *sensitive attributes*. In this work, we consider the setting in which an adversary has black-box access to a target model and show that information about this model's fairness can be exploited by the adversary to enhance his reconstruction of the sensitive attributes of the training data. More precisely, we propose a generic reconstruction correction method, which takes as input an initial guess made by the adversary and corrects it to comply with some user-defined constraints (such as the fairness information) while minimizing the changes in the adversary's guess. The proposed method is agnostic to the type of target model, the fairness-aware learning method as well as the auxiliary knowledge of the adversary. To assess the applicability of our approach, we have conducted a thorough experimental evaluation on two state-of-the-art fair learning methods, using four different fairness metrics with a wide range of tolerances and with three datasets of diverse sizes and sensitive attributes. The experimental results demonstrate the effectiveness of the proposed approach to improve the reconstruction of the sensitive attributes of the training set.**

***Index Terms*—Reconstruction attack, privacy, fairness, machine learning, constraint programming.**

I. Introduction

The growing use of machine learning models in high-stakes decision-making raises several ethical issues such as the risk of discrimination. To address this issue, a growing body of work has emerged on how to learn machine learning models under fairness constraints, often expressed with respect to some *sensitive attributes* [1]–[3]. These *sensitive attributes* correspond to characteristics such as gender, age or race [4], which should not be taken into account in decision-making processes impacting individuals, for legal, ethical, social or philosophical reasons [1]. While fair models usually do not use such sensitive attributes at inference time to avoid disparate treatment [5], they still require access to them at training time [6]. The fact that these models are learnt with the objective to meet specific constraints regarding these sensitive attributes indicates that fair models intrinsically contain information about them.

Another fundamental aspect of responsible machine learning is the protection of privacy. Indeed, machine learning models are often trained on large amounts of personal data. Here, the main challenge is ensuring that these models learn useful generic patterns without leaking private information about individuals. In this context, *inference attacks* [7]–[9] aim at leveraging the output of a computation (*e.g.*, a trained model) to retrieve information regarding its inputs (*e.g.*, a training dataset). Our work belongs to the category of *dataset reconstruction attacks*, in which an adversary tries to recover part of a model's training data [9]. More precisely, we study the setting in which an adversary aims at retrieving the entire column of sensitive attributes of the training set.

Depending on the available *auxiliary knowledge*, several strategies can be adopted by an adversary to reconstruct the sensitive attributes of the training set. The proposed approach is a post-processing method that we coin as *reconstruction correction*, which takes as input an initial reconstruction performed by an adversary, optionally associated with confidence scores for each guess. The reconstruction correction method then minimally updates the adversary's initial guess to satisfy some user-defined constraints. Our work focuses on the scenario in which these are fairness constraints and the adversary leverages the fact that a model is known to be fair to improve his initial reconstruction. Such *fairness information* can for instance be the results of legal requirements, such as the "80 percent rule" for Statistical Parity [10] stated by the US Equal Employment Opportunity Commission (EEOC) [11].

The tensions between fairness and privacy in machine learning have been studied in recent years, mainly through the theoretical [12], [13] and technical [14]–[16] conflicts existing between statistical fairness metrics and Differential Privacy (DP). For instance, it was proved theoretically impossible to learn models under fairness constraints while respecting DP [12], [13]. Furthermore, DP was shown to have unfair effects on the model's performances [14] and it was observed that fairness led to an increased privacy risk [15]. We refer the interested reader to a recent survey [16] summarizing the different causes and consequences of this conflict. Our work takes a different direction but also demonstrates that enforcing statistical fairness can endanger the privacy of sensitive attributes. More precisely, our contributions are as follows:

- We propose a novel reconstruction attack pipeline, in which a *reconstruction correction* is applied as post-processing to an initial adversary's guess to enforce some user-defined constraints (*e.g.*, fairness constraints).
- We show that declarative programming approaches can be applied to implement a generic reconstruction correction. The proposed integer programming model includes statistical fairness constraints but is general enough to also work for a wide range of user-defined constraints.
- We derive an efficient reconstruction correction model with polynomial search space, suitable to formulate any rate constraints (such as statistical fairness constraints).
- We empirically demonstrate the effectiveness of the pro-

posed reconstruction correction method for two fairness-enhancing techniques that intervene at different stages of the learning pipeline, three datasets with diverse characteristics and sensitive attributes, four statistical fairness metrics as well as a wide range of unfairness tolerances.
- We discuss possible countermeasures to mitigate the proposed reconstruction correction method. In particular, we show that even when the fairness information is not disclosed, the adversary can estimate it and that the performance of reconstruction correction remains high.

The outline of the paper is as follows. First, we introduce in Section II the necessary background notions and review the related work on reconstruction attacks. Afterwards, we describe in Section III our proposed reconstruction correction strategy before evaluating its empirical effectiveness in Section IV. Finally, we discuss possible countermeasures in Section V before concluding.

II. Background and Related Work

In this section, we first introduce the considered supervised machine learning setup and the associated notations. Then, we explain how fairness can be quantified in machine learning before reviewing related work on reconstruction attacks.

A. Supervised Machine Learning & Fairness

Let M be the number of *non-sensitive attributes* characterizing an example. For $j \in \{1..M\}$, $\mathcal{X}_j$ denotes the domain of possible values for attribute j, which can be either categorical or numerical, and $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_M$. Similarly, let $\mathcal{S}$ (respectively $\mathcal{Y}$) be the domain of a (categorical) *sensitive attribute* (respectively *label*). Such sensitive attribute corresponds to personal information such as age, gender or race, which should not be used for a decision-making process due to legal, ethical, social or philosophical reasons [1].

$D = (X, S, Y)$ is a dataset drawn from the true (unknown) distribution over $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$. Let N be the number of *examples* (*i.e.*, datapoints) in D, with $e_{i \in \{1..N\}} = (x, s, y) \in \mathcal{X} \times \mathcal{S} \times \mathcal{Y}$. The objective of a supervised machine learning algorithm is to learn a *classifier* $\mathcal{L}(D) = h$ mapping the attributes space to the label space. The explicit use of a sensitive attribute (such as gender, age or race [4]) is usually prohibited by law to avoid *disparate treatment* [5]. Thus, we assume that the sensitive attribute is not used for inference, which means that $h : \mathcal{X} \mapsto \mathcal{Y}$, with $\hat{Y} = h(X)$ being the predictions of the machine learning model. In line with the fairness literature, we consider the task of binary classification in this work: $\mathcal{Y} = \{0, 1\}$. Nonetheless, our framework could easily be extended to non-binary classification provided that fairness constraints are formulated in this more general setting.

To ensure that machine learning algorithms do not reproduce or create *undesirable biases* (*e.g.*, leading to discrimination), different fairness notions have been proposed in the literature [3]. Three main approaches have emerged [20], namely statistical fairness metrics, individual fairness and causal fairness. *Statistical fairness measures* [17] aim at equalizing some value (function of the confusion matrix of a classifier, *e.g.*, true positive rate) between several *protected groups* (usually defined by the sensitive attributes). *Individual fairness* has as rationale that similar examples should be treated similarly [17]. Finally, *causal fairness* approaches analyze the causal relationships between the different attributes and the outcome of a classifier, possibly mitigating those deemed as discriminatory.

Many methods have been proposed in recent years [1]–[3] to produce *fair* models, which can be divided into three categories depending on which step of the machine learning pipeline they intervene [21]. *Pre-processing* methods aim at removing undesired correlations from the training data before applying standard learning techniques on the sanitized data [22] while *in-processing* techniques directly adapt the learning procedure to produce inherently fair models. Finally, *post-processing* techniques [19] modify the outputs of a trained classifier to achieve fairness.

In this paper, we consider the setting in which fairness is expressed using statistical fairness notions. Our framework is agnostic to the type of fairness-enhancing technique used. This means that the step of the machine learning pipeline in which the fairness intervention occurs does not impact our attack as the latter simply relies on the predictions vector of the model along with the fairness information. For our experiments, we consider four metrics widely used in the literature, namely Statistical Parity [17], Predictive Equality [18], Equal Opportunity [19] and Equalized Odds [19]. Table I provides a summary of these statistical fairness metrics, along with the measure being equalized across the different protected groups and the corresponding mathematical expression.

B. Reconstruction Attacks

One fundamental objective in privacy protection is to ensure that the output of a computation over a dataset D cannot be used to retrieve private information about this dataset [23]. Our proposed framework lies in the category of inference attacks [7], [8], which precisely aim at retrieving information regarding the dataset D by only observing the outputs of the computation. In the machine learning field, the computation being performed is usually a learning algorithm whose output is a trained model.

Different types of inference attacks have been proposed against machine learning models [9]. For instance, membership inference attacks [24], [25] try to infer whether individuals whose profiles are known from the adversary were present in the training set of the model. Our proposed inference attack is rather a *reconstruction attack* [7]–[9], sometimes called *model inversion attack*. Inference attacks against machine learning often consider two distinct adversarial settings [8], [9]. In the *black-box setting*, the adversary does not know the actual trained model's parameters and can only query it through an API. In contrast, in the *white-box* setting, the adversary has full knowledge of the model parameters. Between these two scenarios, different *gray-box* settings are possible. Our attack only requires black-box access to the trained fair model and is agnostic to the actual type of the model, the training algorithm and the fairness mitigation procedure.

TABLE I
SUMMARY OF THE CONSIDERED STATISTICAL FAIRNESS METRICS

Ref.	Metric	Equalized Measure	Constraint Expression
[17]	Statistical Parity (SP)	Probability of positive prediction	$\forall s,\ \lvert\mathbb{P}(\hat{y}=1)-\mathbb{P}(\hat{y}=1 \mid s)\rvert \leq \epsilon$
[18]	Predictive Equality (PE)	False Positive Rate	$\forall s,\ \lvert\mathbb{P}(\hat{y}=1 \mid y=0)-\mathbb{P}(\hat{y}=1 \mid s,\ y=0)\rvert \leq \epsilon$
[19]	Equal Opportunity (EO)	True Positive Rate	$\forall s,\ \lvert\mathbb{P}(\hat{y}=1 \mid y=1)-\mathbb{P}(\hat{y}=1 \mid s,\ y=1)\rvert \leq \epsilon$
[19]	Equalized Odds (EOdds)	False Positive Rate and True Positive Rate	Conjunction of Predictive Equality and Equal Opportunity

Reconstruction attacks have been studied in the context of database access mechanisms since the early 2000s. In the considered setup, a database contains records about individuals, which each record being composed of non-private information along with a private bit (one per individual) [7]. The adversary performs queries to a database access mechanism, whose outputs are aggregate and noisy statistics about private bits of individuals in the database. Such reconstruction attacks were introduced and formalized in [23], along with some fundamental reconstruction results based on the adversary's capabilities. An efficient linear program for reconstructing private bits of a database leveraging counting queries was also proposed. This linear program was later improved and extended to handle different query types [26]. The practical effectiveness of the proposed attacks was demonstrated by a large-scale study carried out by the US Census Bureau in 2018 [27] and was part of its motivation to adopt differential privacy for future data releases. The linear reconstruction program was also used successfully to break the Diffix commercial database access mechanism [28]. Pursuing the same goal, another attack [29] exploited Diffix's data-dependent noise (*i.e.*, sticky noise as well as the addition of static and dynamic noise) to infer private attributes of individuals in a dataset.

One fundamental difference between this line of work and ours lies is the nature of the mechanism accessing the private data. In the machine learning (respectively, database access) setup, such mechanism is the learning algorithm (respectively, database access mechanism), and its output is the trained model (respectively, answers to queries). Indeed, database access mechanisms use the private information to compute the answer to each query. On the contrary, in our setup, the training set sensitive attributes are not accessed anymore at inference time, and all the information regarding them is released at once (with the model itself or its predictions). However, our objective is similar to these works: we aim at retrieving a *column* of the dataset by leveraging the output of some computation involving this column (query answers in the previously depicted works, trained fair model in ours).

Other previous works have also tackled reconstruction problems in various settings. For example in the context of online learning, a reconstruction attack was proposed to infer the *updating set* (newly-collected data used to re-train the deployed model) information using a generative adversarial network leveraging the difference between the model before and after its update [30]. In collaborative deep learning, it was also shown that an adversarial server can exploit the collected gradient updates to recover parts of the participants' data [31]. In the pharmacogenetics field, machine learning models are learnt to propose medical treatments specific to a patient's genotype and background. In this sensitive context, a reconstruction attack was proposed, taking advantage of the correlation between the sensitive attributes, the non-sensitive ones, and the output of a trained model. More precisely, the attack takes as input a trained model and some demographic (non-private) information about a patient whose records were used for training and predicts the patient's sensitive attributes [32]. Subsequent work proposed model inversion attacks leveraging confidence values output by several ML models to infer private information about training examples given some information about them [33]. The attack has been shown to be effective against several models and applications, namely decision trees for lifestyle surveys and neural networks for facial recognition. In the white-box setting, an attack was introduced that exploits the structure of an interpretable machine learning model to reconstruct a probabilistic (uncertain) version of a database [34]. While being different both in terms of techniques and objectives, such inference attack still lies in the category of reconstruction attacks. Finally, other works have studied the intended [35] and unintended [36] training data memorization of machine learning models, along with different ways to exploit it in a white-box or black-box setting.

More closely related to this paper are the works of [37] and [38]. On the one side, [37] proposes an attack to infer the sensitive attribute of an example given the model's output for this example. It is the only attack considering the scenario in which the sensitive attribute is not used for inference (what we also assume in this paper). In a nutshell, the adversary trains a machine learning model using a separate *attack set* for which the sensitive attributes are known. This attack roughly corresponds to the baseline adversaries introduced in section IV-A. On the other side, [38] proposes a mechanism whose principle is related to ours, but considers a very particular setup [38]. The fair training process is done in a distributed manner, with a learner wanting to build a fair model on some training dataset for which it does not know the sensitive attributes, and a third-party which owns them. The learner iteratively sends models parameters to the third-party, which then tells him whether the current model is fair. The learner then knows, for an entire set of models, whether they satisfy the fairness constraint or not. Afterwards, he uses Integer Programming techniques to encode this information and perform the reconstruction of the training set sensitive attributes. While the intuition is similar, our work covers a more general setting (with no assumption on the underlying fairness-enhancing method) in a considerably less favourable attack setup (as the adversary only knows that the final model satisfies the fairness constraint).

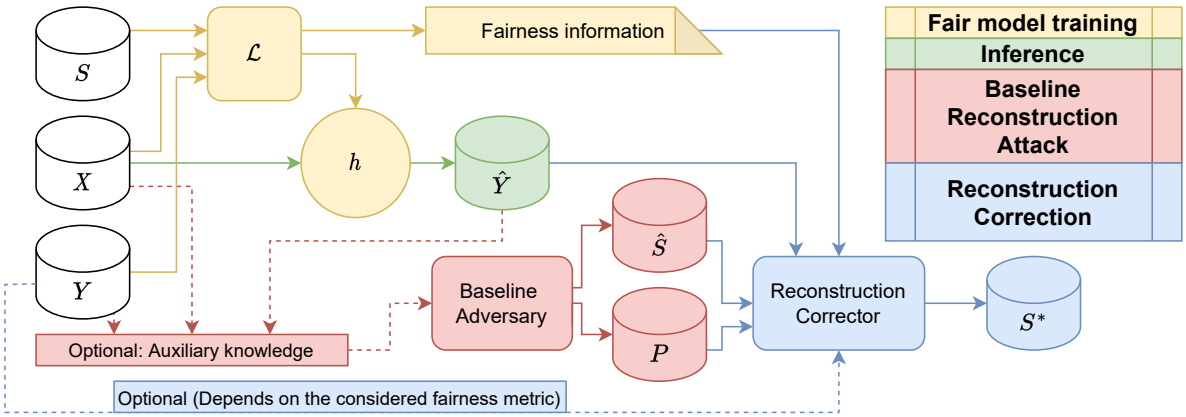

Fig. 1. The proposed attack framework. A model h is learnt by the fair learning procedure $\mathcal{L}$ and used for inference. Then, a *Baseline Adversary* tries to reconstruct the sensitive attributes S of h's training set. Our contribution lies in the *Reconstruction Corrector* component, which takes as input the Baseline Adversary's guess $\hat{S}$ and corrects it to satisfy some fairness property.

III. Leveraging Fairness to Improve Sensitive Attributes Reconstruction

In this section, we first introduce our proposed framework to enhance the reconstruction of sensitive attributes by leveraging the information about the target model's fairness. Afterwards, we describe a general model that can be used to correct any adversary's guess about the sensitive attributes vector, given some knowledge expressed as constraints over this vector. We show how this model can be reformulated to improve scalability in the case of statistical fairness metrics. Finally, we discuss how the proposed models can be generalized to handle other metrics and sensitive attribute values.

A. Attack Pipeline

Fig. 1 illustrates the different components of the considered framework. Given a training dataset $D = (X, S, Y)$, a model h is trained using a fair learning algorithm $\mathcal{L}$, which ensures that h is fair on D according to some statistical fairness metric with respect to the sensitive attribute S. Note that h does not use the sensitive attribute S for inference to prevent disparate treatment [5]. Thus once trained, h can be used for inference based only on non-sensitive attributes X. Our approach does not make any assumption on the underlying fairness-enhancing technique $\mathcal{L}$ used. Indeed, the only requirement of our attack is the knowledge of the fairness information.

The attack itself aims at retrieving the training set sensitive attributes vector S. In the considered pipeline, S is only used by $\mathcal{L}$ to ensure h's fairness (and never used again). In the first step of the attack, a *Baseline Adversary* makes a *guess* $\hat{S}$ on S, based on some auxiliary knowledge. The adversary also outputs a probability vector P, illustrating his confidence for each component of the guess vector $\hat{S}$. Our attack does not assume anything about the form of the auxiliary knowledge. If the adversary does not compute confidence scores, the confidence vector can simply be set to the identity vector.

In the second step of the attack, a *Reconstruction Corrector* component takes as input the baseline adversary's guess and confidence vectors ($\hat{S}$ and P). It outputs a new reconstruction guess S^* minimizing the (confidence-weighted) changes to the adversary's guess while satisfying some given properties, such as statistical fairness constraints. To ensure the respect of such constraints, the *Reconstruction Corrector* component also needs as input the fairness information, the target model's predictions on the training set $\hat{Y}$ as well as (depending on the particular statistical fairness metric at hand, *cf.* Table I) the true labels Y. Importantly, if the actual fairness information is unknown, it can still be estimated as discussed later in Section V-B. As stated previously, our attack does not make any assumptions about the target model h, which can be seen as a black-box as it only requires access to its predictions.

The success of the attack pipeline can be evaluated as the *reconstruction accuracy* of S^* (*i.e.*, proportion of elements of S correctly predicted in S^*). The core contribution of our attack lies in the Reconstruction Corrector component, which, by incorporating solely the fairness information, is able to significantly improve the quality of the reconstruction of the sensitive attribute. Such improvement can be quantified by comparing the reconstruction accuracy of the initial adversary's guess $\hat{S}$ and that of the corrected one S^*.

B. General Reconstruction Correction Model

We now introduce $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$, a general Integer Programming model implementing the Reconstruction Corrector component of Fig. 1, for the binary sensitive attributes setting. Its objective is to modify the adversary's guess for the sensitive attributes of the training examples to satisfy some constraints while minimizing the (confidence-weighted) changes to the adversary's original guess. Here, the constraints implement the fairness information.

a) Inputs:

- $\hat{s}_i \in \{0, 1\},\ i = 1, \dots, N$ (adversary's initial guesses)
- $p_i \in \{0, 1\},\ i = 1, \dots, N$ (adversary's confidence for $\hat{s}_i$)
- $\hat{y}_i \in \{0, 1\},\ i = 1, \dots, N$ (target model h's predictions)
- Fairness information: h satisfies fairness constraints for some metric (*e.g.*, SP) and some tolerance ϵ

b) Decision variables:

- $s_i^* \in \{0,1\},\ i = 1,\dots,N$ (corrected guess for the sensitive attributes vector)

c) Model $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$:

$$\min \sum_{i=1}^{N} (p_i \cdot (1-\hat{s}_i) \cdot s_i^*) + \sum_{i=1}^{N} (p_i \cdot \hat{s}_i \cdot (1-s_i^*)) \quad (1)$$

$$s.t.: \sum_{i=1}^{N} s_i^* > 0 \quad (2)$$

$$\sum_{i=1}^{N} (1-s_i^*) > 0 \quad (3)$$

$$-\epsilon \leq \frac{\sum_{i=1}^{N} \hat{y}_i}{N} - \frac{\sum_{i=1}^{N} \hat{y}_i \cdot s_i^*}{\sum_{i=1}^{N} s_i^*} \leq \epsilon \quad (4)$$

$$-\epsilon \leq \frac{\sum_{i=1}^{N} \hat{y}_i}{N} - \frac{\sum_{i=1}^{N} \hat{y}_i \cdot (1-s_i^*)}{\sum_{i=1}^{N} (1-s_i^*)} \leq \epsilon \quad (5)$$

The objective (1) aims at minimizing the confidence-weighted changes to the original adversary's guess $\hat{S}$. Each modification of a component $\hat{s}_i$ of the original adversary's guess is penalized with cost p_i and the model minimizes the total cost. Constraints (2) and (3) simply ensure that the reconstruction contains at least one example from each protected group. Finally, constraints (4) and (5) encode the fairness constraint for the Statistical Parity metric. Here, constraint (4) (respectively, constraint (5)) ensures that the Positive Prediction Rate (PPR) on group 1 (respectively, group 0) is no further than ϵ from the PPR on the overall dataset.

The key idea here is that fairness is ensured by modifying the reconstruction of the sensitive attributes. This differs from the typical case of fair model training, in which the sensitive attributes are known and fairness is ensured by modifying the model's predictions $\hat{y}_i$ (which, in turn, are fixed here, and exploited to build the sensitive attributes s_i^*).

Finally, an optimal solution to our general reconstruction correction model $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$ is an assignment of the binary variables s_i^* that minimizes (1) while satisfying constraints (2) to (5). This assignment S^* corresponds to the minimum (confidence-weighted) changes to the original adversary guess $\hat{S}$ in order to meet the fairness requirement. If the performed changes are correct most of the time (which is to be expected if the adversary provides good confidence scores), then the overall reconstruction accuracy will be improved. In any case, the algorithm is guaranteed to find a solution satisfying the fairness constraint - which is not the case of the baseline adversary. Indeed, as it is able to modify the sensitive attributes guess of all training examples, the model could actually set any fairness value regarding the sensitive attributes corrected reconstruction. Thus, the knowledge of the exact training unfairness value (rather than a simple upper bound) could easily be used to reduce the set of acceptable reconstructions and enhance the performance of the reconstruction correction. Finally, because it explicitly encodes each training example's sensitive attribute, $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$ can be used to formulate *any* constraint using such attributes.

C. Efficient Model for Statistical Fairness

The search space of the reconstruction correction model $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$ grows exponentially with the number of training examples N. As each element of the sensitive attributes vector S is considered independently from the others (and represented as a binary decision variable), the search space of this model is $O(2^N)$, which limits its scalability. However, when considering statistical fairness metrics, one does not need such granularity. More precisely to satisfy the fairness constraint, the reconstruction corrector may consider exactly four different moves: flipping an element of the reconstructed sensitive attributes $\hat{s}_i$ from 0 to 1 (or the contrary), for an example with prediction $\hat{y} = 1$ (or 0). Then, for the chosen move, the model will always select the example with the lowest confidence score (and then, eventually, the second lower and so on), which drastically reduces the size of the search space as we explain below.

Let n_1^+ be the number of training examples positively predicted by the target model and assigned to group 1 by the initial adversary's guess: $n_1^+ = \sum_{i=1}^{N} \hat{s}_i \cdot \hat{y}_i$. Similarly, let $n_0^+ = \sum_{i=1}^{N} (1-\hat{s}_i) \cdot \hat{y}_i$, $n_1^- = \sum_{i=1}^{N} \hat{s}_i \cdot (1-\hat{y}_i)$, and $n_0^- = \sum_{i=1}^{N} (1-\hat{s}_i) \cdot (1-\hat{y}_i)$. The four numbers n_1^+, n_0^+, n_1^- and n_0^- are the cardinalities of the four groups of examples defining the four possible moves from a fairness perspective. For each group, we sort and cumulate the confidence scores associated to its examples and obtain the following arrays: T_{1^+}, T_{0^+}, T_{1^-} and T_{0^-}. For instance, T_{1^+} contains the confidence scores associated to the n_1^+ training examples positively predicted by the target model and assigned to group 1 by the initial adversary's guess. $T_{1^+}[i]$ is the sum of the i lowest confidence scores among this group. Indeed, $T_{1^+}[i]$ is the exact minimal cost of switching the final reconstruction guess from 1 to 0 for i examples positively predicted by the target model. We use four positive integer decision variables, modeling the number of times each of the four moves is performed to correct the reconstruction. We now define our efficient model for sensitive attributes reconstruction correction: $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$.

a) Inputs:

- Original guesses cardinalities n_1^+, n_0^+, n_1^- and n_0^-.
- Arrays of sorted and cumulated adversary's probabilities for each original guess : T_{1^+}, T_{0^+}, T_{1^-} and T_{0^-}.
- Fairness information: h satisfies fairness constraints for some metric (*e.g.*, SP) and some tolerance ϵ

b) Decision variables:

- $s_{01}^+ \in [0, n_0^+]$: number of changes of $\hat{s}_i$ from 0 to 1, for examples such that $\hat{y}_i = 1$.
- $s_{10}^+ \in [0, n_1^+]$: number of changes of $\hat{s}_i$ from 1 to 0, for examples such that $\hat{y}_i = 1$.
- $s_{01}^- \in [0, n_0^-]$: number of changes of $\hat{s}_i$ from 0 to 1, for examples such that $\hat{y}_i = 0$.
- $s_{10}^- \in [0, n_1^-]$: number of changes of $\hat{s}_i$ from 1 to 0, for examples such that $\hat{y}_i = 0$.

c) Model $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$:

$$\min \; T_{0^+}[s_{01}^+] + T_{1^+}[s_{10}^+] + T_{0^-}[s_{01}^-] + T_{1^-}[s_{10}^-] \tag{6}$$

$$s.t.: \; n_0^+ + n_0^- - s_{01}^+ - s_{01}^- + s_{10}^+ + s_{10}^- > 0 \tag{7}$$

$$n_1^+ + n_1^- - s_{10}^+ - s_{10}^- + s_{01}^+ + s_{01}^- > 0 \tag{8}$$

$$-\epsilon \leq \frac{\sum_{i=1}^{N} \hat{y}_i}{N} - \frac{n_1^+ - s_{10}^+ + s_{01}^+}{n_1^+ + n_1^- - s_{10}^+ - s_{10}^- + s_{01}^+ + s_{01}^-} \leq \epsilon \tag{9}$$

$$-\epsilon \leq \frac{\sum_{i=1}^{N} \hat{y}_i}{N} - \frac{n_0^+ - s_{01}^+ + s_{10}^+}{n_0^+ + n_0^- - s_{01}^+ - s_{01}^- + s_{10}^+ + s_{10}^-} \leq \epsilon \tag{10}$$

Similarly to the general model, the objective (6) minimizes the confidence-weighted sum of the changes. It can be efficiently implemented using `element` constraints within a Constraint Programming (CP) solver. Such constraints are used to access a data array at index given by the value of a variable: $T_{0^+}[s_{01}^+] = \texttt{element}(T_{0^+}, s_{01}^+)$. Furthermore, when minimizing only the number of changes, one could simply sum the four decision variables. The objective then becomes linear as the whole model which can be solved using off-the-shelf Mixed Integer Linear Programming solvers.

Constraints (7) and (8) simply ensure that the reconstruction contains at least one example from each protected group. Finally, constraints (9) and (10) encode the fairness constraint for the Statistical Parity metric. More generally, $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$ could be used to encode any rate constraints on the target model's outputs (using the sensitive attributes), including (but not restricted to) all statistical fairness metrics.

Once the model is solved, optimal assignments of the four decision variables define the (confidence-weighted) minimal number of moves that must be done to ensure fairness. In a post-processing step, the associated moves are performed to the corresponding examples in an increasing order of the confidence scores (so that the overall cost is exactly the objective value (6) of the solved model). This results in the corrected reconstruction vector S^*. One can notice that S^* is also an optimal solution to the general reconstruction correction model $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$. Indeed, as stated in Theorem 1, both models share the same set of optimal solutions, even though their encodings of such solutions differ. The difference is that some non-optimal solutions to the general model $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$ do not correspond to any solution to our efficient model $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$ (*i.e.*, they are simply not part of its search space). Such solutions are all the assignments in which the corrector makes one of the four aforementioned moves but does not select the example with the lowest confidence score (which in this context does not make sense).

Theorem 1 (Equivalence of models): In the context of statistical fairness constraints, the general reconstruction correction model $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$ and the efficient one $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$ share the same set of optimal solutions.

Proof: (a) Any optimal solution to $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$ corresponds to a solution to $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$. Let S^* be an optimal solution to $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$. Then, count the number of performed changes of each type between S^* and $\hat{S}$ (*i.e.*, for an example i with $\hat{y}_i = 0$ (or 1), switching $\hat{s}_i$ from 0 to 1 (or the contrary)). When performing such changes, the solver must have chosen the examples with the lowest confidence scores, or else another solution also satisfies the fairness constraint and has a better objective function value, which contradicts the optimality hypothesis. Afterwards, S^* corresponds to a solution to $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$, represented by the counts for the four moves. Indeed, application of the aforementioned post-processing procedure then allows to retrieve S^*.

(b) Any solution to $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$ corresponds to a solution to $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$. Consider a solution to $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$ and then apply the post-processing step aforementioned. The obtained reconstruction vector is a solution to $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$.

(c) The objective function value of any solution of $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$ is the same in $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$ and $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$. Consider a solution to $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$ with objective value o and apply the aforementioned post-processing step before plugging the resulting reconstruction vector into $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$. By construction, the objective value of this solution of $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$ will be exactly o.

Overall, by *(a)*, *(b)*, and *(c)*, each optimal solution to one of the models is also an optimal solution to the other. ■

Model $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$ uses four variables whose total sum cannot exceed N. Its search space is then $O(N^4)$, which is polynomial in the training set cardinality. Our resolution method also requires some polynomial $O(N \cdot log(N))$ pre-processing and $O(N)$ post-processing computations, which does not modify the overall solving complexity. Overall, for statistical fairness constraints, solving our new model is equivalent to solving the general one, but with polynomial search space instead of exponential one. In practice, this will lead to running times smaller by several orders of magnitude.

D. Generalizing the Reconstruction Correction

The proposed models directly encode the Statistical Parity fairness constraints, but can also be used to correct sensitive attributes reconstructions from all the other metrics of Table I. Recall that the Predictive Equality (PE) metric equalizes the False Positive rates (across the protected groups), which is equivalent to satisfying Statistical Parity over the negatively-labelled subset of the training set. Then, one can simply use the reconstruction correction model on the negatively-labelled subset of the training set. Indeed, PE gives no information on the positively-labelled subset of the training set. Similarly, Equal Opportunity equalizes the True Positive rates, and reconstruction can be achieved using the proposed model on the positively-labelled subset of the training set. Finally, dealing with the Equalized Odds metric can be done by successively applying Predictive Equality and Equal Opportunity reconstruction corrections. Overall, the model proposed for the Statistical Parity metric can actually be used for any of the statistical fairness metrics of Table I, by applying the reconstruction correction on the appropriate data slice.

Observe that even though $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$ is proposed for the binary sensitive attributes setting, it could easily be gen-

eralized by adapting the domains of the s_i^* variables and adding the appropriate cardinalities and fairness constraints for the additional groups. Extending $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$ can also be done by declaring additional variables and constraints. Appendix A depicts how both models can be extended to the general case of multi-valued sensitive attributes, along with a discussion regarding the resulting complexity.

IV. Experiments

In this section, we present our large experimental study regarding the proposed reconstruction framework. We consider a wide range of scenarios using two fair learning algorithms intervening at different stages of the machine learning pipeline, three datasets of various sizes with diverse sensitive attributes, four fairness metrics and a variety of unfairness tolerances. First, we describe our baseline adversaries before detailing the experimental setup and the results obtained.

A. Baseline Adversaries Initial Reconstruction

We instantiate the framework described in Fig. 1 with two different baseline adversaries, $\mathcal{A}$ and $\mathcal{A}'$. Both have access to an auxiliary *attack set*, $D_A = (X_A, S_A, Y_A)$ which is drawn from the same distribution as the actual training set. This attack set models the knowledge of an approximation of the underlying distribution. In line with the reconstruction literature [23], [26], [28], [29], we consider that the dataset contains a *"large amount of nonprivate identifying information and a secret bit, one per individual"* [7]. Here, the private bit of every individual i is its sensitive attribute s_i. Both adversaries hence know the training set non-sensitive attributes vector X and ground truth labels Y (*i.e.*, all dataset columns except the *secret* one, which is the sensitive attribute in our case). Adversary $\mathcal{A}'$ also knows the target model's predictions on the training set $\hat{Y} = h(X)$ and on the attack set $\hat{Y}_A = h(X_A)$.

Adversary $\mathcal{A}$ (respectively $\mathcal{A}'$) relies on the attack set to train an *attack model* to predict S_A from (X_A, Y_A) (respectively from $(X_A, Y_A, \hat{Y}_A)$). Adversary $\mathcal{A}$ (respectively $\mathcal{A}'$) then uses his trained *attack model* to predict S from (X, Y) (respectively from $(X, Y, \hat{Y})$). On the one hand, adversary $\mathcal{A}$ can be used to estimate to what extent general knowledge about the distribution can be used to reconstruct the sensitive attributes of the training set. Indeed, $\mathcal{A}$ does not have any knowledge about the sensitive attributes singularities of the training set, as such sensitive attributes were not used directly or indirectly for any of its inputs. On the other hand, $\mathcal{A}'$ has access to all information that our reconstruction correction will later use, which constitutes the strongest baseline possible to compare against our reconstruction correction. The use of an *attack set* to train an *attack model* is in line with the literature, and $\mathcal{A}'$ corresponds to the adversary proposed in [37].

The attack model performs binary classification, hence its confidence scores lie between 0.5 and 1.0. Using these scores directly to weight our reconstruction correction problem would imply that modifying a prediction with confidence 1.0 (the attacker was certain about it) is better than modifying two predictions with confidence 0.51 (the attacker was unsure). To encourage the reconstruction correction to target the predictions with the lowest scores, we normalize all confidence scores and exponentiate them in order to enlarge their differences. In practice, all the normalized scores are set to the power of k, in which k is chosen to maximize reconstruction correction accuracy on part of the attacker's data used as a validation set. However, other confidence scores processing techniques are possible and may improve the reconstruction correction step. For instance, an adversary could learn how to best discriminate the confidence scores between correct and incorrect predictions on his attack set. Overall, each adversary outputs a guess $\hat{S} = \{\hat{s}_{i \in \{1 \ldots N\}}\}$ for the sensitive attributes vector, along with a confidence vector $P = \{p_{i \in \{1 \ldots N\}}\}$.

B. Setup

1) Datasets: To obtain sufficiently diverse scenarios, we consider three datasets of the fairness literature with different sizes, each with a different binary sensitive attribute. The first one is the UCI Adult Income dataset [39], which gathers records about the 1994 US Census database, with the classification task being to predict whether individuals earn more than $\$50,000$ per year. The considered sensitive attribute is gender (female/male). We also consider two datasets built from the American Community Survey (ACS) Public Use Microdata Sample (PUMS) of the US Census Bureau. More precisely, the datasets are built from data collected in the Texas state in 2018. The second dataset, ACSPublicCoverage [4], contains data about individuals under the age of 65, with an income of less than $\$30,000$, with the classification task being to predict whether they are covered by public health insurance. Here, age is used as the sensitive attribute (younger quartile/others). The third dataset, ACSIncome [4], gathers records about individuals above the age of 16, who reported usual working hours of at least 1 hour per week in the past year, and an income of at least $\$100$. Similar to the original UCI Adult Income dataset, the classification task is to predict whether individuals earn more than $\$50,000$ per year. We rely on the binarized race (white/others) as the sensitive attribute.

Table II summarizes the datasets used in our experiments. For all experiments, each dataset is split between a training set $\left(\frac{1}{3}\right)$, a test set $\left(\frac{1}{3}\right)$ and an attack set $\left(\frac{1}{3}\right)$. The test set is only used to ensure that the fair target model is trained appropriately (in particular, to show that it does not overfit). The attack set is known by the baseline adversary (see Section IV-A).

2) Target Fair Models: To validate our approach, we have tested two off-the-shelf fair learning methods implemented in the `Fairlearn` library [40]: one in-processing method, ExponentiatedGradient [41], as well as a post-processing method, ThresholdOptimizer [19]. In a nutshell, ExponentiatedGradient [41] formulates the fair classification problem as a sequence of cost-sensitive classification problems. Given a cost-sensitive base learner, it follows a two-player game structure in which one player trains the base learner while the other adapts the training examples weights. ThresholdOptimizer [19] takes as input a trained (possibly unfair) classifier and computes group-specific thresholds on the outputs of the classifier to

TABLE II
SUMMARY OF THE DATASETS USED IN OUR EXPERIMENTS

Ref.	Dataset	Binary Prediction Task	#Datapoints	#Non-Sensitive Features	Sensitive Feature
[39]	UCI Adult Income	Income above $50K	45,222	7 categorical, 6 numerical	Gender (Male/Female)
[4]	ACSPublicCoverage*	Coverage from public health insurance	98,928	17 categorical, 1 numerical	Age (First Quartile/Others)
[4]	ACSIncome*	Income above $50K	135,924	7 categorical, 2 numerical	Race Code (White/Other)

*(Texas State, 2018)

adjust its predictions. The thresholds are optimized to enforce some fairness constraints while having minimal impact on classification accuracy. By using two fair learning techniques intervening at different steps of the machine learning pipeline, we want to emphasize that our method is completely agnostic to the type of fairness intervention. Indeed, the only information used by our reconstruction correction strategy is the final fairness information, along with the predictions of the model. For both methods, we use `scikit-learn` [42] Decision Tree classifiers as base learners with the maximum depth being set to 8 and all other parameters left to their default values.

3) Fairness Metrics: We run experiments for the four fairness metrics presented in Table I. Experiments using the ExponentiatedGradient method use 49 different values of the unfairness tolerance ϵ, ranging non-linearly from 0.0 (exact fairness) to 0.20 (loose constraint). The ThresholdOptimizer method modifies the initial model's predictions to approximate 0.0 unfairness, so we cannot vary the unfairness tolerance here.

4) Attack Models: The attack models used by our baseline adversaries are `scikit-learn` [42] Random Forest classifiers, which are known to be resistant to overfitting and generalize well in many situations. This hypothesis class was chosen based on thorough preliminary experiments. To handle sensitive attributes imbalance [37], we use a class-balanced loss. The Random Forest hyperparameters are optimized using the `HyperOpt-Sklearn` framework [43], with a maximum of 100 evaluations for its Tree of Parzen Estimators search algorithm. This setup ensures that the baseline adversary implements a strong baseline and is in line with the literature.

5) Reconstruction Correction: Our efficient reconstruction correction model $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$ (depicted in Section III-C) is implemented and solved using the `IBM ILOG CP Optimizer Version 12.10`[1] via the `DOcplex`[2] Python Modeling API (version 2.21.207) and its default configuration. The number of threads used in CP Optimizer is set to 1 and the optimality tolerance (absolute and relative) is set to 0.0. Indeed, due to the probabilities exponentiation process presented in Section IV-A, some values can be very small and would lie below the solver's default optimality tolerance. Our reconstruction correction method is implemented as a Python module and will be released publicly upon acceptance.

6) Experimental Parameters: We set a one minute timeout for the reconstruction correction step (model creation and solving). It was never reached in practice, and all models were solved to optimality in less than a few seconds (less than one second in average). Each experiment is repeated 100 times, with different seeds for the data split process and the random state of the algorithms. The results are averaged over the 100 runs and the standard deviation is reported. All experiments are run on a computing cluster over a set of homogeneous nodes using Intel Xeon E5-2683 v4 Broadwell @ 2.1GHz CPU.

[1]https://www.ibm.com/analytics/cplex-cp-optimizer

[2]http://ibmdecisionoptimization.github.io/docplex-doc/

C. Results

1) Experiments using the ExponentiatedGradient technique: Results of our experiments using the ExponentiatedGradient [41] method are displayed for the different datasets in Fig. 2, 3 and 4. The training and test performances of the target fair models are shown in Appendix B, and show that they do not overfit. As expected, training accuracy and unfairness both increase when the fairness constraint is relaxed (*i.e.*, ϵ increases). Due to the models' good generalization, test accuracy and unfairness follow the same trends.

The reconstruction accuracy results displayed in Fig. 2, 3 and 4 for the three considered datasets and the four fairness metrics demonstrate the effectiveness of the proposed approach. In this section, we report the results for adversary $\mathcal{A}'$. Results for adversary $\mathcal{A}$, which are provided in Appendix C, are almost perfectly identical and follow the same trends. As the adversary $\mathcal{A}'$ exploits all the information that our reconstruction correction uses, any further improvement in the reconstruction accuracy can only be explained by the semantics of the fairness constraint integrated in our Reconstruction Corrector model. Recall that the reconstruction accuracy is the proportion of training examples e_i for which the sensitive attribute $s_i \in S$ was correctly reconstructed (in the baseline attacker original guess $\hat{s}_i \in \hat{S}$ or in the corrected one $s_i^* \in S^*$).

One can observe that the corrected reconstruction is always more accurate than the adversary's original guess, which means that the changes made by the reconstruction correction model are correct most of the time. Furthermore, the corrected reconstruction accuracy gets better as the fairness constraint becomes tighter (*i.e.*, lower values of the unfairness tolerance ϵ). Indeed, the reconstruction accuracy improvement is related to the amount of bias mitigated by the fair learning technique, which in turn depends on the considered fairness metric, the unfairness tolerance and the original data bias. For tight fairness constraints, we observe reconstruction accuracy absolute improvements up to 0.06, as in the experiments using the Statistical Parity metric on the ACSIncome dataset (Fig. 4, top left). Such improvements are due to the fairness information, which is the only constraint of our correction models.

Recall that the Predictive Equality (respectively Equal Opportunity) metric only applies to the negatively-labelled (re-

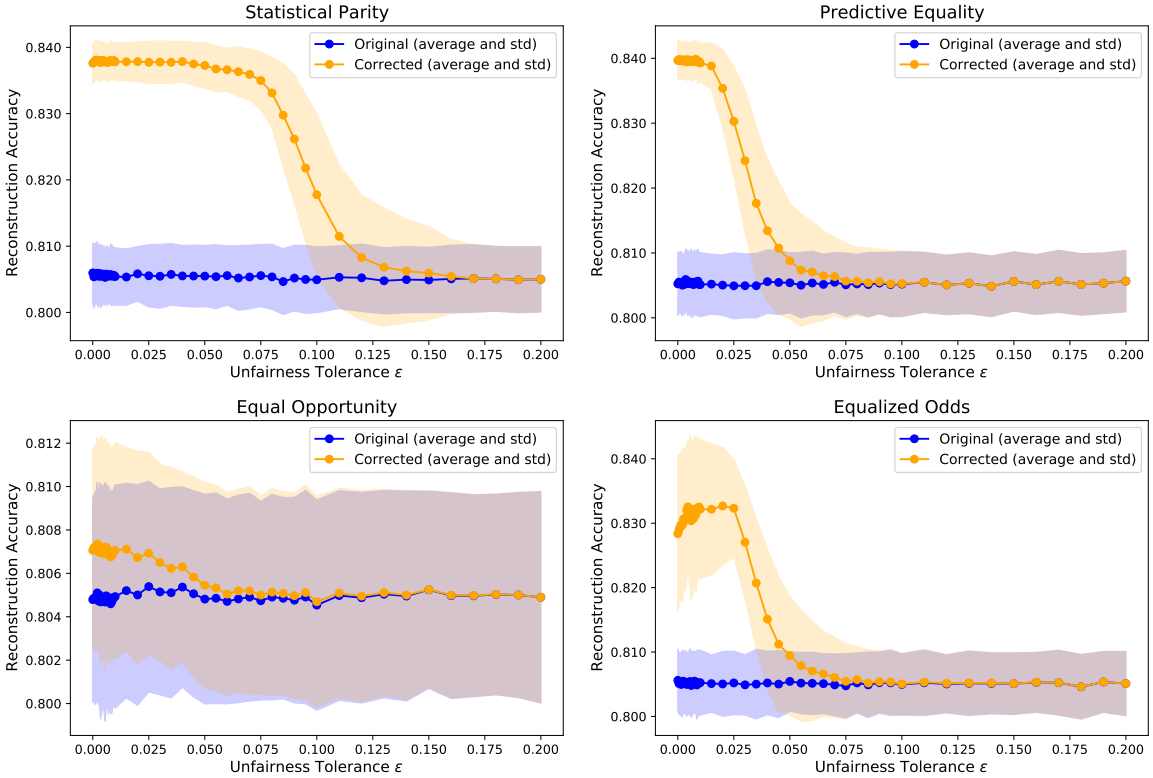

Fig. 2. Corrected and original (adversary $\mathcal{A}'$) reconstruction quality, for our experiments using the UCI Adult Income dataset.

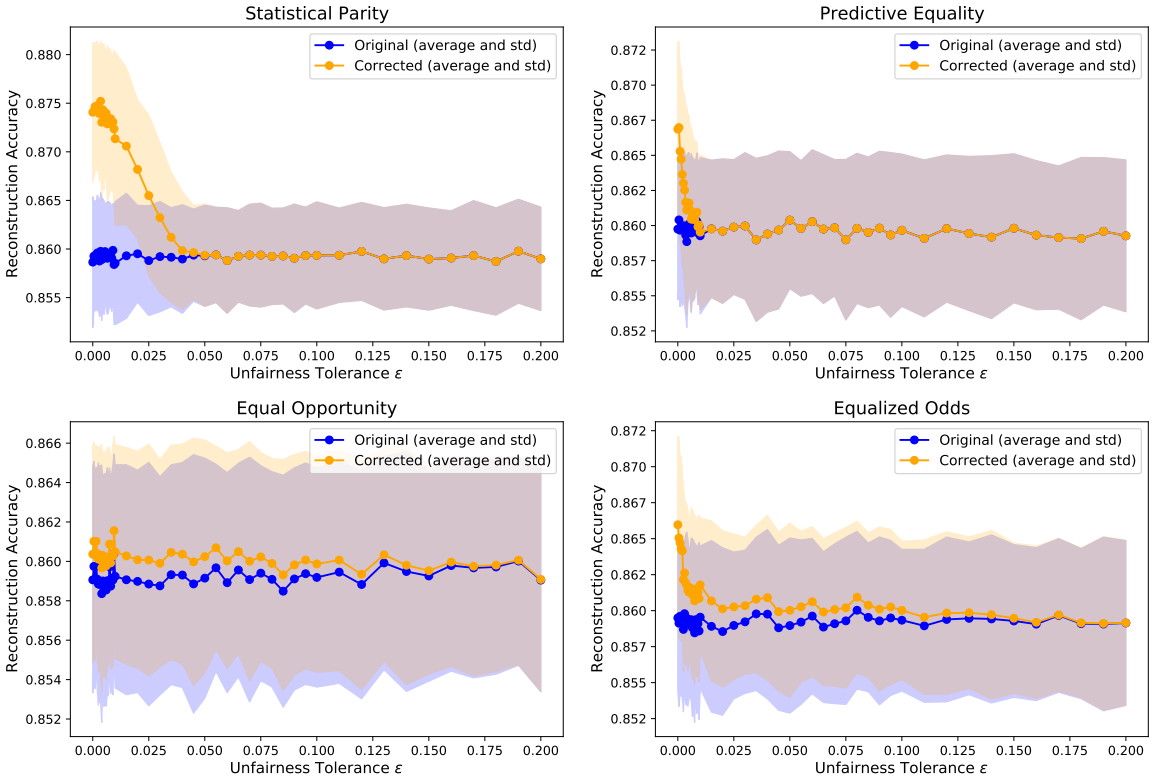

Fig. 3. Corrected and original (adversary $\mathcal{A}'$) reconstruction quality, for our experiments using the ACSPublicCoverage dataset.

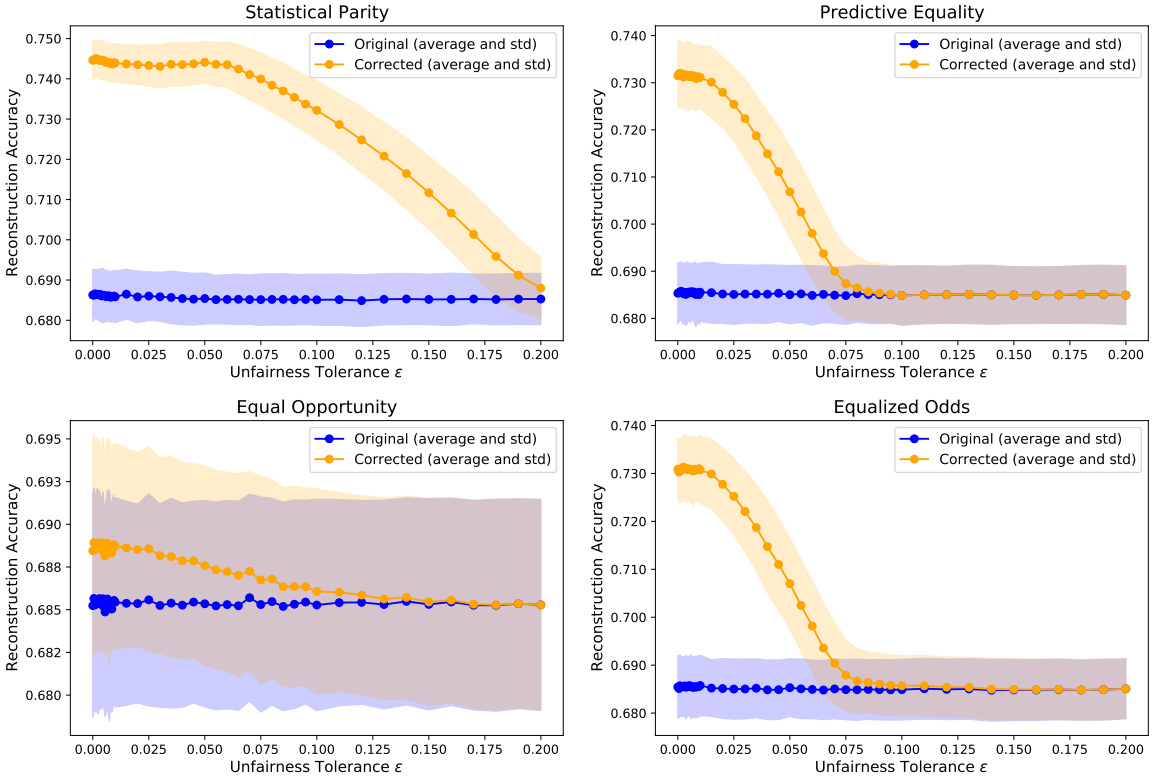

Fig. 4. Corrected and original (adversary $\mathcal{A}'$) reconstruction quality, for our experiments using the ACSIncome dataset

spectively positively-labelled) training examples. This means that such metrics can only help in partially correcting the adversary's guess (as described in Section III-D). Because the datasets used are imbalanced, with the majority of training examples belonging to the negative class, the Equal Opportunity metric relates only to a minority of training examples. As a result, the reconstruction accuracy improvement is more modest than for the remaining metrics. Indeed, even with a close rate of correct modifications, the number of corrections applied (and thus the overall improvement) is smaller.

When varying the unfairness tolerance ϵ, the only input of the reconstruction methods that is modified is the fair model's predictions $\hat{Y}$ (and the fairness information). The fact that the reconstruction accuracy of the baseline adversary $\mathcal{A}'$ is rather constant across variations of ϵ shows that the fair model's predictions $\hat{Y}$ are not used a lot by the learnt attack models. In contrast, as our method knows exactly how to interpret the fairness information with respect to $\hat{Y}$, it is able to exploit it to significantly improve the final reconstruction accuracy.

Finally, the empirical results show that our reconstruction correction method is able to considerably improve the reconstruction accuracy of the training set sensitive attributes, even when the original adversary is as informed as our method.

2) Experiments using the ThresholdOptimizer technique: Results of our experiments using the ThresholdOptimizer [19] fair post-processing method are displayed in Table III. The observed trends are similar to that of the previous subsection, which demonstrates that the type of fairness intervention does not impact our framework. One can observe that the performances of both baseline adversaries are very close. As he possesses more information than $\mathcal{A}$, $\mathcal{A}'$ always performs better on the attack set (used to train the attack models). However, his generalization is sometimes poorer, resulting in worse reconstruction performances when used on the target model training set. This may be due to the distribution of the target fair model's predictions on its own training set $\hat{Y}$ being different from that on the adversary's attack set $\hat{Y}_A$.

Importantly, we observe that the reconstruction correction step always improves the reconstruction accuracy. Indeed, the improvement obtained depends on the considered fairness metric and on the original bias of the reconstruction (which is related to the inherent bias of the original training set). The reconstruction accuracy improvements over the two baseline adversaries are of the same magnitude than with the ExponentiatedGradient method. Again, reconstruction correction using the Equal Opportunity metric offers modest improvements due to the fact that it applies to a minority of training examples.

V. Discussion on Countermeasures

We have seen that the proposed reconstruction correction is able to exploit the fairness information to significantly improve the reconstruction accuracy, even with an informed adversary. In this section, we discuss possible countermeasures to limit the effectiveness of the reconstruction correction step.

TABLE III
SUMMARY OF THE RESULTS OF OUR EXPERIMENTS USING A POST-PROCESSING METHOD FOR FAIRNESS

Metric	Target model (under attack)				Baseline Reconstructions		Corrected Reconstructions	
	Train Acc.	*Test Acc.*	*Train Unf.*	*Test Unf.*	$\mathcal{A}$	$\mathcal{A}'$	$\mathcal{A}$	$\mathcal{A}'$
UCI Adult Income dataset								
SP	0.820 ± 0.008	0.808 ± 0.009	0.003 ± 0.002	0.005 ± 0.003	0.808 ± 0.005	0.814 ± 0.006	0.851 ± 0.003	$\mathbf{0.858 \pm 0.005}$
PE	0.849 ± 0.005	0.836 ± 0.006	0.002 ± 0.001	0.003 ± 0.003	0.808 ± 0.005	0.807 ± 0.005	0.843 ± 0.003	$\mathbf{0.844 \pm 0.004}$
EO	0.857 ± 0.005	0.845 ± 0.005	0.005 ± 0.005	0.041 ± 0.023	0.808 ± 0.005	0.805 ± 0.005	$\mathbf{0.810 \pm 0.005}$	0.807 ± 0.005
EOdds	0.846 ± 0.006	0.834 ± 0.007	0.007 ± 0.006	0.037 ± 0.021	0.808 ± 0.005	0.807 ± 0.004	0.839 ± 0.008	$\mathbf{0.840 \pm 0.009}$
ACSPublicCoverage dataset								
SP	0.861 ± 0.003	0.851 ± 0.003	0.001 ± 0.001	0.003 ± 0.002	0.861 ± 0.005	0.860 ± 0.006	0.874 ± 0.005	$\mathbf{0.875 \pm 0.007}$
PE	0.861 ± 0.002	0.853 ± 0.002	0.001 ± 0.000	0.003 ± 0.002	0.861 ± 0.005	0.860 ± 0.005	0.864 ± 0.005	$\mathbf{0.870 \pm 0.007}$
EO	0.851 ± 0.005	0.843 ± 0.004	0.002 ± 0.002	0.022 ± 0.011	0.861 ± 0.005	0.859 ± 0.006	$\mathbf{0.862 \pm 0.004}$	0.861 ± 0.006
EOdds	0.841 ± 0.004	0.833 ± 0.004	0.003 ± 0.002	0.023 ± 0.011	0.861 ± 0.005	0.860 ± 0.005	$\mathbf{0.862 \pm 0.004}$	0.861 ± 0.005
ACSIncome dataset								
SP	0.788 ± 0.003	0.776 ± 0.003	0.002 ± 0.001	0.005 ± 0.004	0.690 ± 0.007	0.715 ± 0.010	0.756 ± 0.005	$\mathbf{0.764 \pm 0.006}$
PE	0.797 ± 0.002	0.785 ± 0.002	0.001 ± 0.001	0.004 ± 0.003	0.690 ± 0.007	0.688 ± 0.007	$\mathbf{0.736 \pm 0.007}$	0.735 ± 0.006
EO	0.796 ± 0.003	0.784 ± 0.003	0.001 ± 0.001	0.010 ± 0.007	0.690 ± 0.007	0.685 ± 0.006	$\mathbf{0.693 \pm 0.007}$	0.689 ± 0.006
EOdds	0.795 ± 0.003	0.783 ± 0.003	0.002 ± 0.001	0.010 ± 0.006	0.690 ± 0.007	0.688 ± 0.007	$\mathbf{0.737 \pm 0.007}$	0.735 ± 0.006

A. Differential Privacy

Differential Privacy (DP) [44], [45] is considered to be one of the state-of-the-art methods for preventing inference attacks against machine learning models. While it may affect the performances of a baseline adversary, DP cannot be an effective countermeasure to our proposed reconstruction correction step. Indeed, it is designed to ensure that the output of a mechanism does not rely too much on any single example, but rather on general patterns. However, statistical fairness metrics are measured over an entire dataset and do not specifically rely on individual examples. Thus, as our reconstruction correction method only relies on group-level statistics, DP cannot effectively affect its performances [46].

Additionally, DP is incompatible with the strict respect of any statistical fairness measure [12], [13]. Indeed, releasing a model along with information regarding its strict respect of any statistical fairness constraint is intrinsically non-DP compliant.

B. Hiding the Fairness Information

Intuitive countermeasures consist in perturbing the fairness information (type of fairness metric used or unfairness tolerance parameter ϵ). Note that this may not be possible when a particular fairness requirement is also a legal requirement, as for the "80 percent rule" for Statistical Parity [10] stated by the US Equal Employment Opportunity Commission (EEOC) [11]. When possible, releasing noisy or empty fairness information may be a reasonable defense mechanism. However, adversaries may still use diverse strategies to infer both the fairness metric that was optimized and the unfairness tolerance parameter. Depending on the adversarial knowledge, such property inference attacks [9] might give a good estimation to the adversary, which we can expect would still allow reasonable reconstruction correction performances from our approach. Indeed, recall that our proposed method only needs information regarding the model's predictions fairness and the reconstruction correction still works even if the set fairness constraint is not the one that was used for training.

Using our baseline adversaries $\mathcal{A}$ or $\mathcal{A}'$, a simple strategy would be to quantify the target model unfairness on the attack set D_A for the different considered metrics. Then, one can select the metric with the smallest measured unfairness, and consider that the model is fair for this metric with unfairness tolerance ϵ equal to the measured unfairness. To assess its effectiveness, we implemented this fairness information estimation strategy and performed our experiments again.

TABLE IV
SUMMARY OF THE RESULTS OF OUR EXPERIMENTS USING A POST-PROCESSING METHOD FOR FAIRNESS, FOR THE SIMPLE COUNTERMEASURE OF NOT REVEALING THE FAIRNESS INFORMATION.
Reconstruction results have to be compared with those of Table III

Metric	Estimated Constraint		Corrected Reconstr. (Estimated Constraint)	
	Metric Detect.	*Average Tolerance*	$\mathcal{A}$	$\mathcal{A}'$
UCI Adult Income dataset				
SP	0.95	0.004 ± 0.003	0.848 ± 0.009	0.856 ± 0.011
PE	0.97	0.003 ± 0.002	0.841 ± 0.006	0.843 ± 0.007
EO	0.26	0.018 ± 0.010	0.829 ± 0.012	0.828 ± 0.013
EOdds	0.00	0.005 ± 0.005	0.841 ± 0.006	0.843 ± 0.007
ACSPublicCoverage dataset				
SP	1.00	0.002 ± 0.002	0.873 ± 0.005	0.873 ± 0.009
PE	1.00	0.003 ± 0.002	0.863 ± 0.005	0.865 ± 0.007
EO	0.28	0.008 ± 0.005	0.862 ± 0.005	0.862 ± 0.005
EOdds	0.00	0.002 ± 0.002	0.868 ± 0.006	0.869 ± 0.007
ACSIncome dataset				
SP	0.80	0.003 ± 0.003	0.743 ± 0.026	0.754 ± 0.020
PE	0.86	0.003 ± 0.003	0.729 ± 0.016	0.728 ± 0.016
EO	0.73	0.008 ± 0.006	0.704 ± 0.019	0.700 ± 0.020
EOdds	0.00	0.002 ± 0.002	0.723 ± 0.021	0.721 ± 0.022

Results for the experiments using the ThresholdOptimizer [19] method are reported in Table IV. More precisely, we report the performances of the fairness constraint estimation process, namely the rate of correct metric identification, and the average unfairness tolerance inferred. Due to the simple estimation process, the Equalized Odds metric can never be identified as its violation is the maximum of the Predictive Equality and Equal Opportunity violations (hence it can never be the smallest value). However, for the other metrics we observe that even this simple estimation process is often able to correctly identify the optimized metric.

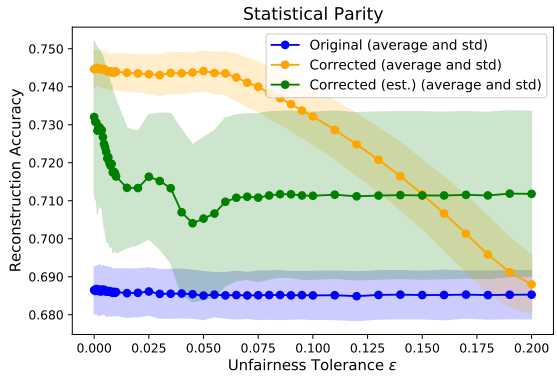

Fig. 5. Original (adversary $\mathcal{A}'$), corrected (from actual fairness constraint, and from estimated one (`est.`)) reconstruction quality, for experiments using the ACSIncome dataset

Several trends can be noted when comparing the reconstruction results with those of Table III, in which the reconstruction correction is done using the actual fairness constraint. A first situation occurs when the fairness constraint is correctly inferred, which is the case in most experiments using the Statistical Parity or Predictive Equality metrics. For instance, when using the ACSIncome dataset, the Statistical Parity metric was correctly identified in all our experiments. In this scenario, the reconstruction correction still brings important improvement - slightly weakened by the fact that the estimated tolerance is usually not as tight as the actual one. A second interesting situation is when the fairness metric is not correctly identified, which is the case for all experiments using the Equalized Odds metric. Nonetheless, the fairness information estimation process can still come with a valid fairness constraint (even if it is not the one that was optimized during training), which can effectively be leveraged by the reconstruction correction step. When the fairness estimation proposes a metric more informative (in terms of number of involved examples) than the actual one (*e.g.*, for some experiments with the Equal Opportunity metric), the reconstruction improvement can sometimes be better than with the original constraint. For instance, consider the experiment using the UCI Adult Income dataset with the Equal Opportunity metric. In 74% of the runs, the fairness constraint estimation process came up with a Predictive Equality constraint. Even though this is not the actual constraint that was optimized during training, this constraint is approximately valid and the corresponding metric relates to a greater number of examples. As a consequence and somewhat counter-intuitively, the final reconstruction is better than with the actual constraint (see Table III). Finally, one important drawback of the fairness estimation process is that the performances of the reconstruction correction step are more variable as shown by greater standard deviation values.

Results using the ExponentiatedGradient method [41] are provided in Fig. 5 for the experiment using the ACSIncome dataset with the Statistical Parity metric and baseline adversary $\mathcal{A}'$, and in Appendix D for the remaining ones. They show similar trends as those using the ThresholdOptimizer method: estimating the fairness constraint still allows for good reconstruction correction performances but leads to a greater variability in the final reconstruction accuracy. Here also, inferring a fairness constraint different from the actual one can improve reconstruction correction, especially when the original tolerance is larger than the actual bias contained in the data (*i.e.*, large values of ϵ). In such cases, the adversary's baseline reconstruction already meets the actual fairness requirement and the reconstruction correction process cannot improve it. In contrast, the fairness constraint estimation process can infer a tighter value, allowing some reconstruction improvement.

Overall, we see that the knowledge of the actual fairness constraint is not necessary as estimations can provide comparable-quality reconstruction correction performances. Using the proposed fairness constraint estimation process, we provide in Appendix E additional reconstruction experiments using a pre-processing method for enhancing fairness. Results demonstrate the effectiveness of the proposed reconstruction correction approach, even when fairness metrics are not directly optimized and no fairness information is available.

VI. Conclusion

In this work, we have proposed a novel approach using declarative programming to improve the reconstruction performances of any baseline adversary by incorporating user-defined constraints. While the general problem may be computationally challenging, we have demonstrated that in the case of statistical fairness metrics (and, more generally, group-level constraints), it can be reformulated and solved efficiently. In addition, our thorough experimental study shows that due to the use of the sensitive attribute information to ensure fairness of the built model, fairness-enhancing learning techniques inherently leak information about it. Even if such information is at the group level, it can be leveraged by an adversary to improve baseline reconstructions of the sensitive attributes. Furthermore, the tighter the fairness requirement, the more significant the reconstruction improvement.

We additionally observed that, even if the fairness information is not available, an adversary can still try to infer it, and obtain good (and sometimes, even better) reconstruction correction performances. While the fairness information is simply an input of our proposed reconstruction correction component, this finding demonstrates the applicability of our approach. It also illustrates the fact that due to their use of the sensitive attributes information, statistical fairness metrics intrinsically conflict with protecting the privacy of such attributes.

Future work includes combining our reconstruction correction attack with different baseline adversaries, optimizing the adversary confidence vector P processing as well as applying our framework in the wider context of non-binary sensitive attributes. One of the key points of our framework is the declarative nature of the reconstruction correction step, which allows considering a wide range of constraints. Extending our proposed pipeline to improve baseline reconstruction attacks by enforcing other constraints (*e.g.*, proportion constraints, rate constraints, ...) is also an interesting research direction.

References

[1] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2019, http://www.fairmlbook.org.

[2] S. Caton and C. Haas, "Fairness in machine learning: A survey," *arXiv preprint arXiv:2010.04053*, 2020.

[3] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 115:1–115:35, 2021. [Online]. Available: https://doi.org/10.1145/3457607

[4] F. Ding, M. Hardt, J. Miller, and L. Schmidt, "Retiring adult: New datasets for fair machine learning," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 6478–6490.

[5] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California Law Review*, vol. 104, no. 3, pp. 671–732, 2016. [Online]. Available: http://www.jstor.org/stable/24758720

[6] I. Zliobaite and B. Custers, "Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models," *Artif. Intell. Law*, vol. 24, no. 2, pp. 183–201, 2016. [Online]. Available: https://doi.org/10.1007/s10506-016-9182-5

[7] C. Dwork, A. Smith, T. Steinke, and J. Ullman, "Exposed! a survey of attacks on private data," *Annual Review of Statistics and Its Application*, vol. 4, no. 1, pp. 61–84, 2017. [Online]. Available: https://doi.org/10.1146/annurev-statistics-060116-054123

[8] M. Rigaki and S. Garcia, "A survey of privacy attacks in machine learning," *CoRR*, vol. abs/2007.07646, 2020. [Online]. Available: https://arxiv.org/abs/2007.07646

[9] E. D. Cristofaro, "An overview of privacy in machine learning," *CoRR*, vol. abs/2005.08679, 2020. [Online]. Available: https://arxiv.org/abs/2005.08679

[10] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, L. Cao, C. Zhang, T. Joachims, G. I. Webb, D. D. Margineantu, and G. Williams, Eds. ACM, 2015, pp. 259–268. [Online]. Available: https://doi.org/10.1145/2783258.2783311

[11] T. U. EEOC., "Uniform guidelines on employee selection procedures," March 2, 1979.

[12] R. Cummings, V. Gupta, D. Kimpara, and J. Morgenstern, "On the compatibility of privacy and fairness," in *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, ser. UMAP'19 Adjunct. New York, NY, USA: Association for Computing Machinery, 2019, p. 309–315. [Online]. Available: https://doi.org/10.1145/3314183.3323847

[13] S. Agarwal, "Trade-offs between fairness and privacy in machine learning," in *IJCAI 2021 Workshop on AI for Social Good*, 2021.

[14] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov, "Differential privacy has disparate impact on model accuracy," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 15 453–15 462.

[15] H. Chang and R. Shokri, "On the privacy risks of algorithmic fairness," in *IEEE European Symposium on Security and Privacy, EuroS&P 2021, Vienna, Austria, September 6-10, 2021*. IEEE, 2021, pp. 292–303. [Online]. Available: https://doi.org/10.1109/EuroSP51992.2021.00028

[16] F. Fioretto, C. Tran, P. V. Hentenryck, and K. Zhu, "Differential privacy and fairness in decisions and learning tasks: A survey," *CoRR*, vol. abs/2202.08187, 2022. [Online]. Available: https://arxiv.org/abs/2202.08187

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

[18] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[19] M. Hardt, E. Price, N. Srebro *et al.*, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, 2016, pp. 3315–3323.

[20] S. Verma and J. Rubin, "Fairness definitions explained," in *Proceedings of the International Workshop on Software Fairness*, ser. FairWare '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1–7. [Online]. Available: https://doi.org/10.1145/3194770.3194776

[21] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. T. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *CoRR*, vol. abs/1810.01943, 2018. [Online]. Available: http://arxiv.org/abs/1810.01943

[22] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 2012.

[23] I. Dinur and K. Nissim, "Revealing information while preserving privacy," in *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 9-12, 2003, San Diego, CA, USA*, F. Neven, C. Beeri, and T. Milo, Eds. ACM, 2003, pp. 202–210. [Online]. Available: https://doi.org/10.1145/773153.773173

[24] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*. IEEE Computer Society, 2017, pp. 3–18. [Online]. Available: https://doi.org/10.1109/SP.2017.41

[25] H. Hu, Z. Salcic, G. Dobbie, and X. Zhang, "Membership inference attacks on machine learning: A survey," *CoRR*, vol. abs/2103.07853, 2021. [Online]. Available: https://arxiv.org/abs/2103.07853

[26] C. Dwork, F. McSherry, and K. Talwar, "The price of privacy and the limits of lp decoding," in *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, ser. STOC '07. New York, NY, USA: Association for Computing Machinery, 2007, p. 85–94. [Online]. Available: https://doi.org/10.1145/1250790.1250804

[27] S. Garfinkel, J. M. Abowd, and C. Martindale, "Understanding database reconstruction attacks on public data: These attacks on statistical databases are no longer a theoretical danger." *Queue*, vol. 16, no. 5, p. 28–53, oct 2018. [Online]. Available: https://doi.org/10.1145/3291276.3295691

[28] A. Cohen and K. Nissim, "Linear program reconstruction in practice," *J. Priv. Confidentiality*, vol. 10, no. 1, 2020. [Online]. Available: https://doi.org/10.29012/jpc.711

[29] A. Gadotti, F. Houssiau, L. Rocher, B. Livshits, and Y. de Montjoye, "When the signal is in the noise: Exploiting diffix's sticky noise," in *28th USENIX Security Symposium, USENIX Security 2019, Santa Clara, CA, USA, August 14-16, 2019*, N. Heninger and P. Traynor, Eds. USENIX Association, 2019, pp. 1081–1098. [Online]. Available: https://www.usenix.org/conference/usenixsecurity19/presentation/gadotti

[30] A. Salem, A. Bhattacharya, M. Backes, M. Fritz, and Y. Zhang, "Updates-leak: Data set inference and reconstruction attacks in online learning," in *29th USENIX Security Symposium, USENIX Security 2020, August 12-14, 2020*, S. Capkun and F. Roesner, Eds. USENIX Association, 2020, pp. 1291–1308. [Online]. Available: https://www.usenix.org/conference/usenixsecurity20/presentation/salem

[31] L. T. Phong, Y. Aono, T. Hayashi, L. Wang, and S. Moriai, "Privacy-preserving deep learning: Revisited and enhanced," in *Applications and Techniques in Information Security - 8th International Conference, ATIS 2017, Auckland, New Zealand, July 6-7, 2017, Proceedings*, ser. Communications in Computer and Information Science, L. Batten, D. S. Kim, X. Zhang, and G. Li, Eds., vol. 719. Springer, 2017, pp. 100–110. [Online]. Available: https://doi.org/10.1007/978-981-10-5421-1_9

[32] M. Fredrikson, E. Lantz, S. Jha, S. M. Lin, D. Page, and T. Ristenpart, "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing," in *Proceedings of the 23rd USENIX Security Symposium, San Diego, CA, USA, August 20-22, 2014*, K. Fu and J. Jung, Eds. USENIX Association, 2014, pp. 17–32. [Online]. Available: https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/fredrikson_matthew

[33] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, October 12-16, 2015*, I. Ray, N. Li, and C. Kruegel, Eds. ACM, 2015, pp. 1322–1333. [Online]. Available: https://doi.org/10.1145/2810103.2813677

[34] S. Gambs, A. Gmati, and M. Hurfin, "Reconstruction attack through classifier analysis," in *Data and Applications Security and Privacy XXVI - 26th Annual IFIP WG 11.3 Conference, DBSec 2012, Paris, France, July 11-13,2012. Proceedings*, ser. Lecture Notes in Computer Science, N. Cuppens-Boulahia, F. Cuppens, and J. García-Alfaro, Eds., vol. 7371. Springer, 2012, pp. 274–281. [Online]. Available: https://doi.org/10.1007/978-3-642-31540-4_21

[35] C. Song, T. Ristenpart, and V. Shmatikov, "Machine learning models that remember too much," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, Dallas, TX, USA, October 30 - November 03, 2017*, B. M. Thuraisingham, D. Evans, T. Malkin, and D. Xu, Eds. ACM, 2017, pp. 587–601. [Online]. Available: https://doi.org/10.1145/3133956.3134077

[36] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The secret sharer: Evaluating and testing unintended memorization in neural networks," in *28th USENIX Security Symposium, USENIX Security 2019, Santa Clara, CA, USA, August 14-16, 2019*, N. Heninger and P. Traynor, Eds. USENIX Association, 2019, pp. 267–284. [Online]. Available: https://www.usenix.org/conference/usenixsecurity19/presentation/carlini

[37] J. Aalmoes, V. Duddu, and A. Boutet, "Dikaios: Privacy auditing of algorithmic fairness via attribute inference attacks," *arXiv preprint arXiv:2202.02242*, 2022.

[38] H. Hu and C. Lan, "Inference attack and defense on the distributed private fair learning framework," in *The AAAI Workshop on Privacy-Preserving Artificial Intelligence*, 2020.

[39] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[40] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker, "Fairlearn: A toolkit for assessing and improving fairness in AI," Microsoft, Tech. Rep. MSR-TR-2020-32, May 2020. [Online]. Available: https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[41] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. M. Wallach, "A reductions approach to fair classification," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 60–69. [Online]. Available: http://proceedings.mlr.press/v80/agarwal18a.html

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[43] B. Komer, J. Bergstra, and C. Eliasmith, "Hyperopt-sklearn: automatic hyperparameter configuration for scikit-learn," in *ICML workshop on AutoML*, vol. 9. Citeseer, 2014, p. 50.

[44] C. Dwork, "Differential privacy: A survey of results," in *Theory and Applications of Models of Computation*, M. Agrawal, D. Du, Z. Duan, and A. Li, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 1–19.

[45] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3–4, p. 211–407, aug 2014. [Online]. Available: https://doi.org/10.1561/0400000042

[46] G. Cormode, "Individual privacy vs population privacy: Learning to attack anonymization," *CoRR*, vol. abs/1011.2511, 2010. [Online]. Available: http://arxiv.org/abs/1011.2511

Appendix A
Reconstruction Correction Models for Multi-Valued Sensitive Attributes

In this appendix section, we discuss the most general setting in which the sensitive attribute is multi-valued and takes one of $|\mathcal{S}|$ values (hence effectively defining $|\mathcal{S}|$ protected groups). We explain how both models can be extended to handle multi-valued sensitive attributes reconstruction and discuss the complexity cost induced by this extension. We begin with the general reconstruction correction model, which is suitable to encode any constraint on the protected attributes. We then treat the efficient model, which can be used to encode any rate constraints on the protected attributes (such as, but not restricted to, statistical fairness constraints).

A. General Reconstruction Correction Model

The general reconstruction correction model $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$ uses exactly one decision variable to encode each training example's sensitive attribute. Extension to the general multi-valued sensitive attributes case hence requires modifying the domains of such variables to match that of the sensitive attributes (with $|\mathcal{S}|$ different possible values). The N decision variables now have domain of cardinality $|\mathcal{S}|$. The objective function sums the (weighted) changes in the adversary's sensitive attributes guess, as was done in the binary case in (1). $|\mathcal{S}|$ constraints ensure that there is at least one example from each protected group (as was done with (2) and (3) for the binary sensitive attribute setting). Finally, one fairness constraint is declared for each protected group (sensitive attribute value), ensuring that its positive prediction rate is no further than ϵ from that of the entire dataset (as was done with (4) and (5) for the binary sensitive attribute setting).

Overall, the size of the search space of $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$ is $O(|\mathcal{S}|^N)$, which generalizes the binary sensitive attribute case for which it was $O(2^N)$.

B. Efficient Model for Statistical Fairness

The efficient reconstruction correction model $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$ uses one decision variable to count the number of changes from one sensitive attribute value to another, for each pair of sensitive attributes values. Extension to the general multi-valued sensitive attributes case hence requires declaring $O(|\mathcal{S}|^2)$ variables. To ensure that each example is counted only once, $O(|\mathcal{S}|)$ constraints must be declared. Furthermore, to quantify the total cost of the performed changes, $O(|\mathcal{S}|^2)$ `element` constraints have to be summed in the objective function, as was performed in (6) in the binary sensitive attributes case. $|\mathcal{S}|$ constraints ensure that there is at least one example from each protected group (as was done with (7) and (8) for the binary sensitive attribute setting). Finally, one fairness constraint is declared for each protected group (sensitive attribute value), ensuring that its positive prediction rate is no further than ϵ from that of the entire dataset (as was done with (9) and (10) for the binary sensitive attribute setting).

Overall, the size of the search space of $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$ is $O(N^{|\mathcal{S}|^2})$, which generalizes the binary sensitive attribute case for which it was $O(N^4)$.

Appendix B
Detailed Results: Target Models Performances

In this section, we provide the training and test performances (accuracy and unfairness) of the (target) trained fair models for our experiments using the ExponentiatedGradient [41] method (Section IV-C1). Fig. 6, 7 and 8 display these results for our experiments on the three datasets, for the four fairness metrics.

As expected, we observe that both accuracy and unfairness decrease when the fairness constraint is tightened (ϵ diminishes). The trained fair models generalize rather well and so similar trends are observed on the test sets.

Appendix C
Detailed Results: Reconstruction Accuracy using Baseline Adversary $\mathcal{A}$

In this section, we provide the experimental results (reconstruction accuracy) for the experiments using the baseline adversary $\mathcal{A}$, for a target model trained using the ExponentiatedGradient [41] method.

These results are displayed in Fig. 9, 10 and 11. They show the same trends as the experiments using the baseline adversary $\mathcal{A}'$ (Fig. 2, 3 and 4, presented in Section IV-C1). In particular, we observe that the corrected reconstruction always has better accuracy than the original one made by the baseline adversary. More precisely, the tighter the fairness constraint (*i.e.*, the smaller the unfairness tolerance ϵ), the greater the reconstruction correction step improvement.

Appendix D
Detailed Results: Reconstruction Correction Performances from Estimated Fairness Constraints

In this section, we provide the reconstruction correction performances for our experiments using the ExponentiatedGradient [41] method, including a scenario in which the actual fairness constraint is not known. In such case, the adversary has to estimate it as described in Section V-B. Fig. 12, 13 and 14 display these results for our experiments on the three datasets, for the four fairness metrics, with baseline adversary $\mathcal{A}'$. Fig. 15, 16, and 17 display these results for baseline adversary $\mathcal{A}$. Results for both adversaries are very similar and show the same phenomenons.

As discussed in Section V-B, we observe several trends. When the fairness constraint is tight enough, the estimation process usually estimates it correctly. In this situation, the reconstruction correction step then exhibits slightly weaker performances as the provided tolerance estimation is not as tight as the actual constraint. However, when the unfairness tolerance ϵ is large enough, its actual value is not informative, while the estimated one is usually tighter leading to more accurate reconstruction results. Finally, the fairness constraint estimation process sometimes comes up with a fairness metric differing from the actual optimized one. When the proposed metric is more informative (in terms of number of examples involved in its computation), the reconstruction performance can even be better than with the actual constraint.

Appendix E
Additional Experiment: Reconstruction Performances Using a Pre-Processing Method for Fairness

In this appendix section, we provide results for additional experiments using a pre-processing method for fairness: the CorrelationRemover method, implemented in the `Fairlearn` library [40]. In a nutshell, the CorrelationRemover transforms the training set unsensitive attributes in order to remove their correlations with the sensitive ones. A traditional machine learning algorithm is then used on the sanitized data (pre-processed unsensitive attributes) to produce a fair model.

The CorrelationRemover does not optimize statistical fairness metrics explicitly. Indeed, bias against sensitive attributes is removed before training the model, in the data preprocessing step. Hence, in order to perform sensitive attributes reconstruction correction, one has to infer some fairness information. To do so, we use the strategy described in section V-B: the attacker measures the target model's unfairness on its own attack set, and chooses the metric with the smallest value. The experimental setup is similar to that of section IV-B. However, because the CorrelationRemover method does not optimize a particular fairness metric nor a particular tolerance value, we only perform one experiment for each dataset (repeated 100 times with different random seeds).

The results presented in Table V show that even in this context, the reconstruction correction step still provides significant reconstruction accuracy improvements. In all situations, the attacker was able to infer a valid fairness constraint and to leverage it to improve the initial sensitive attributes reconstruction.

Finally, these additional experiments confirm that the type of fairness intervention does not influence the performances of our proposed reconstruction correction step. The key factor for allowing reconstruction correction is that the predictions of the target model should be more fair than the original data. In this situation, the original attacker's reconstruction will likely be more biased than the (fair) target model's predictions, which will allow some reconstruction correction.

TABLE V

SUMMARY OF THE RESULTS OF OUR EXPERIMENTS USING A PRE-PROCESSING METHOD FOR FAIRNESS, WITH THE ATTACKER INFERRING THE FAIRNESS INFORMATION. WE REPORT THE ACCURACY PERFORMANCES OF THE TRAINED (TARGET) MODEL, THE RESULTS OF THE FAIRNESS CONSTRAINT ESTIMATION PROCESS (INFERRED METRICS AND AVERAGE INFERRED TOLERANCE), AND THE RECONSTRUCTION PERFORMANCES.

Target model (under attack)		**Estimated Constraint**		**Baseline Reconstructions**		**Corrected Reconstructions**	
Train Acc.	***Test Acc.***	***Estimated Metric***	***Estimated Tolerance***	$\mathcal{A}$	$\mathcal{A}'$	$\mathcal{A}$	$\mathcal{A}'$
UCI Adult Income dataset							
0.860 ± 0.003	0.848 ± 0.003	PE (68%), EO (32%)	0.023 ± 0.013	0.808 ± 0.005	0.806 ± 0.005	0.828 ± 0.013	$\mathbf{0.827 \pm 0.014}$
ACSPublicCoverage dataset							
0.862 ± 0.001	0.852 ± 0.002	PE (92%), SP (8%)	0.006 ± 0.004	0.861 ± 0.005	0.860 ± 0.006	0.863 ± 0.005	$\mathbf{0.872 \pm 0.010}$
ACSIncome dataset							
0.798 ± 0.002	0.785 ± 0.003	PE (100%)	0.056 ± 0.016	0.690 ± 0.007	0.685 ± 0.008	0.704 ± 0.014	$\mathbf{0.763 \pm 0.009}$

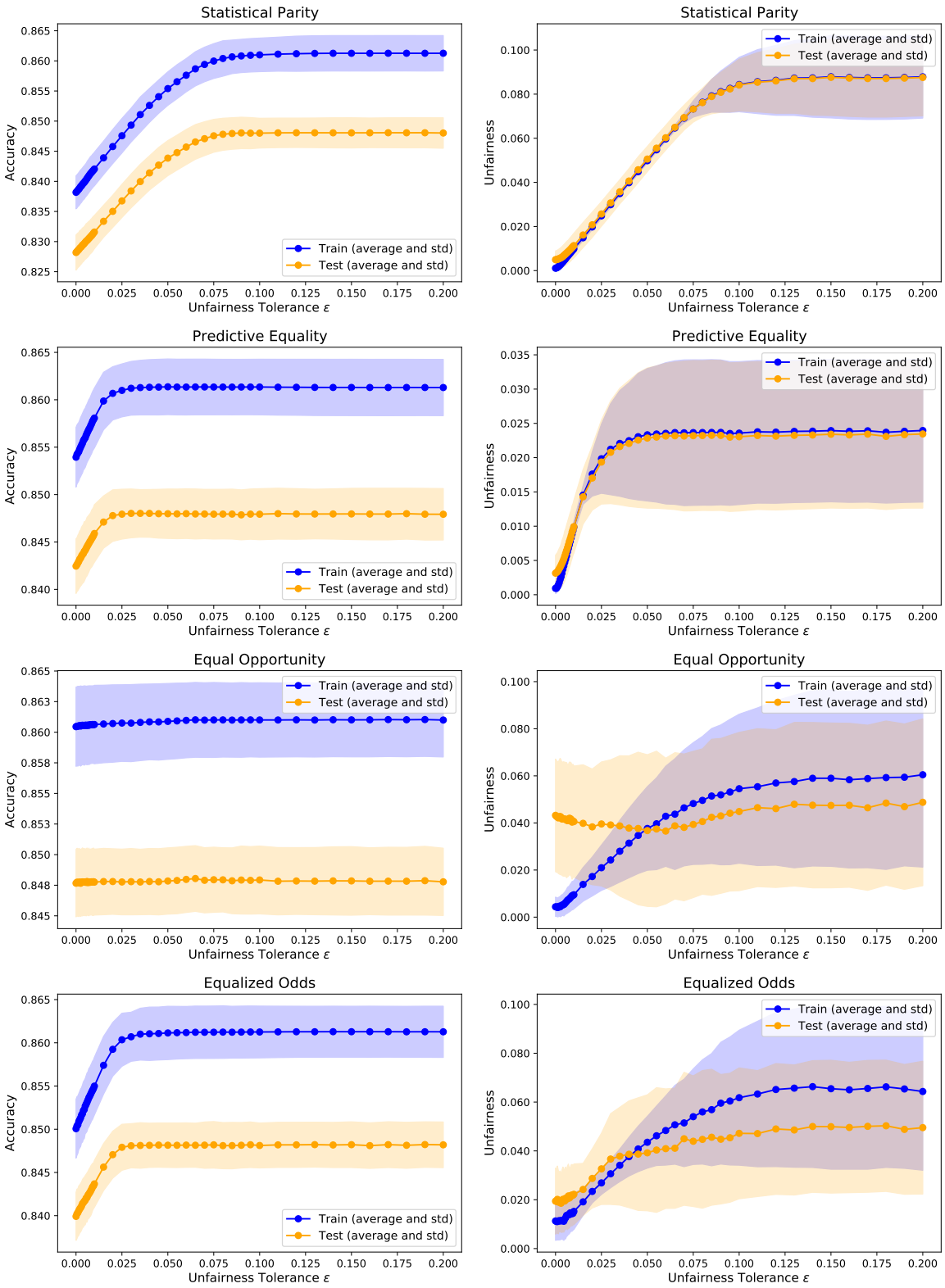

Fig. 6. Target models performances for our experiments using the UCI Adult Income dataset.

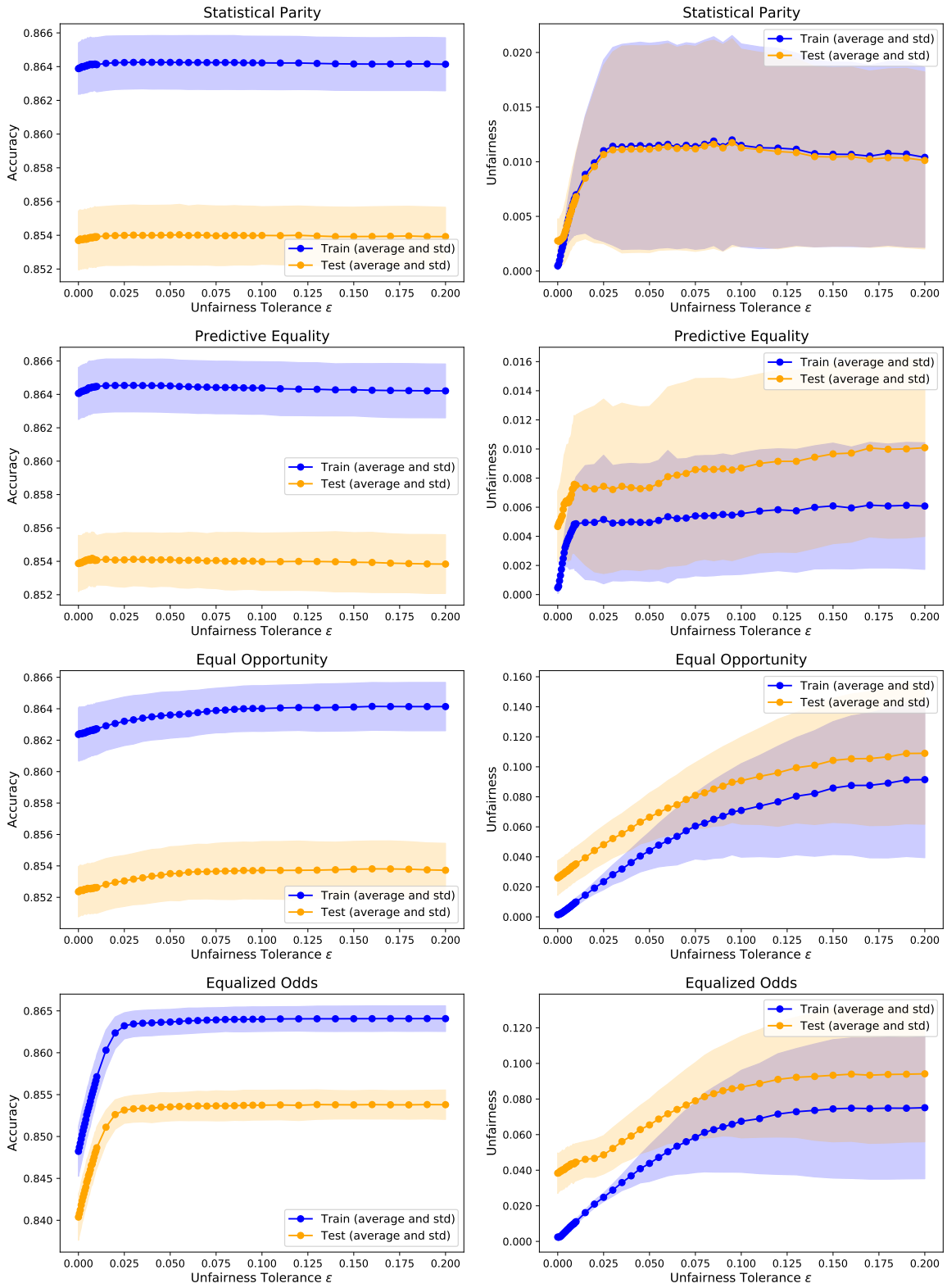

Fig. 7. Target models performances for our experiments using the ACSPublicCoverage dataset.

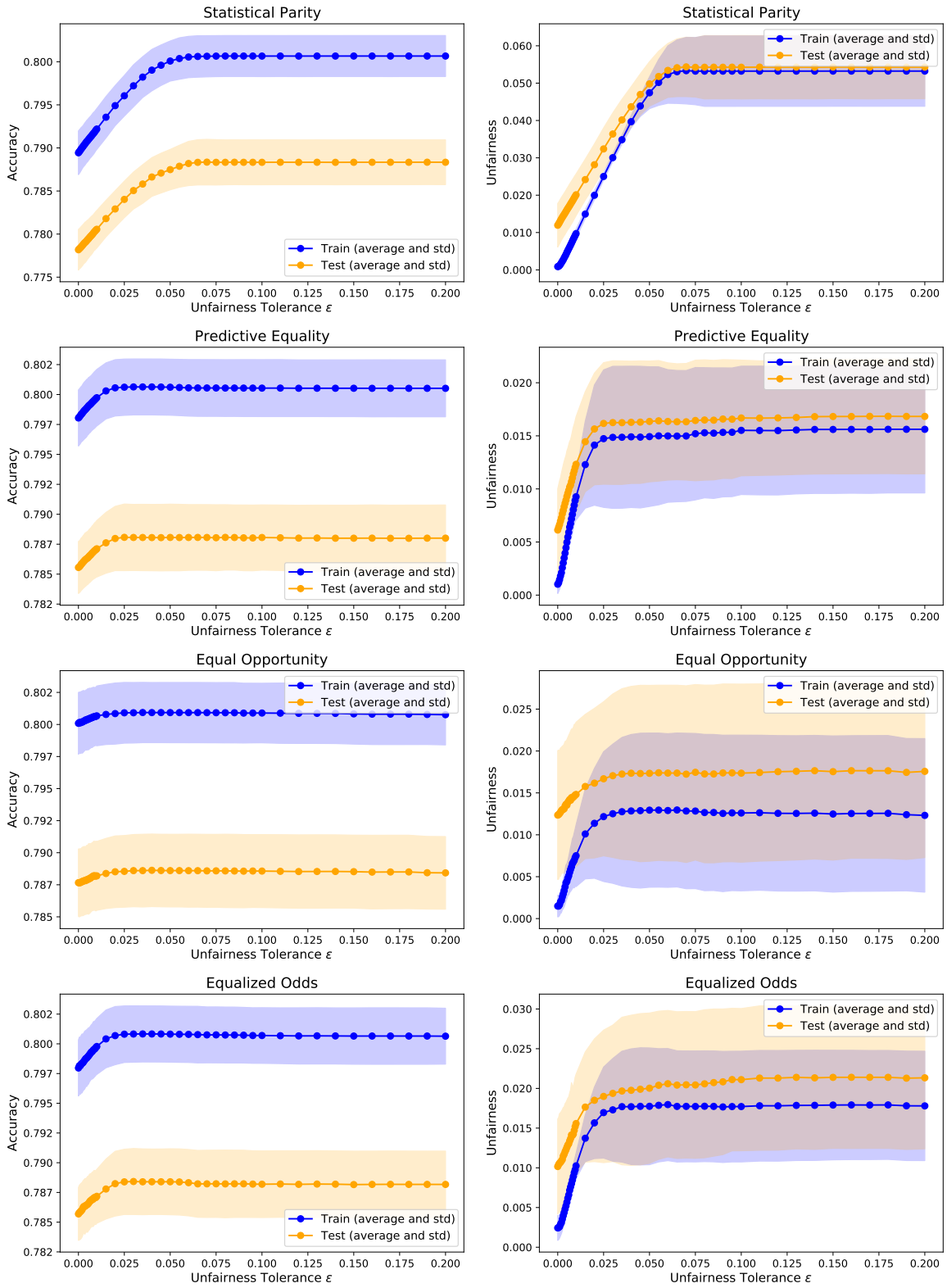

Fig. 8. Target models performances for our experiments using the ACSIncome dataset.

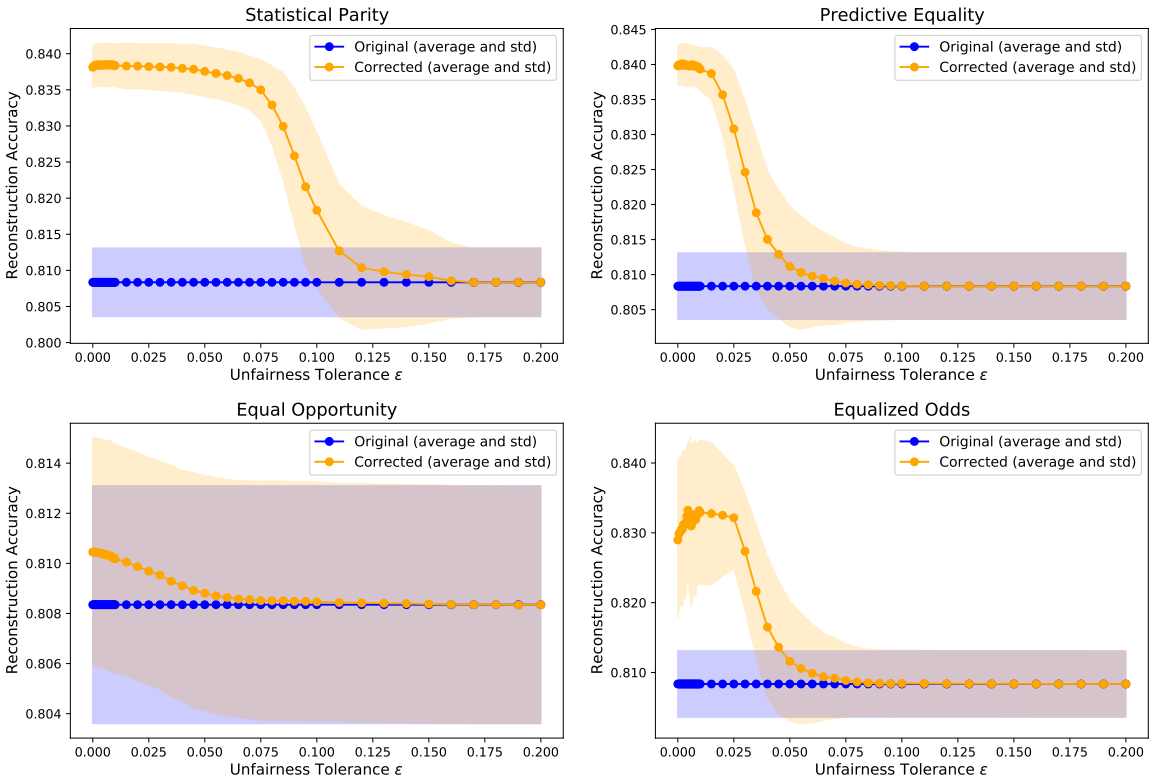

Fig. 9. Corrected and original (adversary $\mathcal{A}$) reconstruction quality, for our experiments using the UCI Adult Income dataset.

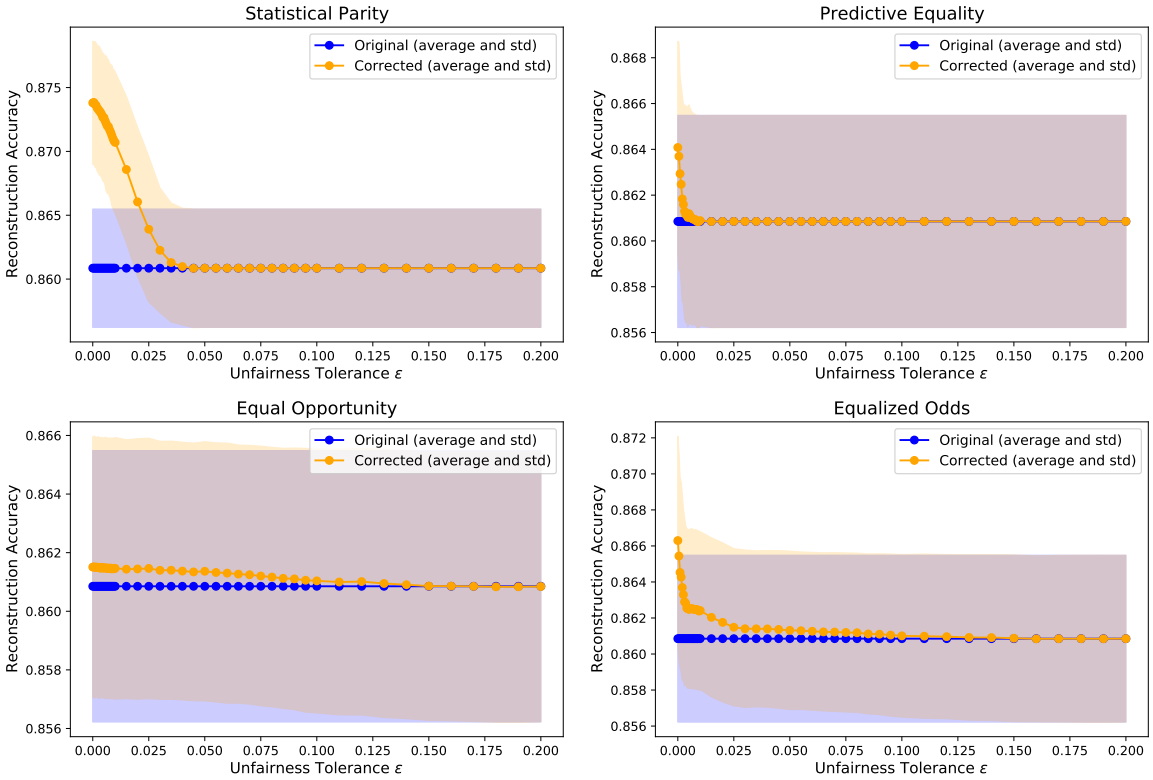

Fig. 10. Corrected and original (adversary $\mathcal{A}$) reconstruction quality, for our experiments using the ACSPublicCoverage dataset

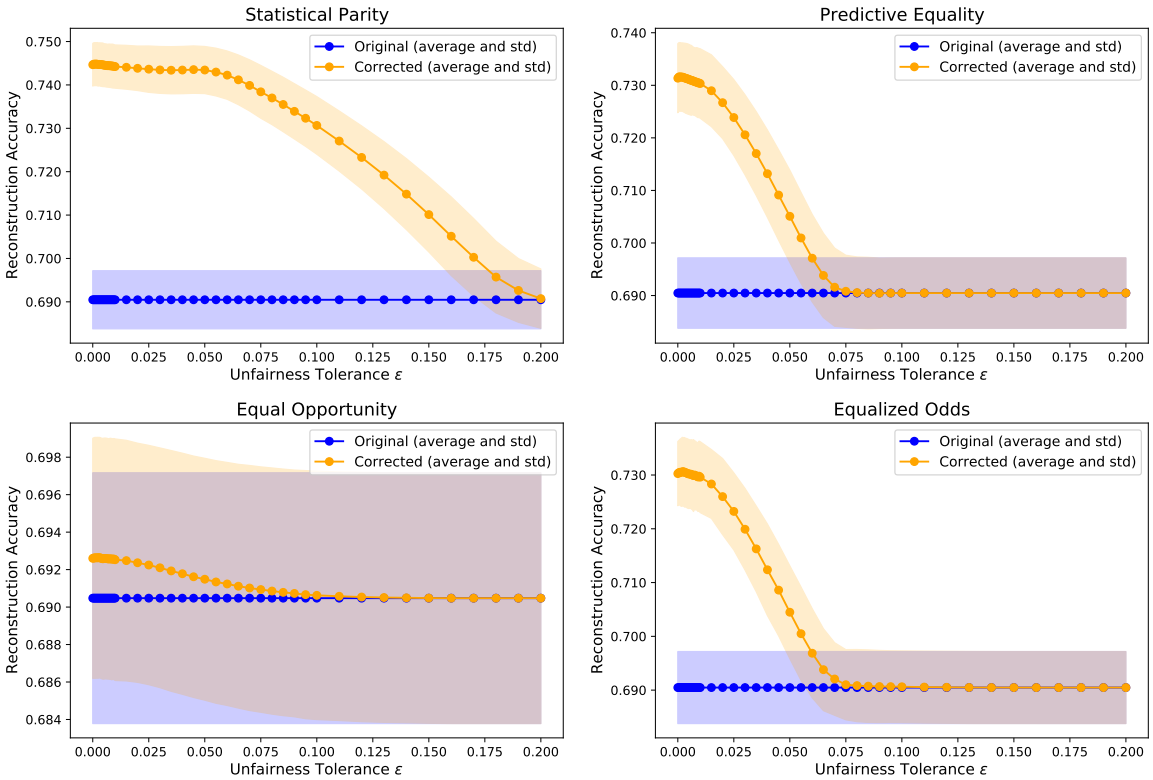

Fig. 11. Corrected and original (adversary $\mathcal{A}$) reconstruction quality, for our experiments using the ACSIncome dataset

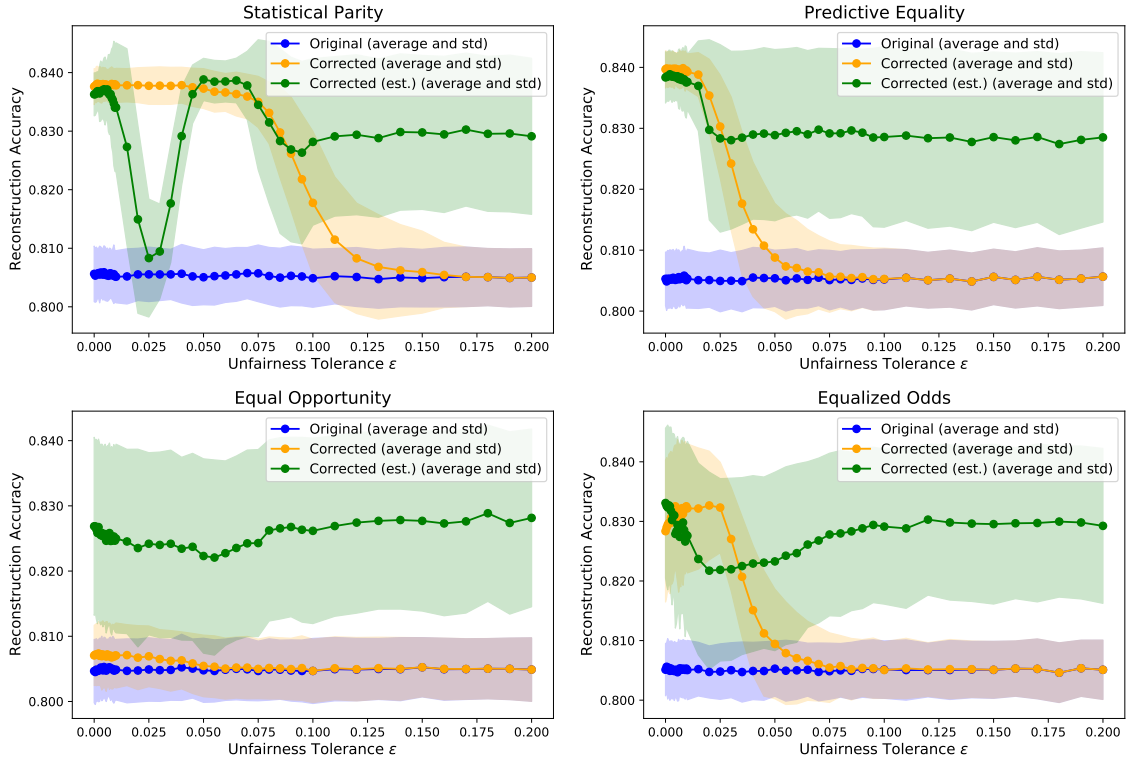

Fig. 12. Original (attacker $\mathcal{A}'$), corrected (from actual fairness constraint, and from estimated one (`est.`)) reconstruction quality, for our experiments using the UCI Adult Income dataset

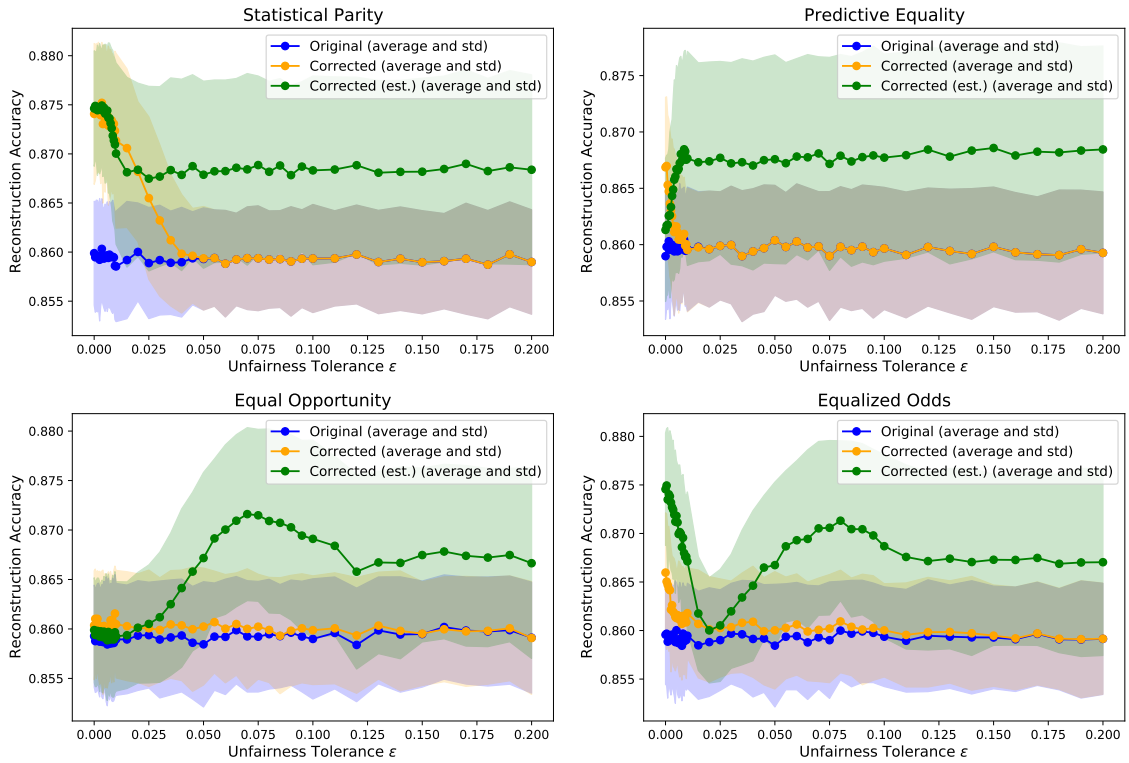

Fig. 13. Original (attacker $\mathcal{A}'$), corrected (from actual fairness constraint, and from estimated one (`est.`)) reconstruction quality, for our experiments using the ACSPublicCoverage dataset

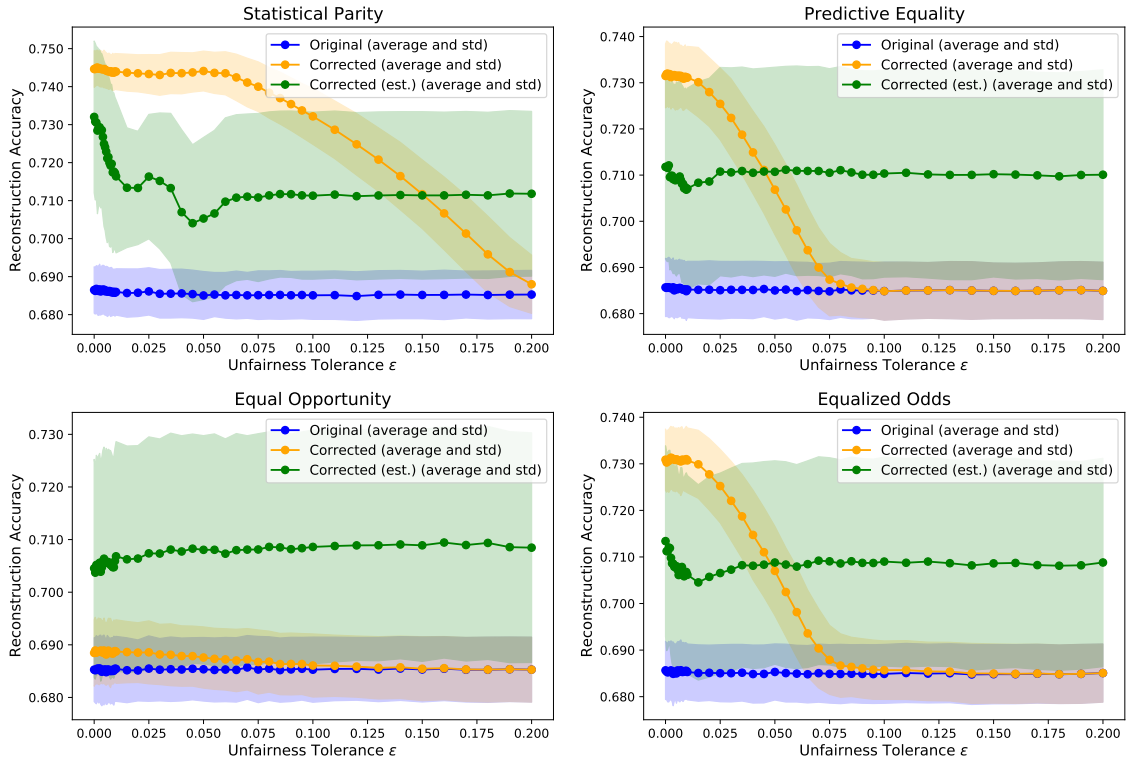

Fig. 14. Original (attacker $\mathcal{A}'$), corrected (from actual fairness constraint, and from estimated one (`est.`)) reconstruction quality, for our experiments using the ACSIncome dataset

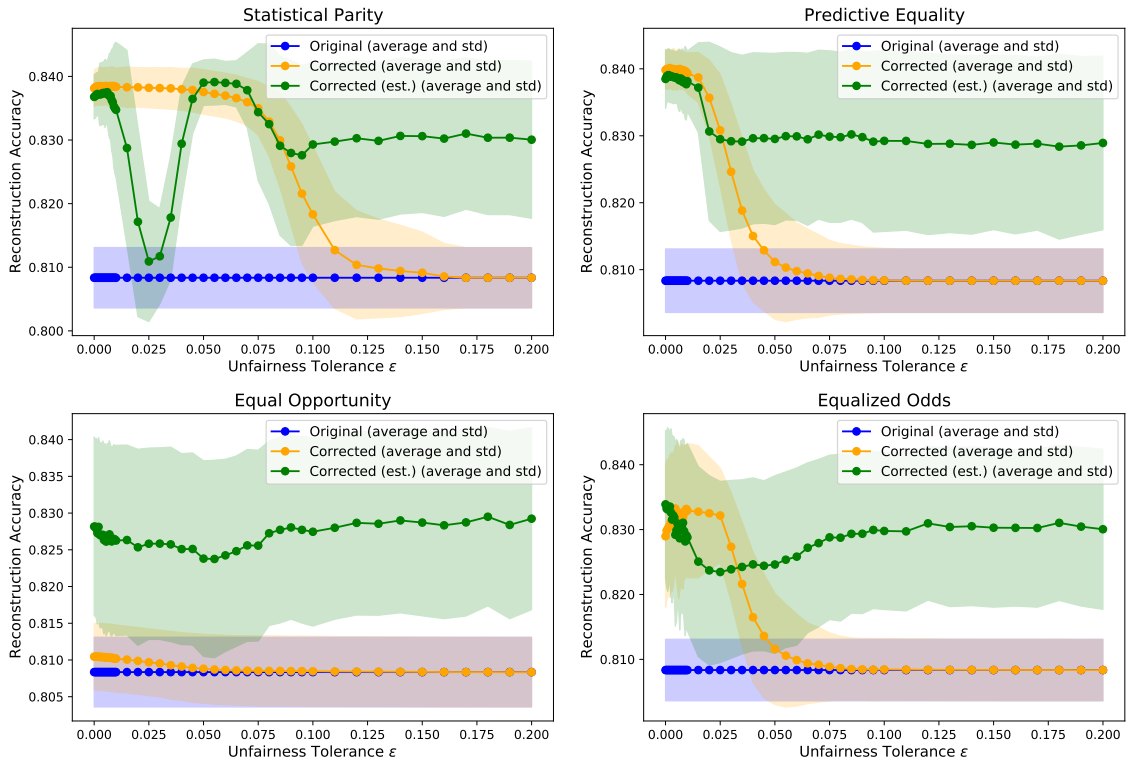

Fig. 15. Original (attacker $\mathcal{A}$), corrected (from actual fairness constraint, and from estimated one (`est.`)) reconstruction quality, for our experiments using the UCI Adult Income dataset

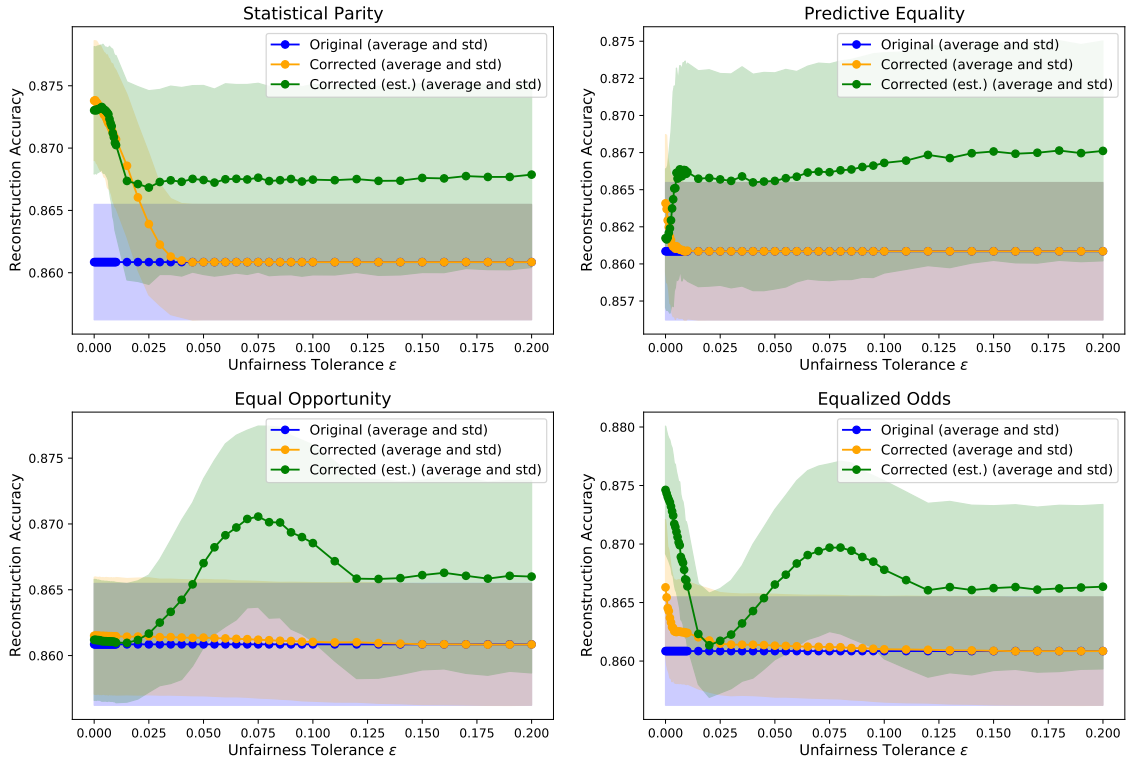

Fig. 16. Original (attacker $\mathcal{A}$), corrected (from actual fairness constraint, and from estimated one (`est.`)) reconstruction quality, for our experiments using the ACSPublicCoverage dataset

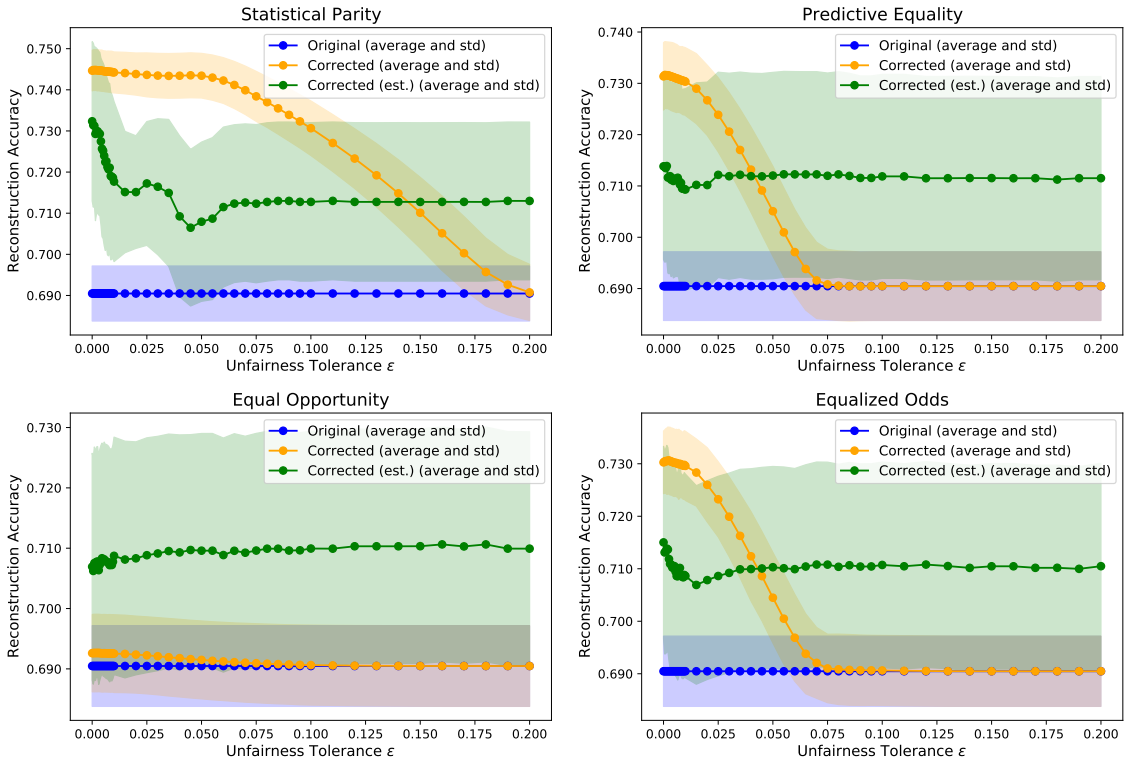

Fig. 17. Original (attacker $\mathcal{A}$), corrected (from actual fairness constraint, and from estimated one (`est.`)) reconstruction quality, for our experiments using the ACSIncome dataset

Reviews from Prior Submission

Anonymous Author(s)

***Abstract*—This document provides the reviews of the former version of our paper, which received the *Reject and Resubmit* decision. *Text in italics is the (unedited) text written by the reviewers.* Text in normal font corresponds to the authors' response. We first provide the meta-review, along with an explanation regarding how each specific point is addressed in the current version of the paper. Then, we provide the detailed reviews. Finally, we include the authors' rebuttal response (which was limited to 500 words).**

I. META REVIEW

Overall, the reviewers had a positive reaction to this work. All felt the paper was very well written and the evaluations were thorough. The reviewers did feel that the following suggestions would improve the paper and help it to be ready for publication.

1. A discussion of what the performance is when an adversary uses the wrong fairness metric? How does this error propagate through your system?

In the current version of the paper, Section V-B addresses the scenario in which the adversary does not know the fairness information in which case, he tries to infer it. We performed extensive experiments (on three datasets with diverse characteristics and sensitive attributes, for four fairness metrics and two fair learning algorithms) using a simple process to estimate the fairness metric and tolerance used for training the target model.

The results show that even when the fairness information inferred is not the actual one that was used for training, our reconstruction correction method could still lead to some improvements. Indeed, any approximately correct constraint that can be encoded in our framework can help reconstruction. The best-case results may not be as good as with the actual constraint, because the later is usually tighter, but there is still room for improvement.

Going further, this finding opens the door towards a wider use of our attack: even if the fair model was not trained using a statistical fairness metric, our post-processing reconstruction correction step may still improve the overall attack accuracy, by using an estimated statistical fairness metric.

The current version of the paper includes additional experiments (provided in Appendix E and briefly discussed at the end of section V-B) using a pre-processing method for enhancing fairness. This fair learning method removes correlations between the unsensitive attributes and the sensitive one, and produces a sanitized dataset on which traditional machine learning models can be learnt. Using this process, we obtain no guarantee regarding the satisfaction of a particular fairness constraint. Even in this setting, the results provided in Appendix E show that our simple fairness constraint estimation process can come up with valid constraints. For the three considered datasets, the inferred constraints allowed an improvement of the reconstruction accuracy using our reconstruction correction approach.

2. The gains from your method seem to be low. Please discuss more as why you nevertheless consider this approach promising and how might you extract higher gains.

The reconstruction accuracy improvements brought by our method are of several percentage points. Such improvements are in line with the literature on inference attacks, in which even a very tight advantage is considered relevant because it suggests a potential privacy breach. In our experiments, applying our simple post-processing step allowed for absolute improvements up to 6 pps (relative improvements close to 9 %) while never requiring more than a second to compute for real-size datasets (while the baseline attackers' training required several minutes).

3. Discuss potential issues with broadening beyond binary attributes for sensitivity. Why do you believe your work will generalize? What issues might arise and/or make this hard? What scalability challenges are there? Can you do this for one case of a non-binary metric?

Our reconstruction correction models are described for the binary sensitive attribute setting, which is the scenario considered in our experiments, and in many fairness works (including [1], which corresponds to the baseline adversaries used in our experiments).

However, both the general and efficient reconstruction correction models can be extended to the general case of multi-valued sensitive attributes. This is briefly discussed in section III-D, and we include a dedicated appendix section precisely describing how our models can be extended and what would be the incurred computational cost.

4. How general are your fairness metrics? Are they all part of a small family? Please discuss.

As discussed in section II-A, there are three main families of fairness notions. Our paper focuses on statistical fairness metrics, which have been by far the most studied and used. We consider the most common metrics from this family. All methods and results discussed also hold for any statistical fairness metric (and, more generally, for any rate constraints on a classifier's output). Dealing with new (different) such metrics would only require modifying the fairness constraints declared in the reconstruction correction model.

In the revised paper, we have clarified this aspect and clearly defined the notions that can be encoded in both models (sections III-B and III-C):

- Our "general" reconstruction correction model is suitable to encode *any* constraint on the sensitive attributes, be-

cause it explicitly encodes each one of them.

- Our "efficient" reconstruction correction model is suitable to encode constraints formulated as rates (possibly conditioned on the protected attributes). This includes (but is not restricted to) all statistical fairness metrics.

5. The authors should justify the choice of the attack model as the baseline to evaluate their correction method. It is unclear whether the improvement of their method comes from the choice of a poor attack model as the baseline. We recommend the authors use state-of-the-art reconstruction models as the new baseline or, alternatively, design experiments that show that their attack baseline is comparable or superior to existing reconstruction attacks.

The choice of a machine learning model as baseline attacker is in line with the literature [1]. Indeed, the attack of [1] is the only one in the literature considering the particular (but realistic) case where one wants to reconstruct an attribute which is not used for the target model's inference. Then, our considered ML-based baseline adversary $\mathcal{A}'$ is the strongest possible attacker, which in our setup, is given exactly the same information as our reconstruction correction model.

In our previous experiments, the attack model type and hyper-parameters were set based on preliminary experiments. For the revised version, we have followed the reviewer's advice and performed the following changes to our experimental setup:

- The proportion of data given to the adversary has been increased: currently, the same amount of data is used to train the attack model than to train the (fair) target model. More precisely, one third of the data is used to train the target model, one third constitutes the target model test set (useful to verify that the target model did not overfit), and the last third is given to the adversary to train its attack model (previously, this attack set only represented 20% of the overall dataset).
- Thorough preliminary experiments using the `hyperopt-sklearn` hyperparameters optimization library motivated the choice of Random Forests as attack model.
- The attack model's hyperparameters are optimized using the `hyperopt-sklearn` hyperparameters optimization library, with a maximum of 100 iterations for its Tree of Parzen Estimators search algorithm.

Overall, the performances of our previous baseline attacker were already good and the performed changes allowed for some minor improvements of the baseline attacker's performances. However, the improvement brought by our reconstruction correction step is of the same magnitude and the findings of the paper are unchanged. Furthermore, compared to the only paper in the literature tackling the same problem (without fairness considerations) [1], our baseline attack models offer considerable performances.

II. Detailed Reviews

A. Reviewer #1

1) Paper summary: This paper describes a way to use fairness constraints on top of an integer programming formulation to perform a reconstruction attack on the sensitive attribute of the training data of machine learning models. The proposed way is computationally expensive, but feasible for statistical notions of fairness, and is shown to significantly improve the adversary's accuracy in the reconstruction in several cases, as long as the fairness constraint is tight enough. The paper also shows that even when the fairness constraints are not known for sure by the adversary, they can be inferred from the use of the model as a black box with enough accuracy to allow for the attack to be performed with significant gains. An extensive experimental evaluation is performed, and thorough discussion is provided.

2) Comments for author: This paper is extremely well written and presented. It was a pleasure to read it for its clarity, conciseness, and competent coverage of the background needed. I even think that this paper's introductory sections can be used as a summary of what is going on in the privacy vs. fairness area for a student or researcher who is new to the topic.

Regarding the technical contributions, the paper is solid. The proposed framework for improving the accuracy of inference attacks is conceptually simple, but novel and insightful. Moreover, it is elegantly formalized and clearly presented. The technical difficulties for scalability are clearly discussed, and situations in which it can be achieved (more precisely, when statistical notions of fairness are used) are presented and experimentally tested. The experimental results are thorough, clearly presented, and convincingly show that this type of attack-correction based on fairness constraints must be considered seriously by experts, and possible novel interesting research may result from this observation.

As a minor (non-mandatory) suggestion, I'd encourage the authors to expand the information of Table 1 to add a sentence with the intuitive explanation of each fairness notion. The reason for this suggestion is that the introductory sections cover so thoroughly the background, that this is the only addition I think a novice to the field who's reading this paper would need to have a first understanding of the area.

3) Areas to address in response: Reading the paper, I was curious about in which situations the lack of knowledge, from the part of the adversary, about the fairness notion employed during training would impact the accuracy of the corrected prediction. In particular, I wondered if it was possible, in principle, that if the adversary used a wrong fairness measure then the extra-step would make the "corrected" prediction worse. Experiments were performed with the correct fairness measure being employed, and then there was a discussion on how the adversary can gain good confidence of the fairness notion used and exploit it. But this discussion assumes the fairness metric is chosen from a family of well known metrics. However, is there anything else to say about situations in

which the fairness metric cannot be inferred? Is it possible that the correction step makes the final prediction worse?

B. Reviewer #2

1) Paper summary: In the context of fair machine learning, this paper proposes a technique to improve an adversary's inferences about a sensitive attribute of the training set. The authors point out the apparent incompatibility between the fairness and privacy objectives and show that a black-box adversary with information about the fairness guarantee can learn the sensitive attribute—even when such attribute is not present in the training dataset. The adversary considered in this paper makes an initial guess by training an attack model on the outputs of the fair model (the target model), and solves an integer program that corrects the initial guess on low-confidence data points. The authors evaluate the correction technique empirically on a wide range of datasets, fairness metrics, and tolerance values. The results show a modest increase in the accuracy of the attack over the initial guess. The authors also discuss potential mitigation strategies and evaluate them against their inference attack.

2) Comments for author: The paper is well written and easy to follow. The proposed technique is simple but technically sound, and its evaluation is well designed; the authors choose relevant scenarios, and use appropriate baselines and metrics.

A concern with this paper is its small contribution to the field. Its novelty lies in the correction method, which basically flips low-confidence decisions such that the model satisfies a fairness definition. The results in Table 3 show that the difference in the accuracy of the correction over the baseline ranges between approx. 0-7pp. Although significant, this is a modest improvement.

My main concern with the paper is that since the attack model implicitly uses fairness information, as its training data are the predictions of the fair model, the adversary's initial guess already incorporates fairness information. Thus, it is unclear whether this gap between the correction and the baseline is artificial and could actually be reduced by refining the attack model or using more data.

Unfortunately, I do not think this paper meets the bar for publication and recommend rejecting it. Unless the authors can show that the attack model cannot be significantly improved, my concerns are fundamental and would require reframing the paper. Some minor comments follow:

- It is unclear until late in the paper that the adversary makes their first guess by training an attack model.

- I am not sure why the authors make reference to "declarative programming approaches." I am only familiar with this concept in the context of programming languages, not regarding mathematical optimization techniques.

- The assumptions that the authors make require a deeper evaluation and discussion. For example, they assume that the adversary has reliable confidence scores. This assumption is critical for the effectiveness of the correction method, but it is unclear how the adversary can obtain such reliable confidence scores. In particular, the authors use random forests as the attack model's underlying learning algorithm, which is known to be uncalibrated by default.

- Section 4.1 should give more detail about the attack models. It is unclear where the attack models of adversaries A and A' differ. For example, how does A' use the extra information about the sensitive attributes?

- I would disagree with the points the authors make in Section 5.1. The authors argue that differential privacy is not an effective countermeasure because "statistical fairness metrics are measured over the entire dataset and do not specifically rely on individual examples." (lines 1157-1158). However, the value to be protected is the sensitive attribute, not the fairness information, which is individual information. A fairness definition should protect the sensitive attribute information and not the fairness information—a different question is whether the fairness notion can be satisfied under such a DP protection.

Nitpicks ————-

- Even though Table 1 defines the fairness notions in terms of probabilities, the authors ensure the fairness notions on the rates (e.g., false positive rate) as calculated on the testing dataset. The rates approximate the probabilities, but it is not always true that by ensuring the fairness definitions on the rates, the model will meet the fairness definition. Perhaps, it would be a bit more accurate to rewrite Table 1 in terms of the rates.

*- I do not think that constraints (4) and (5) in the definition of the optimation problem match the definition of statistical parity in Table 1: note that the constraints only ensure that the difference is less than 2*epsilon (e.g., think of constraint (4) being equal to epsilon and constraint (5) equal to -epsilon). The solution is to take epsilon'=epsilon / 2.*

- I expect the confidence bands in the graphs for EO to be much tighter since the stdev values in Table 3 are so small.

- Graphs lack consistency (please, label lines with A, A', etc.).

3) Areas to address in response: My most pressing concerns are on top of the bullet points

C. Reviewer #3

1) Paper summary: The paper proposes an integer programming technique that exploits the constraints from fairness-aware learning algorithms to more accurately reveal sensitive attributes from the underlying training data. To do so, the approach augments the standard surrogate model pipeline for inversion with an integer programming problem that improves the initial prediction of the attacker's surrogate model by ensuring they meet various constraints imposed by statistical fairness measures. An optimization is proposed to reduce the overall complexity of the problem to polynomial time in the number of samples by taking advantage of aggregate counts and confidence scores across the allowable 'corrections.' The approach is evaluated across a number of datasets and fairness constraint settings, where the results

generally show improved prediction accuracy when fairness constraints are relatively tight.

2) Comments for author: Strengths: + Clever example of using training constraints to reveal sensitive information + Very clearly written paper and description of the techniques + Experiments generally follow standard datasets and methodologies in the fairness space

Weaknesses: - The overall contribution of the attack feels somewhat low given findings of previous work in database reconstruction (e.g., [12, 20, 28, 30]), model inversion (e.g., [26, 27]), and even reconstruction with fairness constraints (e.g., [A]) - Significant concerns about how well the attack would scale in practice (beyond binary variables) - Claims of generality feel a bit overblown given that all fairness constraints examined were related - Improvement in sensitive attribute prediction appears to be relatively small or non-existent in many cases

General Comments: Overall, I thought the paper was really well written and engaging. The background section, in particular, provided good discussion of the techniques and set up the notation nicely for the rest of the paper. Likewise, the description of the integer programming problem was clear and well-motivated. While I generally felt that the experiments examined toy problems, they do seem to fall in line with many of the other papers in the fairness and model inversion space (e.g., [A, 26]), which makes it difficult to complain about it in this paper. The broad takeaway of the paper, which highlights that training constraints can exacerbate data leakage, reinforces findings of previous work – that means the results are probably not particularly surprising, but could still help socialize the problem to the broader community.

There are, however, quite a few big concerns that I have with the paper and how its contributions are framed. Most notably, there is a lot of time spent highlighting the generality of the attack to various fairness constraints and sensitive attribute settings, however when we reveal the underlying mechanism these claims come off as overblown. As an example, the claims that the attack works against a number of fairness constraints really turn out to be limited to a family of very closely related constraints, and many other fairness approaches (e.g., individual fairness) or even other statistical constraints (e.g., conditional equality definitions) do not seem to be covered.

Additionally, even in the relatively narrow setting tackled in this paper, there are some pretty big questions surrounding the practicality of the approach for any non-trivial data. Despite many claims in this paper (see Section 3.4) that approaches demonstrated for single binary sensitive attributes can be generalized to more realistic attributes (numeric or categorical), the discussion in Section 3.3 seems to imply otherwise. In fact, it is difficult to imagine how any non-trivial categorical attribute (where values exist independently of each other) could be encoded in a similarly efficient optimization problem. There needs to be additional work to demonstrate that this is effective outside of toy problem settings.

Probably the last major complaint is that the experimental results are simply not particularly impressive given the amount of computational effort that goes into the integer programming approach. In the case of the equal opportunity constraint, there is no significant benefit at all, while in other cases it only provides a 6 or 7

Aside from these larger issues, there were a few smaller items that I will enumerate below that I think need to be addressed, but which are primarily editorial in nature. For example, the paper needs more clarity around the adversarial model, including why we believe the adversary would have knowledge of the epsilon parameter and what happens to the viability of the attack when there is a mismatch there.

[A] Hu, Hui, and Chao Lan. "Inference attack and defense on the distributed private fair learning framework." The AAAI Workshop on Privacy-Preserving Artificial Intelligence. 2020.

*3) Areas to address in response: * Tone down the claims of general attacks against a broad set of fairness constraints. Instead, clarify the exact class of fairness constraints that this can operate on, and be clear about what constraints (even within just statistical fairness) that the attack will not naturally handle. It should be clear that the attack focuses on a single family of related fairness constraints, and that there are a large number of others that the current instantiation will not naturally handle.*

** Make more effort to demonstrate and analyze the impact of moving from simple binary sensitivity attributes to numeric and categorical attributes with realistic cardinality. What would the integer programming problem look like in those settings, and would there be a way to make them efficient?*

** Provide a more explicit adversarial model. This should include assumptions about adversary's knowledge, access to data and the target model, and other knowledge about the training procedure. As it stands, this information is spread throughout the paper in various sections, and I think it would help tremendously to state these in a succinct discussion early in the paper.*

** Further to the point of the adversarial model, some discussion needs to be included about why the adversary should be expected to have precise knowledge of the epsilon constraint parameter used during training. Either provide examples for where these parameters are shared in practice, or provide evaluation of what happens to the attack when there is a mismatch between attacker epsilon and that which is used in training.*

** Provide additional context and examples from the literature demonstrating that the lift in attack accuracy is substantial and that the range of epsilon values in which it operates are realistic/common.*

** While it is not exactly the same as the attack presented in this paper, I found a paper that uses similar techniques to attack fairness-aware models trained in a distributed setting. It would be worthwhile to discuss the proposed approach in the context of this other paper [A].*

[A] Hu, Hui, and Chao Lan. "Inference attack and defense on the distributed private fair learning framework." The AAAI Workshop on Privacy-Preserving Artificial Intelligence. 2020.

D. Reviewer #4

1) Paper summary: This paper shows how using information obtained from the fairness evaluation of an ML model can be used by an adversary to improve the performance of reconstruction attacks to recover the entire column of sensitive attributes of the training set. The proposed method is agnostic to the fairness-enhancing technique and the type of machine learning model.

2) Comments for author: The paper is very well-written. The introduction motivates the problem well. The Background and Related work Section is comprehensive and makes the right connections between this paper and the related work.

The paper considers binary classification tasks and assumes that a fair learning algorithm preserves fairness by not being trained on sensitive attributes.

The attack model assumes that the adversary knows an approximation of the underlying distribution of training datasets. This assumption seems to be accepted by the research community as the authors have provided references.

The idea is to augment the reconstruction model, which is an optimization problem, by adding constraints obtained from fairness properties. Using these constraints can reduce the search space too.

Evaluation setup is formally defined. Three datasets of the fairness literature with different sizes have been used. Two fairness models from Fairlearn have been tested, and Random Forest classifiers are used as machine learning algorithms.

While this approach improves the performance of the reconstruction attack, this improvement is very little, in the range of 0.02-0.07. Not sure if this is a considerable improvement, as the authors have claimed. The authors should discuss acknowledge and discuss this.

Some good insights are provided about when and how the attacks perform better.

A discussion on countermeasures has been provided, where it is discussed that DP is not a good solution, and an intuitive approach is implemented to perturb the fairness information. However, the results again do not show much improvement, at best about 0.04.

In summary, this is a timely paper addressing a fundamental problem. The attack model and approach are explained very well. The evaluation is sound. The results are discussed and interpreted well, even though the evaluation results do not show that much improvement in the attacks' performances. The discussion of countermeasures and the implementation of an intuitive method is excellent, but not sure if it is meaningful when the performance of the attack and defense methods is low.

III. Rebuttal Authors Response

We thank the reviewers for their insightful comments and suggestions.

- All reviewers:
  - Adversary A knows an approximated data distribution, while A' also leverages the outputs of the target model. We will use standard security game formalism to specify each adversarial framework
  - Our experiments demonstrate that our approach achieves attack accuracy improvements of several pps (up to 8pps) compared to an informed baseline. Since random guess achieves 50% accuracy, such improvements are considerable and in line with the literature on privacy attacks
  - The reconstruction correction step is computationally efficient: for realistic datasets ($> 100,000$ examples), solving our ILP takes fractions of seconds (see Section 4.2.6 - detailed running times can be reported in Appendix).
  - Minor:
    - fairness metrics will be intuitively explained
    - in practice, fairness constraints are enforced with empirical rates instead of probabilities
    - group-wise difference can reach 2*epsilon
- Reviewer#A:
  - While the reconstruction correction could theoretically worsen the prediction, it is unlikely to occur in practice as the correction focuses on low-confidence elements, and simple fairness estimation processes are able to infer (approximately) valid constraints.
- Reviewer#B:
  - Part of the fairness information is implicitly given to the adversary A' through the target model's predictions on the attacker's data. However, it is not completely accurate as fairness measured on the attack set may differ from that on the training data. Our approach can bring improvements whenever the original attacker's guess does not satisfy the actual fairness constraint
  - The baseline attack model was selected after thorough preliminary experiments on diverse machine learning models and different proportions of data given to the adversary using hyperopt-sklearn, which confirmed that Random Forests are the best performing models
  - DP will not perturb the fairness information by more than the sensitivity of the fairness constraints with respect to the individual sensitive attributes (to be protected). As our attack relies only on this global information, we do not expect its performance to degrade significantly
- Reviewer#C:
  - Our "efficient" model can encode any rate constraints on the protected groups (including all statistical fairness metrics). We can use our "general" model to formulate any constraint on the protected attributes, as it explicitly encodes them
  - Our "efficient" model uses one variable to count the number of changes from one sensitive attribute value to another. This requires $O(l^2)$ variables and $O(l)$ constraints, with l the number of possible

values for the sensitive attribute, which results in a polynomial search space (details can be included in the Appendix).

- Epsilon is publicly known when fairness implements a legal requirement (e.g., the 80% rule for statistical parity). This is in line with [A], where the adversary knows the metric and the tolerance. Furthermore, even if none of these are known, we have demonstrated that simple estimation strategies still allow significant reconstruction accuracy improvements
- The range of values for epsilon are in line with the literature. The smallest implement exact fairness while the largest actually do not constrain the learning
- Although the considered setup is different, [A] has similar objectives and uses close techniques (which will be discussed in Section 2).

REFERENCES

[1] J. Aalmoes, V. Duddu, and A. Boutet, "Dikaios: Privacy auditing of algorithmic fairness via attribute inference attacks," *arXiv preprint arXiv:2202.02242*, 2022.