# On the Alignment of Group Fairness with Attribute Privacy

Jan Aalmoes, Vasisht Duddu, Antoine Boutet

# On the Alignment of Group Fairness with Attribute Privacy

Jan Aalmoes[1 (✉) [0000−0003−0075−8965]], Vasisht Duddu[2 [0000−0003−2138−4341]], and Antoine Boutet[1 [0000−0002−4057−416X]]

[1] Univ Lyon, INSA Lyon, Inria, CITI, Lyon, France - jan.aalmoes@inria.fr
[2] University of Waterloo, Waterloo, Canada

**Abstract.** Machine learning (ML) models have been adopted for applications with high-stakes decision-making like healthcare and criminal justice. To ensure trustworthy ML models, the new AI regulations (e.g., AI Act) have established several pillars such as privacy, safety and fairness that model design must take into account. Designing such models requires an understanding of the interactions between fairness definitions with different notions of privacy. Specifically, the interaction of group fairness (i.e., protection against discriminatory behaviour across demographic subgroups) with attribute privacy (i.e., resistance to attribute inference attacks—AIAs), has not been comprehensively studied. In this paper, we study in depth, both theoretically and empirically, the alignment of group fairness with attribute privacy in a blackbox setting. We first propose AdaptAIA, which outperforms existing AIAs on real-world datasets with class imbalances in sensitive attributes. We then show that group fairness theoretically bounds the success of AdaptAIA, which depends on the choice of fairness metrics (e.g., demographic parity or equalized odds). Through our empirical study, we show that attribute privacy can be achieved from group fairness at no additional cost other than the already existing trade-off with utility. Our work has several implications: i) group fairness acts as a defense against AIAs, which **is** currently lacking, ii) practitioners do not need to explicitly train models for both fairness and privacy to meet regulatory requirements, iii) AdaptAIA can be used for blackbox auditing of group fairness.

## 1 Introduction

Machine learning (ML) models have been adopted for several high-stakes decision-making applications, such as criminal justice and healthcare. To govern this massive deployment of ML models, new AI regulations have highlighted the design of trustworthy models. Trustworthy models rely on several pillars such as privacy, safety, fairness [20, 18]. The design of models ensuring all these (potentially conflicting) properties remains an open challenge and requires an understanding of the relationship among them.

For instance, to avoid models susceptible to discriminatory behaviour [25], group fairness algorithms train an ML model optimized for a fairness metric (e.g.,

equalized odds, demographic parity) to ensure equitable behaviour across different demographic subgroups [35, 1]. This optimization ensures the conditional independence between the model's predictions and sensitive attributes [33, 17]. However, training with group fairness may conflict with different notions of privacy [10]. For instance, group fairness increases the susceptibility to membership inference attacks [6] and conflicts with differential privacy [8, 13]. However, there is limited literature on the interaction of group fairness with privacy of sensitive attributes, as measured using *attribute inference attacks* (AIAs) where an adversary infers sensitive attributes (e.g., `Race` and `Sex`) from model predictions [15, 9, 32, 29, 23, 24]. Ferry et al. [12] indicate a conflict in a restricted setting: fair models are susceptible to AIAs when adversary *knows the fairness metric* that the model was optimized on. This assumption is unlikely in practice since companies do not reveal their proprietary training procedures. Hence, the interaction in a blackbox setting where the adversary has no knowledge about the target model is more realistic, but missing in the literature. On the contrary, Zhang et al. [36] only speculate the alignment of group fairness with attribute privacy without any evaluation. Hence, it is still unclear how different fairness metrics influence the interaction [36]. Despite the GDPR's emphasis on safeguarding individuals against attribute inference, this specific privacy risk has not been thoroughly evaluated concerning its trade-off with fairness in ML models.

To address this lack of understanding, we study the following research question: ***How does group fairness interact with attribute privacy?*** We formally define attribute privacy as the indistinguishability in the model's predictions for different sensitive attribute values [36]. Empirically, we evaluate this by checking whether the AIAs are close to random guess. Intuitively, the conditional independence of sensitive attributes on using group fairness should be equivalent to indistinguishability in model predictions. Since this meets the requirement for attribute privacy, we conjecture that there is alignment.

Our goal is to validate this conjecture by examining whether the success of AIAs is close to random guessing, which would imply attribute privacy and indicate alignment with group fairness. However, this is challenging. First, none of the current AIAs account for real-world datasets with significant class imbalance in sensitive attributes making them ineffective in practice. Second, group fairness algorithms can either output soft labels (probability that an input belongs to different classes) as seen for adversarial debiasing (AdvDebias) [35, 22] or hard labels (most likely class from soft labels) as seen in exponentiated gradient descent (EGD) [1]. We have to design AIAs for both settings. We address these by proposing a state-of-the-art AIA, AdaptAIA, to measure attribute privacy. We theoretically analyze the bounds of different fairness metrics to protect against AIAs, and validate these bounds through an extensive experimental analysis using several state-of-the-art datasets.

## 2 Background and Related Work

### 2.1 Machine Learning Classifiers

**Training.** An ML classifier is a function $f^{\theta}_{trg}$ (omit $\theta$ for simplicity) parameterized by $\theta$ that map inputs with corresponding classification labels. $\theta$ is updated using a training dataset ($\mathcal{D}_{tr}$) with the objective to minimize the loss incurred on predicting the classification label for inputs from $\mathcal{D}_{tr}$. We remove sensitive attributes such as `Race` or `Sex` from $\mathcal{D}_{tr}$ to censor them [29]. Consequently, $f_{trg}$ is trained on non-sensitive attributes.

Formally, consider a probability space $(\Omega, \mathcal{T}, \mathcal{P})$, measurable spaces $(E, \mathcal{U})$, $(\{0,1\}, \mathcal{P}(\{0,1\}))$ and $([0,1], \mathcal{B})$ where $\mathcal{B}$ is the Borel tribe on $[0,1]$. We define random variables $X$ for the input data, $Y$ for the classification labels and $S$ for the sensitive attributes: $X : (\Omega, \mathcal{T}, P) \longrightarrow (E, \mathcal{U}),$, $Y : (\Omega, \mathcal{T}, P) \longrightarrow (\{0,1\}, \mathcal{P}(\{0,1\})),$, and $S : (\Omega, \mathcal{T}, P) \longrightarrow (\{0,1\}, \mathcal{P}(\{0,1\}))$. Then, $f_{trg}$ is a measurable function $f_{trg} : (E, \mathcal{U}) \longrightarrow ([0,1], \mathcal{B})$ which is used to build the statistic approaching $Y$ by updating the parameters $\theta$ on $\mathcal{D}_{tr}$. The prediction of $f_{trg}$ on $X$ is a random variable: $\hat{Y}_h = 1_{[\tau,1]} \circ \hat{Y}_s$ where $\hat{Y}_s = f_{trg} \circ X$ and $\tau \in [0,1]$.
**Inference.** Once training is completed, $X(\omega)$ is passed to $f_{trg}$ to obtain a prediction score $f_{trg}(X(\omega))$ (aka soft labels). The attributes during inference, are sampled from an unseen test dataset $\mathcal{D}_{te}$ disjoint from $\mathcal{D}_{tr}$ to evaluate how well $f_{trg}$ generalizes. We refer to $f_{trg}$'s final predictions and intermediate outputs as *model observables*. Sensitive attributes, although available for different data records, play no role in training or inference. They are reserved solely for designing and evaluating attacks.

### 2.2 Group Fairness

Generally, data records in the minority subgroup, identified by some sensitive attribute (e.g., `Race` or `Sex`), face unfair prediction behaviour compared to data records in the majority subgroup. For instance in criminal justice, `Race` plays a non-negligible role in predicting the likelihood of them re-offending [3]. Group fairness algorithms add constraints during training such that different subgroups (i.e., $S : \Omega \to \{0,1\}$) are treated equally (e.g., AdvDebias [35] and EGD [1]). $S$ is either `Sex` or `Race` (i.e., $S(\omega)$ is 0 for woman and 1 for man, or 0 for black and 1 for white). There are different definitions of group fairness which have been introduced in prior work. We discuss two well-established definitions: demographic parity (DemPar) and equalized odds (EqOdds).

**Definition 1.** *$\hat{Y}_h$ satisfies* DemPar *for $S$ if and only if:* $P(\hat{Y}_h = 0|S = 0) = P(\hat{Y}_h = 0|S = 1)$.

DemPar ensures that the number of correct predictions is the same for each subgroup. However, this may result in different false positive (FPR) and true positive rates (TPR) if the true outcome $Y$ varies with $S$ [11]. EqOdds is a modification of DemPar to ensure that both TPR and FPR are the same for each subgroup [17].

**Definition 2.** *$\hat{Y}_h$, classifier of $Y$, satisfies* EqOdds *for $S$ iff:* $P(\hat{Y}_h = \hat{y}|S = 0, Y = y) = P(\hat{Y}_h = \hat{y}|S = 1, Y = y) \quad \forall(\hat{y}, y) \in \{0, 1\}^2$.

We consider two fairness algorithms: (a) adversarial debiasing (AdvDebias) [22, 35] and (b) exponentiated gradient descent (EGD) [1]. AdvDebias achieves fairness by training $f_{trg}$ to have indistinguishable output predictions in the presence of a discriminator network $f_{disc}$. $f_{disc}$ infers $S$ corresponding to a target data point given $f_{trg} \circ X$. $f_{trg}$ is then trained to minimize the success of $f_{disc}$. AdvDebias outputs *soft labels* (i.e., a probability attached to each value of the sensitive attribute). EGD solves an under-constraint optimization problem to find a collection of optimal measurable functions $t_0, \cdots, t_{N-1}$ and threshold $(\tau_0, \cdots \tau_{N-1}) \in [0, 1]^N$. They are used to create the statistic for predictions $\hat{Y}_h$ to estimate $Y$. A random variable $I : \Omega \to \{0, \cdots, N-1\}$ selects one of the measurable functions and generates a randomized classifier: $\hat{Y}_h = 1_{[\tau_I, 1]} \circ t_I \circ X$. EGD can satisfy different fairness constraints (e.g., DemPar or EqOdds). EGD outputs *hard labels* (i.e., a binary assignment to the sensitive attribute).

### 2.3 Attribute Privacy and Inference Attacks

Attribute privacy has been previously defined by Zhang et al. [36] for *databases* using Pufferfish framework [21, 30]. Attribute privacy is the indistinguishability in model predictions for different values of sensitive attributes. Formally, it bounds the ratio of distribution of mechanism output conditioned on different values of sensitive attributes. Their definition was not designed for ML, but we can adopt it for ML by considering distribution of model predictions and adapt it for attribute privacy of individual data records.

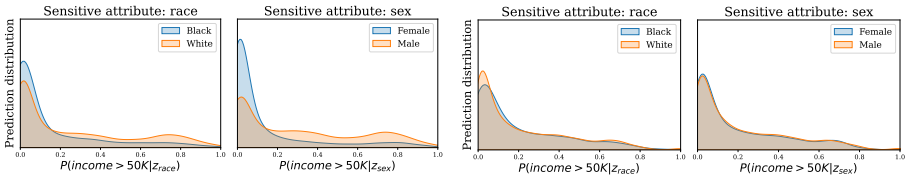
However, this raises a question on *how to verify whether a model satisfies attribute privacy?* To this end, we take inspiration from the literature on auditing differential privacy [28], we can check if a model satisfies differential privacy by distinguishing between models trained on adjacent datasets (referred to as "Differential Privacy Distinguishability"). Similarly, we present *attribute privacy distinguishability* by which we distinguish between model inputs for different values of sensitive attributes. A specific variant of this distinguishing test are AIAs which can be empirically evaluated. AIAs constitute a privacy risk as $\mathcal{A}dv$ learns something about the inputs which would be impossible to learn without access to $f_{trg}$. This differentiates between a privacy risk and simple statistical inference [7]. Hence, we can empirically measure attribute privacy using the resistance to AIAs which exploit distinguishability in $f_{trg}(X)$ for different values of sensitive attributes. Specifically, $f_{trg}$ satisfies attribute privacy if the success of AIA is random guess. Using AIAs, $\mathcal{A}dv$ exploits model observables (e.g., predictions) and background information to infer the specific value of a sensitive attribute corresponding to an input [15, 9, 32, 29, 23, 24]. We assume $\mathcal{A}dv$ has access to auxiliary data $\mathcal{D}_{aux}$ which is sampled from the same distribution as $\mathcal{D}_{tr}$, a standard assumption across all AIAs. Based on $\mathcal{A}dv$'s knowledge, AIAs can be categorized into imputation-based and representation-based attacks.

**Imputation-based attacks** assume $\mathcal{A}dv$ has access to non-sensitive attributes and background information (e.g., marginal prior over sensitive attribute and confusion matrix) in addition to model's predictions. Fredrikson et al. [15], Yeom et al. [32] and Mehnaz et al. [24] assume that $S$ is part of the input of $f_{trg}$ and the targeted data point belongs to $\mathcal{D}_{tr}$. Fredrikson et al.[15] and Mehnaz et al.[24] for a targeted data point, compute $f_{trg}$ for different values of the sensitive attribute to find the most likely one. Yeom et al. [32] predict $S$ using the output of a membership oracle or assuming it follows some distribution. However, these attacks perform no better than data imputation and do not pose an actual privacy risk [19]. Jayaraman and Evans [19] propose a whitebox AIA which is a privacy risk in the setting where $\mathcal{A}dv$ has limited knowledge. We omit this work due to difference in threat model. **Representation-based attacks**, in turn, exploit the distinguishability in model observables for different values of sensitive attributes [29, 9, 23]. For instance, the distribution of $f_{trg} \circ X$ for $S = males$ is different from $S = females$. Song et al. [29] / Mahajan et al. [9] assume that $S$ is not in the input. $\mathcal{A}dv$ only observes $f_{trg} \circ X$. $\mathcal{A}dv$ trains an ML attack model $f_{att}$ to map the output predictions $f_{trg}(X(\Omega))$ to $S(\Omega)$. *Malekzadeh et al. [23]* assume that $\mathcal{A}dv$ can actively introduce a "backdoor" and train $f_{trg}$ to explicitly encode information about $S$ in $f_{trg} \circ X$. We omit comparison with Malekzadeh et al. [23] due to difference in threat model and use Song et al. [29] and Mahajan et al. [9] as our baselines.

**Interactions between privacy and fairness.** Chang et al. [6] show that applying group fairness constraints increases the susceptibility to membership inference attacks, particularly affecting minority subgroups. Further, the definition of individual fairness is a generalization of differential privacy [11]. However, differential privacy and group fairness are at odds indicated by a performance discrepancy between minority and majority subgroups [4, 26, 14]Song et al. [29] consider a setting where the model is split into a local on-device and a remote component. Hence, $\mathcal{A}dv$ has access to some intermediate (censored) representation of the model. They claim that fairness-based censoring of intermediate layer model representation but assuming $\mathcal{A}dv$ has access to non-censored layers. But given access to only censored layers, their AIAs are not successful which supports our observation.

## 3 Problem Statement

Our goal is to comprehensively understand the relation between group fairness and attribute privacy and present tools for further analysis. Intuitively, group fairness ensures the conditional independence of sensitive attributes. This *should be* equivalent to indistinguishability in model predictions. To validate this, we present an illustration using `CENSUS` dataset showing the distribution of output predictions for `Race` and `Sex` in Figure 1. We see that training with fairness constraints results in indistinguishability. Hence, we *conjecture* that group fairness *aligns* with attribute privacy by ensuring indistinguishability in the predictions




(a) Without Fairness Constraints (b) With Fairness Constraints

**Fig. 1.** Group fairness ensures conditional independence of model predictions with sensitive attributes, which is equivalent to indistinguishability in their predictions. Hence, this should satisfy attribute privacy.

for different values of sensitive attributes. However, to effectively validate this, we need to address two other challenges:

**C1** ***Accounting for Class Imbalance***: We evaluate group fairness algorithms against AIAs. If the success of AIAs is a random guess, it indicates that attribute privacy is satisfied. However, none of the current AIAs in literature are effective as they fail to account for real-world datasets with significant class imbalance in sensitive attributes. For instance, the fraction of males and whites in different datasets are 68% and 90% (`CENSUS`), 81% and 51% (`COMPAS`), 53% and 36% (`MEPS`), and 78% and 96% (`LFW`). We have to design AIAs to account for this class imbalance to make them effective.

**C2** ***Accounting for Soft/Hard Labels***: Group fairness algorithms can output either soft labels (i.e., (probability that an input belongs to different classes) or hard labels (i.e., most likely class from soft labels) and we have to design AIAs for both settings.

### 3.1 Threat Model

**$\mathcal{A}dv$'s Knowledge.** We assume a blackbox $\mathcal{A}dv$ with no knowledge of $f_{trg}$'s parameters or architecture. $\mathcal{A}dv$ can query $f_{trg}$ and obtain corresponding predictions as seen in ML as a service. Additionally, $\mathcal{A}dv$ has access to $\mathcal{D}_{aux}$, sampled from the same distribution as $\mathcal{D}_{tr}$, which is split into two disjoint datasets: $\mathcal{D}^{tr}_{aux}$ and $\mathcal{D}^{te}_{aux}$ to design and evaluate the attack respectively. This strong assumption aligns with prior works to favor $\mathcal{A}dv$ [9, 29, 32].

**Attack Methodology.** $\mathcal{A}dv$ first queries $f_{trg}$ using $X'(\omega)$ from $\mathcal{D}^{tr}_{aux}$ to obtain $f_{trg}(X'(\omega))$. $\mathcal{A}dv$ then trains $f_{att}$ to map $f_{trg}(X'(\omega))$ to $S'(\omega)$. Once $f_{att}$ is trained, $\mathcal{A}dv$ performs the attack which is evaluated on inputs from $\mathcal{D}^{te}_{aux}$. We present an overview in Figure 2. Depending on whether $f_{trg}$ outputs hard or soft labels, we identify the following settings: **TM1** (hard labels) $\mathcal{A}dv$ builds a statistic $\hat{S}$ to infer $S$: $\hat{S} = f_{att} \circ \hat{Y}_h \circ X$; **TM2** (soft labels) $\mathcal{A}dv$ builds a statistic $\hat{S}$ to infer $S$: $\hat{S} = 1_{[\upsilon,1]} \circ f_{att} \circ \hat{Y}_s \circ X$. Here, $\upsilon \in [0, 1]$ is a threshold which can be adapted to improve the attack. Note that in both cases, $\mathcal{A}dv$ only uses $f_{trg}$'s outputs as input to $f_{att}$ to infer the value of sensitive attributes.

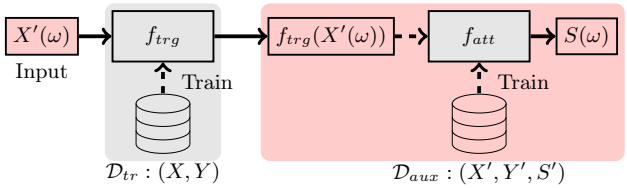



**Fig. 2.** $\mathcal{A}dv$ wants to infer sensitive attributes for an input given its prediction. $\mathcal{A}dv$ trains $f_{att}$ on $\mathcal{D}_{aux}$ to map $f_{trg}(X'(\omega))$ to $S'(\omega)$. Once trained, $\mathcal{A}dv$ only uses $f_{trg}$'s outputs as input to $f_{att}$ to infer sensitive attributes. red indicates accessible by $\mathcal{A}dv$.

### 3.2 Limitations of prior work

Ferry et al. [12] show that sensitive attributes can be inferred from a fair model under the strong assumption that the fairness metrics is known to $\mathcal{A}dv$. However, companies are unlikely to release such proprietary information. We consider a practical threat model where $\mathcal{A}dv$ has no knowledge of $f_{trg}$. Zhang et al. [36] define attribute privacy as indistinguishability across sensitive subgroups using Pufferfish privacy framework but do not not focus on protecting sensitive attributes for individual data records. Further, their definitions and mechanisms to achieve attribute privacy are for databases and not ML models. Finally, they speculate the alignment of group fairness and privacy but without any evaluation. It is not clear how different fairness metrics impact the interaction. Hence, none of the prior works have comprehensively studied this interaction. Our study provides a better understanding, both empirically and theoretically, of the relationship between attribute privacy and group fairness as well as the impact of fairness measures on interactions.

## 4 Interaction: Theoretical Analysis

By theoretically analyzing the bounds on AdaptAIA, we present tools to analyze the impact of different fairness algorithms (e.g., EGD and AdvDebias) and metrics (e.g., DemPar and EqOdds) on the alignment. This was missing in Zhang et al. [36].

### 4.1 EGD

Training a model with EGD outputs hard labels and falls in **TM1**. For EGD, we first consider DemPar in **TM1** and then EqOdds in **TM1** and by consequence in **TM2**, to show how different fairness metrics can impact the overall interaction. We use the definitions of fairness metrics to find an upper bound on the attack accuracy of AIA.

**Theorem 1.** *Maximum attack accuracy achievable by* AIA *in* ***TM1*** *is equal to* $\frac{1}{2}(1 + \text{DemPar-}level\ of\ f_{trg})$.

Hence, we obtain a bound for AIA in **TM1** without any conditions on $f_{trg}$ or datasets. Additionally, we observe that DemPar-*level* $= 0$. Consequently, if $f_{trg}$ satisfies DemPar then no $f_{att}$ will perform better than a random guess. Hence, EGD+DemPar satisfies attribute privacy.

**Theorem 2.** *If $\hat{Y}$ satisfies* EqOdds *for $Y$ and $S$ then the balanced accuracy of* AIA *in **TM1** is $\frac{1}{2}$ iff $Y$ is independent of $S$ or $\hat{Y}$ is independent of $Y$.*

Those two conditions are unlikely to happen in practice. The condition of $Y$ being independent of $S$ was not observed for our datasets. We evaluate $|P(Y = 0|S = 0) - P(Y = 0|S = 1)|$ where a high value indicates $Y$ and $S$ are dependent. For `Race` and `Sex`, we found these values to be 0.05 and 0.27 (`COMPAS`), 0.20 and 0.13 (`CENSUS`) and 0.07 and 0.13 (`MEPS`) respectively. Further, the independence between $\hat{Y}$ and $Y$ means that $f_{trg}$ has random guess utility. Hence, in practice, EqOdds aligns by reducing the risk to AIAs but does not *perfectly align* as seen in DemPar by reducing AIAs to random guessing. ***The choice of fairness metric is important for* EGD *for perfect alignment.*** We now only consider EGD+DemPar in the rest of the paper.

### 4.2 AdvDebias

Training $f_{trg}$ with AdvDebias, outputs soft labels and hence falls in **TM2**. We now show that using AdvDebias bounds the balanced attack accuracy to random guess.

Definition 1 of DemPar can be written synthetically as the following property: $P_{\hat{Y},S} = P_{\hat{Y}} \otimes P_S$. Where $P_{\hat{Y}} \otimes P_S$ is the product measure defined as the unique measure on $\mathcal{P}(\mathcal{Y}) \times \mathcal{P}(\mathcal{S})$ such that $\forall y \in \mathcal{P}(\mathcal{Y}) \forall s \in \mathcal{P}(\mathcal{S}) \quad P_{\hat{Y}} \otimes P_S(y \times s) = P_{\hat{Y}}(y)P_S(s)$. This is equivalent to definition 1 for binary labels and sensitive attribute but more general because when $\hat{Y}$ is not binary as in soft labels, this new definition is well defined.

**Definition 3.** *$\hat{Y}$ satisfies* DemPar *for $S$ if and only if: $P_{\hat{Y},S} = P_{\hat{Y}} \otimes P_S$.*

This definition is the same as the statistical parity introduced for fair regression [2]. Note that we can not derive a quantity similar to DemPar-*level* with this definition but this extended DemPar assures indistinguishably of the sensitive attribute when looking at the soft labels. We have the following theorem:

**Theorem 3.** *The following propositions are equivalent: "$\hat{Y}_s$ is independent of $S$" and "Balanced accuracy of* AIA *in **TM2** is $\frac{1}{2}$"*

These results suggest that AdvDebias, by imposing $P_{\hat{Y},S} = P_{\hat{Y}} \otimes P_S$, reduces susceptibility to AIA in **TM2**. Extended demographic parity, a notion for soft labels implies DemPar for hard labels whatever the threshold is.

# 5 Interaction: Empirical Analysis

## 5.1 Experimental Setup

**Datasets.** We consider four real-world datasets covering different domains: criminal justice (`COMPAS`), income prediction (`CENSUS`), healthcare (`MEPS`), and face recognition (`LFW`), to illustrate the effectiveness of the proposed AIAs. These have been used as benchmarks for privacy [29, 24, 9] and fairness [25]. In all datasets, we consider `Race` and `Sex` as binary sensitive attributes to be inferred. `CENSUS` comprises 30,940 data records with 95 attributes about individuals from 1994 US Census data. The attributes include marital status, education, occupation, job hours per week among others. The classification task is to estimate whether an individual makes an income of 50k per annum. `COMPAS` is used for commercial algorithms by judges and parole officers for estimating the likelihood of a criminal re-offending using seven attributes. The classification task is whether a criminal will re-offend or not, and contains 6,172 criminal defendants in Florida. `MEPS` contains 15,830 records of different patients using medical services by capturing the trips made to clinics and hospitals. The classification task is to predict the utilization of medical resources as 'High' if the sum of the number of office based visits, outpatient visits, ER visits, inpatient nights and home health visits, is greater than ten. `LFW` is an example for face recognition systems and has 8,212 images of people with the classification task to predict whether their age is >35 years.

We use 80% of the dataset as $\mathcal{D}_{tr}$ and the remaining 20% for $\mathcal{D}_{te}$. We use $\mathcal{D}_{te}$ as $\mathcal{D}_{aux}$ and ensure that the distribution of $S$ is uniform between them. We use 80% of $\mathcal{D}_{aux}$ for training $f_{att}$ and 20% for evaluation of the attack.

**Model Architectures:** We use neural networks with four hidden layers with the following dimensions: [32, 32, 32, 32] and ReLU activation functions. We train the models using cross validation where each split is done five times without any overlap. $f_{trg}$ is trained and evaluated five times and $f_{att}$ is trained and validated ten times. We check for statistical significance for the results (i.e., differences in results are significant if p-value $<0.05$). Due to their widespread use, we indicate results for neural networks. But, we also evaluated using random forest and we omit them as the results are similar to neural networks.

**Metrics.** We use standard classification *accuracy* between prediction and ground truth labels to evaluate $f_{trg}$'s *utility*. To evaluate AIA success, we note that accuracy is misleading with class imbalance. Hence, we use *balanced accuracy* which is the average of the proportion of correct predictions of each class of the sensitive attribute individually: $\frac{1}{2}(P(\hat{S}=0|S=0)+P(\hat{S}=1|S=1))$. For fairness, we use DemPar-*level* given by $|P(\hat{Y}=0|S=0)-P(\hat{Y}=0|S=1)|$ and a value close to zero indicates fairness.

**Baselines.** To evaluate the impact of group fairness on attribute privacy, we compare the attack success of AdaptAIA with and without using group fairness (the former case is referred to as `Empirical` while the latter is referred to as `Baseline`). For classifiers with fairness algorithms that output hard labels, we also indicate the theoretical bound (referred to as `Theoretical`).

### 5.2 AdaptAIA: Adaptive Thresholding

We first design AdaptAIA to address the two challenges (**C1** and **C2**) from Section 3. AdaptAIA uses an adaptive threshold over $f_{att}$ to account for class imbalance in sensitive attributes, typical of real-world datasets [16, 27]. Also, AdaptAIA is designed for both soft (AdaptAIA-S) and hard labels (AdaptAIA-H).

**AdaptAIA-S.** Recall from Section 3 that real-world datasets have significant class imbalance in sensitive attributes. We conjecture that this skews $f_{trg}$'s predictions, and none of the existing AIAs are effective. Hence, we have to adapt the threshold over $f_{att}$'s soft labels to correctly infer sensitive attributes instead of the default threshold of 0.5 as in prior AIAs [29, 9]. The use of an adaptive threshold has shown to improve the utility of a classifier [16, 27, 5]. However, none of the prior AIAs account for this.

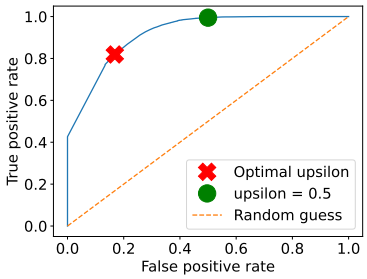



**Fig. 3.** Using $\upsilon^*$ can lower FPR to infer sensitive attributes.

$\mathcal{A}dv$ optimizes the threshold $\upsilon^*$ on $\mathcal{D}^{tr}_{aux}$ which is later fixed during the attack on $\mathcal{D}^{te}_{aux}$. We compute $\upsilon^*$ to balance TPR and FPR. Ideally, a perfect attack would result in no FPR and perfect TPR and $\mathcal{A}dv$'s goal is to design an AIA to approach this optimal value. Formally, $\upsilon^* \in [0, 1]$ where $\upsilon^* = \text{argmin}_\upsilon(1 - TPR_\upsilon)^2 + FPR^2_\upsilon$. For illustration purposes, we plot the ROC curve to infer `Race` in Figure 3. We observe that $\upsilon^*$ results in lower FPR resulting in a more confident attack and does not correspond to the default threshold of 0.5 used in prior AIAs [29, 9].

**AdaptAIA-H.** In **TM1**, it is not necessary to train $f_{att}$. Instead, we consider a set of functions from $\{0, 1\}$ to $\{0, 1\}$ containing four elements: $x \mapsto 0$, $x \mapsto x$, $x \mapsto 1 - x$, and $x \mapsto 1$. Instead of finding $\upsilon^*$ as in AdaptAIA-S, here, we optimize the attack by finding the function which gives the best balanced accuracy on $\mathcal{D}^{tr}_{aux}$. This function is fixed during the attack on $\mathcal{D}^{te}_{aux}$.

**Evaluation.** For **TM1**, we use a neural network over hard labels as our baseline and then compare with AdaptAIA-H. For **TM2**, we consider the attacks by Song et al. [29]/Mahajan et al. [9] as the state-of-the-art baselines which use the default $\upsilon = 0.5$ over $f_{att}$'s predictions. We then compare this with AdaptAIA-S.

**Table 1.** AdaptAIA-H and AdaptAIA-S outperform their respective baselines. We report attack accuracy over ten runs.

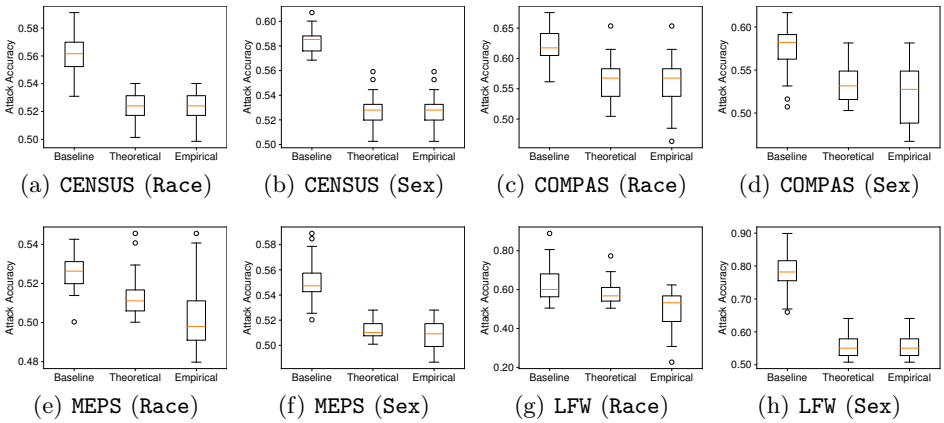
| Dataset | TM1 Baseline ($\upsilon$=0.50) Race \| Sex | TM1 AdaptAIA-H Race \| Sex |
|---|---|---|
| CENSUS | 0.50 ± 0.00 \| 0.50 ± 0.00 | **0.56 ± 0.01** \| **0.58 ± 0.01** |
| COMPAS | **0.62 ± 0.03** \| 0.50 ± 0.00 | **0.62 ± 0.03** \| **0.57 ± 0.03** |
| MEPS | 0.51 ± 0.01 \| **0.55 ± 0.02** | **0.53 ± 0.01** \| **0.55 ± 0.01** |
| LFW | 0.59 ± 0.00 \| 0.64 ± 0.15 | **0.61 ± 0.11** \| **0.78 ± 0.05** |
| **Dataset** | **TM2 Baseline ($\upsilon$=0.50) Race \| Sex** | **TM2 AdaptAIA-S Race \| Sex** |
| CENSUS | 0.50 ± 0.02 \| 0.56 ± 0.04 | **0.61 ± 0.02** \| **0.68 ± 0.01** |
| COMPAS | **0.62 ± 0.03** \| 0.50 ± 0.00 | **0.62 ± 0.03** \| **0.57 ± 0.03** |
| MEPS | 0.52 ± 0.02 \| 0.55 ± 0.02 | **0.60 ± 0.02** \| **0.62 ± 0.02** |
| LFW | 0.50 ± 0.10 \| **0.77 ± 0.07** | **0.61 ± 0.10** \| **0.79 ± 0.05** |

We present the comparison of the baseline with AdaptAIA-H and AdaptAIA-S in Table 1. For all datasets, we see that AdaptAIA-H and AdaptAIA-S are significantly better on average than their baselines. Having successfully addressed 3 and 3, we use them to evaluate the alignment with two well-established group fairness algorithms: EGD and AdvDebias.

### 5.3 EGD against AdaptAIA-H

To evaluate alignment of EGD+DemPar with attribute privacy, we compare the difference in attack accuracy with and without EGD+DemPar against AdaptAIA-H (Figure 4). AdaptAIA-H has significantly lower effectiveness (approaching random guessing, 50%) when utilizing EGD+DemPar compared to the `Baseline`. Additionally, we note that the theoretical bound on attack accuracy matches with the empirical attack accuracy. The theoretical accuracy is equal to the empirical accuracy when the values are $> \frac{1}{2}$. But DemPar-*level* $\geq 0$ implies that $(1 +$ DemPar-*level*$) \geq \frac{1}{2}$. Hence, we observe that the theoretical accuracy is not equal to the experimental when $f_{att}$'s attack accuracy is random guess (under $\frac{1}{2}$). This happens when $f_{trg}$ nearly follows DemPar where $\mathcal{A}dv$'s $f_{att}$ is optimal on $\mathcal{D}_{tr}$ but worse than random guess for $\mathcal{D}_{te}$.

**Sanity Check for Fairness and Trade-off with Utility.** We also confirm if EGD+DemPar indeed results in a fair model and measure the corresponding decrease in utility. For fairness, we report DemPar-*level* with and without EGD+DemPar. For `Race`, DemPar-*level* decreases from 0.12 to 0.01 (`CENSUS`), 0.24 to 0.13 (`COMPAS`), 0.26 to 0.16 (`LFW`), and 0.05 to 0.01 (`MEPS`). For `Sex`, DemPar-*level* decreases from 0.17 to 0.02 (`CENSUS`), 0.14 to 0.07 (`COMPAS`), 0.57 to 0.11 (`LFW`), and 0.10 to 0.00 (`MEPS`). Hence, EGD+DemPar is effective in achieving group fairness indicated by DemPar-*level* closer to zero on using EGD+DemPar. For utility, we report the difference in $f_{trg}$'s accuracy on using EGD+DemPar: 15% (`CENSUS`), 5% (`MEPS`), ∼13% (`COMPAS`), and `LFW` (∼6%). This trade-off with the utility of $f_{trg}$ is inherent to training with EGD [34, 31].




(a) `CENSUS` (`Race`) (b) `CENSUS` (`Sex`) (c) `COMPAS` (`Race`) (d) `COMPAS` (`Sex`)

(e) `MEPS` (`Race`) (f) `MEPS` (`Sex`) (g) `LFW` (`Race`) (h) `LFW` (`Sex`)

**Fig. 4.** For AdaptAIA-H, we observe that EGD reduces the attack accuracy to random guess (∼50%).

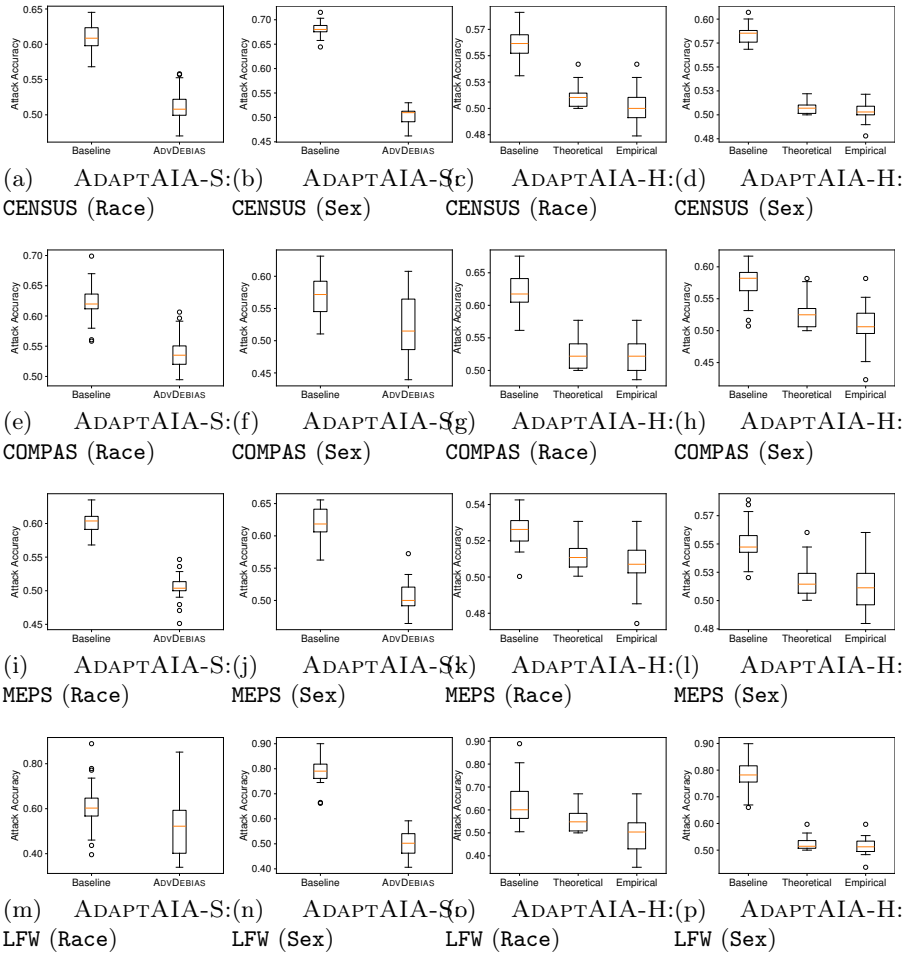
### 5.4 AdvDebias against AdaptAIA

To evaluate the alignment of the fairness constraint imposed by AdvDebias with attribute privacy, we compare the attack success of AdaptAIA-S and AdaptAIA-H on $f_{trg}$ with and without AdvDebias (Figure 5). Results show that for all datasets AdvDebias reduces the attack accuracy close to random guess (i.e., 50%) compared to `Baseline`. This suggests that AdvDebias is aligned with attribute privacy, and acts as a defense against AdaptAIA. Further, the theoretical bound for AdaptAIA-H for DemPar from Section 4.2 matches with the empirical results.

**Sanity Check for Fairness and Trade-off with Utility.** We confirm that the resulting model is indeed fair and measure the corresponding decrease in utility. For fairness, we measure DemPar-*level* with and without training $f_{trg}$ with AdvDebias. $f_{trg}$ with AdvDebias has significantly lower DemPar-*level* for both `Race` and `Sex` which is closer to zero as compared to `Baseline`. For `Race`, DemPar-*level* decreases from 0.12 to 0.02 (`CENSUS`), 0.24 to 0.05 (`COMPAS`), 0.26 to 0.12 (`LFW`), and 0.05 to 0.02 (`MEPS`). For `Sex`, DemPar-*level* decreases from 0.17 to 0.02 (`CENSUS`), 0.15 to 0.05 (`COMPAS`), 0.57 to 0.05 (`LFW`), and 0.10 to 0.04 (`MEPS`). Hence, AdvDebias is effective for group fairness. For utility, we report the difference in $f_{trg}$'s accuracy with AdvDebias: ∼13% (`CENSUS`), ∼17% (`COMPAS`), ∼8% (`MEPS`), and ∼16% (`LFW`). Hence, there is a significant decrease in utility but this trade-off with utility is inherent to AdvDebias [35, 37].

## 6 Conclusion

This paper shows that there are no conflicts between group fairness and the specific notion of attribute privacy, which is lacking in the literature. Specifically, through an extensive empirical evaluation and theoretical guarantees, we show




(a) AdaptAIA-S: CENSUS (Race) (b) AdaptAIA-S: CENSUS (Sex) (c) AdaptAIA-H: CENSUS (Race) (d) AdaptAIA-H: CENSUS (Sex)

(e) AdaptAIA-S: COMPAS (Race) (f) AdaptAIA-S: COMPAS (Sex) (g) AdaptAIA-H: COMPAS (Race) (h) AdaptAIA-H: COMPAS (Sex)

(i) AdaptAIA-S: MEPS (Race) (j) AdaptAIA-S: MEPS (Sex) (k) AdaptAIA-H: MEPS (Race) (l) AdaptAIA-H: MEPS (Sex)

(m) AdaptAIA-S: LFW (Race) (n) AdaptAIA-S: LFW (Sex) (o) AdaptAIA-H: LFW (Race) (p) AdaptAIA-H: LFW (Sex)

**Fig. 5.** For both AdaptAIA-S and AdaptAIA-H, AdvDebias reduces the attack accuracy to random guess (∼50%). For AdaptAIA-H, the theoretical bound on attack accuracy matches with the empirical results.

that group fairness imposed through the use of AdvDebias and EGD satisfying DemPar is aligned with attribute privacy. This alignment means that ensuring group fairness also ensures a protection against the attribute inference attack. However, ensuring fairness remains at the cost of model utility. To perform our extensive evaluation, we also propose new AIAs which outperform prior works. Finally, we also theoretically analyze how different fairness metrics bound AIAs.

We focus on binary sensitive attributes. However, for AdaptAIA-S, the $f_{att}$ can be also trained to learn to infer non-binary attributes. For AdaptAIA-H, while considering non-binary attributes is reasonable for small values of classes and attributes, efficiently finding functions for large values is left as future work.

**Acknowledgment:** This work has been supported by the ANR 22-PECY-0002 IPOP (Interdisciplinary Project on Privacy) project of the Cybersecurity PEPR, and the Trusty-IA project supported by the Auvergne Rhône-Alpes region.

## References

1. Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: International Conference on Machine Learning. vol. 80, pp. 60–69 (2018)
2. Agarwal, A., Dudík, M., Wu, Z.S.: Fair regression: Quantitative definitions and reduction-based algorithms (2019)
3. Alikhademi, K., Drobina, E., Prioleau, D., Richardson, B., Purves, D., Gilbert, J.E.: A review of predictive policing from the perspective of fairness. Artif. Intell. Law **30**(1), 1–17 (mar 2022)
4. Bagdasaryan, E., Poursaeed, O., Shmatikov, V.: Differential privacy has disparate impact on model accuracy. In: Advances in Neural Information Processing Systems, pp. 15479–15488 (2019)
5. Brownlee, J.: A Gentle Introduction to Threshold-Moving for Imbalanced Classification - MachineLearningMastery.com, [Accessed 31-08-2023]
6. Chang, H., Shokri, R.: On the privacy risks of algorithmic fairness. European Security & Privacy pp. 292–303 (2021)
7. Cormode, G.: Personal privacy vs population privacy: Learning to attack anonymization. In: International Conference on Knowledge Discovery and Data Mining. p. 1253–1261 (2011)
8. Cummings, R., Gupta, V., Kimpara, D., Morgenstern, J.: On the compatibility of privacy and fairness. In: Conference on User Modeling, Adaptation and Personalization. p. 309–315 (2019)
9. Divyat Mahajan, Shruti Tople, A.S.: Does learning stable features provide privacy benefits for machine learning models? In: NeurIPS PPML Workshop (2020)
10. Duddu, V., Szyller, S., Asokan, N.: Sok: Unintended interactions among machine learning defenses and risks. arXiv preprint arXiv:2312.04542 (2023)
11. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Innovations in Theoretical Computer Science. p. 214–226 (2012)
12. Ferry, J., Aivodji, U., Gambs, S., Huguet, M., Siala, M.: Exploiting fairness to enhance sensitive attributes reconstruction. In: Conference on Secure and Trustworthy Machine Learning. pp. 18–41 (feb 2023)
13. Fioretto, F., Tran, C., Van Hentenryck, P., Zhu, K.: Differential privacy and fairness in decisions and learning tasks: A survey. In: International Joint Conference on Artificial Intelligence. pp. 5470–5477 (7 2022)
14. Fioretto, F., Tran, C., Van Hentenryck, P., Zhu, K.: Differential privacy and fairness in decisions and learning tasks: A survey. arXiv preprint arXiv:2202.08187 (2022)
15. Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., Ristenpart, T.: Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In: USENIX Conference on Security Symposium. p. 17–32 (2014)
16. Guo, L.Z., Li, Y.F.: Class-imbalanced semi-supervised learning with adaptive thresholding. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) International Conference on Machine Learning. vol. 162, pp. 8082–8094 (17–23 Jul 2022)
17. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems. p. 3323–3331 (2016)

18. House, W.: Guidance for regulation of artificial intelligence applications. In: Memorandum For The Heads Of Executive Departments And Agencies (2020)
19. Jayaraman, B., Evans, D.: Are attribute inference attacks just imputation? arXiv preprint arXiv:2209.01292 (2022)
20. Kazim, E., Denny, D.M.T., Koshiyama, A.: AI auditing and impact assessment: according to the uk information commissioner's office. AI and Ethics (Feb 2021)
21. Kifer, D., Machanavajjhala, A.: Pufferfish: A framework for mathematical privacy definitions. Trans. Database Syst. **39**(1) (jan 2014)
22. Louppe, G., Kagan, M., Cranmer, K.: Learning to pivot with adversarial networks. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
23. Malekzadeh, M., Borovykh, A., Gündüz, D.: Honest-but-curious nets: Sensitive attributes of private inputs can be secretly coded into the classifiers' outputs. In: Conference on Computer and Communications Security. p. 825–844 (2021)
24. Mehnaz, S., Dibbo, S.V., Kabir, E., Li, N., Bertino, E.: Are your sensitive attributes private? novel model inversion attribute inference attacks on classification models. In: USENIX Security Symposium. pp. 4579–4596 (2022)
25. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. Comput. Surv. **54**(6) (jul 2021)
26. Pujol, D., McKenna, R., Kuppam, S., Hay, M., Machanavajjhala, A., Miklau, G.: Fair decision making using privacy-protected data. In: Fairness, Accountability, and Transparency. p. 189–199 (2020)
27. Rajaraman, S., Ganesan, P., Antani, S.K.: Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks. PLoS ONE **17** (2021)
28. Salem, A., Cherubin, G., Evans, D., Köpf, B., Paverd, A., Suri, A., Tople, S., Zanella-Béguelin, S.: Sok: Let the privacy games begin! a unified treatment of data inference privacy in machine learning. In: Security & Privacy. pp. 327–345 (2023)
29. Song, C., Shmatikov, V.: Overlearning reveals sensitive attributes. In: International Conference on Learning Representations (2020)
30. Song, S., Wang, Y., Chaudhuri, K.: Pufferfish privacy mechanisms for correlated data. In: International Conference on Management of Data. pp. 1291–1306 (2017)
31. Veldanda, A.K., Brugere, I., Chen, J., Dutta, S., Mishler, A., Garg, S.: Fairness via in-processing in the over-parameterized regime: A cautionary tale with mindiff loss. Transactions on Machine Learning Research (2023)
32. Yeom, S., Giacomelli, I., Fredrikson, M., Jha, S.: Privacy risk in machine learning: Analyzing the connection to overfitting. In: Computer Security Foundations Symposium. pp. 268–282 (2018)
33. Zafar, M.B., Valera, I., Gomez-Rodriguez, M., Gummadi, K.P.: Fairness constraints: A flexible approach for fair classification. Journal of Machine Learning Research **20**(75), 1–42 (2019)
34. Zhai, R., Dan, C., Kolter, Z., Ravikumar, P.: Understanding why generalized reweighting does not improve over erm (2023)
35. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Conference on AI, Ethics, and Society. p. 335–340 (2018)
36. Zhang, W., Ohrimenko, O., Cummings, R.: Attribute privacy: Framework and mechanisms. In: Fairness, Accountability, and Transparency. p. 757–766 (2022)
37. Zhao, H., Chi, J., Tian, Y., Gordon, G.J.: Trade-offs and guarantees of adversarial representation learning for information obfuscation. In: Advances in Neural Information Processing Systems. vol. 33, pp. 9485–9496 (2020)