Using Copies to Remove Sensitive Data: A Case Study on Fair Superhero Alignment Prediction

Irene Unceta[1,2][0000−0002−7422−1493], Jordi Nin[1,2][0000−0002−9659−2762], and Oriol Pujol[2][0000−0001−7573−009X]

[1] BBVA Data & Analytics, Barcelona, Spain
{irene.unceta, jordi.nin}@bbvadata.com
[2] Dept. of Mathematics and Computer Science
Universitat de Barcelona, Spain
{irene.unceta, jordi.nin, oriol_pujol}@ub.edu

Abstract. Ensuring classification models are fair with respect to sensitive data attributes is a crucial task when applying machine learning models to real-world problems. Particularly in company production environments, where the decision output by models may have a direct impact on individuals and predictive performance should be maintained over time. In this article, build upon [17], we propose copies as a technique to mitigate the bias of trained algorithms in circumstances where the original data is not accessible and/or the models cannot be re-trained. In particular, we explore a simple methodology to build copies that replicate the learned decision behavior in the absence of sensitive attributes. We validate this methodology in the low-sensitive problem of superhero alignment. We demonstrate that this naïve approach to bias reduction is feasible in this problem and argue that copies can be further exploited to embed models with desiderata such as fair learning.

Keywords: Fairness · Superhero alignment · Bias reduction · Copying classifiers.

1 Introduction

Machine learning is rapidly infiltrating critical areas of society that have a substantial impact on people's lives. From financial and insurance markets to medicine, citizen security or the criminal justice system, the tendency has prevailed in recent years to devolve decision making to machine learning models. This tendency is deeply rooted in the idea that algorithms provide an objective approach to social problems, as a reliable alternative to human cognitive biases.

However, while algorithms may escape prejudices, the data with which they are trained does not. Algorithms can only be as good as the data they are trained with and data is often imperfect [8]. Machine learning models that learn from labeled examples are susceptible to inheriting biases existing in the training data. Indeed, they have been shown to reproduce existing patterns of discrimination [4, 12]. So much so that algorithms are often biased against people with certain

protected attributes like race [3, 6, 14, 15], gender [5, 7] or sexual orientation [11]. Studies on analogy generation using word embeddings have demonstrated that the popular Word2Vec space encodes gender biases that are potentially propagated to systems based on this technology [5]. Similarly, machines trained to learn word associations from written texts have been shown to display problematic attitudes towards race or gender [7]. Associations between female names and family or male names and career are particularly worrying consequences of this result. Besides these findings, examples of significant racial disparities in commercial software have also proliferated over the last years.

In light of these findings many works have studied how to create fairer algorithms [4, 9, 13] as well as to benchmark discrimination in various contexts. Fairness-aware learning has, as a matter of fact, received considerable attention in the machine learning community of late, with most solutions being aimed at introducing new formal metrics for fairness and ensuring that classifiers satisfy the desired levels of equity under such definitions.

Solutions to this problem often come in two types. In the first case, an exhaustive data preprocessing removes the ability to distinguish between group membership by getting rid of the sensible information in the training data [10]. This amounts to removing the sensitive attributes themselves, but also to ensuring no residual information is encoded by the remaining data. While simple, this approach often succeeds in repairing the original disparity. In the second case, unfairness is removed by adding corrective terms to the optimization function. A fairness metric [9, 13] is defined and incorporated to the training algorithm. Initially biased models are therefore re-trained ensuring that the fairness measure is optimized together with the defined classification loss.

In this article we propose the use of copies as a technique to mitigate the bias of trained algorithms in circumstances where the original data is not accessible and/or the models cannot be re-trained. Copying [17] corresponds to the problem of building a machine learning model that replicates the decision behavior of another. This process not only reproduces the target decision function, but it may also be used to endow the considered classifier with new characteristics, such as interpretability, online learning or equity features. The use of copies has already been shown to improve model accuracy when ensuring decomposability of attributes in financial production environments where more explainable machine learning models are desirable [16].

Notably, in this paper we explore a potential use of copies in the context of fair prediction. In the simplest scenario, we use copies to remove sensitive attributes from the original classifier while maintaining its performance and reducing the unfairness in the resulting predictions. Further, we argue that desiderata such as equity of learning could be directly imposed upon copies to obtain more sophisticated solutions to the problem of fairness.

We validate our approach in a case study, where we obtain a new decision function based in copies from which the protected features are absent. Due to the generally sensible nature of data in this kind of studies, we carry our experiments using a proxy dataset. In particular, we use the superhero dataset [2], which con-

tains socio-demografic data including gender and race, as well as personal traits and features in the form of superpowers for an extensive list of superheroes and villains. We use this data to predict superhero alignment. Nonetheless, the methodology proposed in this article is readily applicable to other real datasets that satisfy the same constraints (detailed in the experimental settings and discussion of the results). The main contributions of this article are the following:

- We introduce copies as a promising methodology for reducing unfairness in classification models.
- This methodology is agnostic to the internals of the classifiers and can be used in any classification setting where the model is considered as a *black-box*.
- Furthermore, we show that we can reduce the prediction bias even when original data is not available.

The rest of this paper is organized as follows. First, *Section 2* presents a case study using the superhero dataset. *Section 3* describes our proposed methodological approach. In *Section 4*, we carry out a set of experiments and in *Section 5* we describe the results that empirically validate our theoretical proposal. Finally, the paper ends with our conclusions and future work in *Section 6*.

2 Case study

In the following sections we explore how to reduce the bias inherited by a machine learning classifier which has been already trained using sensitive information and which cannot be modified. We do so by means of a fictitious example that nonetheless represents a use case common to many real scenarios. In particular, we use the publicly accessible superhero dataset, which contains information about a few hundred superheroes in the literature, including their physical attributes, powers and alignment. We choose this dataset in order to avoid using real sensitive data, as well as to enable an in-depth study of how the mechanism affects the different variables and instances.

This dataset serves as a good proxy to many real problems where data contains sensitive information. Among the many attributes in the superhero dataset, there are those that account for protected group features. This is the case, for example, of *gender* and *race*. Without the appropriate control, models trained on these attributes can lead to an unfair decision system. For this case study, we assume a classifier has been trained using both *gender* and *race* attributes and that it cannot be modified or re-trained to correct for any bias.

There exists many situations in which a new training may not be advisable, or even possible. This is the case, for example, of company production environments, where the predictive performance of models should be maintained in time. Another situation in which a new training is not an option is when the original training data is not available. This could either be because the data has been lost or because it is subject to privacy constrains or because the server where the data is hosted is not accessible any more. Whatever the cause of this

lack of availability, the fact that the original data points are unknown, makes a new training impossible.

Under these circumstances we propose a methodology to build a copy of the trained model that is able to retain its predictive accuracy and from which we can remove the protected data attributes. Copies are new classifiers built to replicate the decision behavior of their target models. When copying, we can transfer the attributes of the original model to the copy, while at the same time including new characteristics during the process.

3 Methodological proposal

In this section we describe our approach to mitigate the bias induced by the existence of protected data attributes in superhero alignment prediction models. We first present the copying methodology and then describe how this methodology can be exploited to remove protected attributes from copies. Note that the full theoretical background for copying is developed in [17]. We refer the reader to this reference for a full description of the different elements and processes that come into play. We here only provide an overview of this framework.

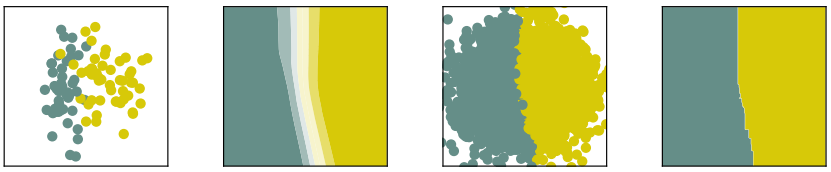
3.1 Copying machine learning classifiers

Let us assume a set of data pairs $\boldsymbol{X} = \{\boldsymbol{x}_i, t_i\}$ for $i = 1, ..., M$, where the $\boldsymbol{x}_i$ are d-dimensional data points and t_i their corresponding labels. In the case of a binary classification problem, we assume these labels to be such that $t_i \in \{0, 1\}$. We define a classifier as a function $f : \boldsymbol{x} \rightarrow \boldsymbol{t}$ from input instances to targets. Under the copying framework, we refer to the set $\boldsymbol{X}$ as the *original dataset* and define the *original model*, $f_{\mathcal{O}}$, as a classifier trained using this data. Copying then corresponds to the problem of building a new model $f_{\mathcal{C}}(\theta)$, parameterized by θ, such that it replicates the behavior of $f_{\mathcal{O}}$.

In order to build this new model, we do not exploit the original training data $\boldsymbol{X}$. Instead, we refer to the original decision function $f_{\mathcal{O}}$. To do this we need to introduce a set of labelled pairs $\boldsymbol{Z} = \{(\boldsymbol{z}_j, y_j)\}$ for $j = 1, ..., N$, where $\{\boldsymbol{z}_1, ..., \boldsymbol{z}_N\}$ are artificially generated data points and $\{y_1, ..., y_N\}$ their labels predicted by the original, so that $y_j = f_{\mathcal{O}}(\boldsymbol{z}_j)$. We refer to $\boldsymbol{Z}$ as the *synthetic dataset* and use it to access the information in $f_{\mathcal{O}}$ through its prediction $f_{\mathcal{O}}(\boldsymbol{z}_j)$ for any given sample. The problem of copying can then be written as

$$\theta^* = \arg\max_{\theta} \int_{\boldsymbol{z} \sim P_Z} P(\theta | f_{\mathcal{O}}(\boldsymbol{z})) dP_Z, \tag{1}$$

Under the empirical risk minimization framework we can cast the equation above into a dual optimization problem where we simultaneously optimize the model parameters θ, the synthetic dataset $\boldsymbol{Z}$ and a probability distribution P_Z over the sampling space. Given a defined empirical loss $R_{emp}(f_{\mathcal{C}}, f_{\mathcal{O}})$ we can define such a problem as



(a) (b) (c) (d)

Fig. 1. Example plots for (a) the *original dataset*, (b) the *original decision function*, (c) the *synthetic dataset* and (d) the *copy decision function*.

$$\begin{aligned} \underset{\theta, \boldsymbol{Z}}{\text{minimize}} \quad & \Omega(\theta) \\ \text{subject to} \quad & \|R^{P_Z}_{emp}(f_{\mathcal{C}}, f_{\mathcal{O}}) - R^{P_Z}_{emp}(f^{\dagger}_{\mathcal{C}}, f_{\mathcal{O}})\| < \epsilon, \end{aligned} \tag{2}$$

for ϵ a defined tolerance, $\Omega(\theta)$ a measure of the capacity of the copy model and $f^{\dagger}_{\mathcal{C}}$ the copy model obtained as a solution to the corresponding unconstrained problem. Roughly speaking, the capacity of a model describes how complex a relationship it can model. A direct way to estimate the capacity of a model is to count the number of parameters. The more parameters, the higher the capacity in general, although this rule can be wrong in many situations.

In this article we restrict to the simplest approach to this problem, the *single-pass copy* [17]. We cast the simultaneous optimization problem into one where we only use only a single iteration of an alternating projection optimization scheme. Thus, we effectively split the problem in two independent sub-problems. We first find an optimal set of synthetic data points $\boldsymbol{Z}^*$ and then optimize for the copy parameters θ^*.

In Figure 1 we show an example of the different steps during a single-pass copy. The data points in Figure 1(a) represent a randomly generated binary classification dataset. We learn this data using a multilayer perceptron that outputs the decision function displayed in Figure 1(b). We sample the original attribute domain following a random normal distribution and label the resulting data points according to the predictions of the original classifier. The resulting synthetic dataset is shown in Figure 1(c). Finally, we fit this data using a decision tree classifier. The decision function learned by this model, the form of which replicates that of the original classifier, is shown in Figure 1(d).

3.2 Using copies to remove sensitive data

When copying, we can transfer model features from one model to another [17, 16]. For one thing, we can ensure the original accuracy is retained by building a copy that replicates the original decision behavior to a high degree of fidelity.

Moreover, this can be done by imposing the new characteristics such as considering only self-explanatory features or removing biased attributes upon the copy. In doing so, we can understand copying as a mechanism to correct pre-existing biases.

Our proposed solution is of particular importance when models are trained in the absence of an external auditing or which are now subject to a regulation that they were previously excluded from. Existing techniques to remove bias or correct unfairness often rely in a new training of the model: change the optimization function to ensure certain constraints are satisfied or remove sensitive variables. However, this is not always possible. For example, in models in production we may not have the specifics or even access to the original data.

In the simplest approach, in order to ensure prediction equity, we require the model not to have access to sensitive data with the additional constraint that this information not be leaked through the remaining attributes. In the copying framework this can be accomplished by changing the input space of the copy, characterized by $\boldsymbol{Z}$. During the synthetic sample generation process, we can remove the sensitive data attributes that were present in the original training dataset $\boldsymbol{X}$. In removing this attributes, we ensure the copy has no access no the sensitive information. Because the copying model replicates the behavior of the original black-box model one expects the copy to maximize its performance even in the lack of the sensitive data.

4 Experiments

We use superheroes dataset [1], which describes characteristics such as demographics, powers, physical attributes and studio of origin of every superhero and villain in SuperHeroDb [2]. In what follows we describe the experimental set up, including the original dataset preprocessing, original model training, synthetic sample generation and, finally, the copy model building; we well as the metrics we use to evaluate our results.

Original dataset The dataset contains information about 177 attributes for 660 superheroes. We remove all entries with an unknown alignment label. We also discard all attributes for which the number of missing values exceeds the 20% of the total size of the dataset. For the remaining columns, we set all missing values to the median for numerical attributes and to *other*, in the case of categorical variables. For the latter, we also group under the general category *other* all values with a count below a certain threshold. We set this threshold to 1% for variable *eye color* and to 10% for *publisher*. In the case of *race* we group all entries under the general categories *human*, *mutant*, *robot*, *extraterrestial* and *other*. Additionally, we impose that for superhero powers the sum be above the 1% of the total number of entries. Finally, we convert nominal attributes to numerical by means of one-hot encoding and re-scale all variables to zero-mean and unit variance. The resulting dataset contains 135 variables. We use this data to define a binary classification problem choosing superhero alignment as

the target attribute. We label as *good* all superheroes marked as so and identify as *bad* all other entries, including those labelled as bad, neutral or unknown. In other words, we assume every superhero not explicitly labelled as good to be bad. The distribution of target labels is slightly unbalanced, with a third of the dataset set to the positive label, *good*, and the remaining two thirds labelled as *bad*. We split this data into stratified 80/20 training and test sets.

Original model We use the resulting binary classification dataset to train a fully-connected artificial neural network with 4 hidden layers, each consisting of 128, 64, 32 and 16 neurons. We use $SeLu$ activations, a drop-out of 0.6 and a softmax cross entropy loss optimized using $Adam$ optimizer for a learning rate equal to $1e-3$. We train the network from a random initialization of weights and without any pretraining. We use balanced batches with a fixed size of 32. We assume this model as as baseline biased model.

Synthetic sample generation process We generate the synthetic dataset using different sampling strategies for numerical and categorical attributes. For the first, we directly generate synthetic data points in the original attribute domain by sampling a random normal distribution with mean 0 and standard deviation 1. In the case of categorical variables, we sample uniformly at random the original category set. When generating new synthetic values for superhero powers, we ensure that the relationships among the original data attributes is kept. To do so, we sample uniformly at random the *n_powers* variable and then randomly distribute the total count over the individual power attributes. Following the guidelines in [17], we generate a balanced synthetic dataset consisting of $1e6$ labelled data pairs, from which we extract the two problematic attributes. We use this dataset as a training set.

Copy model We use the lower-dimensional synthetic dataset to learn a new artificial neural network with the same architecture and training protocol as that of the original model, with a fixed batch size to 512 and a drop-out rate of 0.9.

Performance metrics We measure the extent to which the copy replicates the original decision behavior following the metrics described in [17]. In particular, we report the empirical fidelity error over both the synthetic dataset, $R_{\mathcal{F}}^{\boldsymbol{Z}}$, and the original dataset, $R_{\mathcal{F}}^{\boldsymbol{X}}$; and the copy accuracy, $\mathcal{A}_{\mathcal{C}}$. The first two give a level of disagreement between original and copy over a common set of data points, while the latter corresponds to the generalization performance of the copy in the original data environment. Additionally, we measure the presence of bias by evaluating the difference in accuracy over the *gender* and *race* groups.

Validation We run each experiment 10 times and evaluate the performance of copies using test sets comprised of $1e6$ synthetic points. For validation purposes, we assume both the original accuracy and the original dataset to be known in all cases and report average metrics over all repetitions.

5 Discussion of results

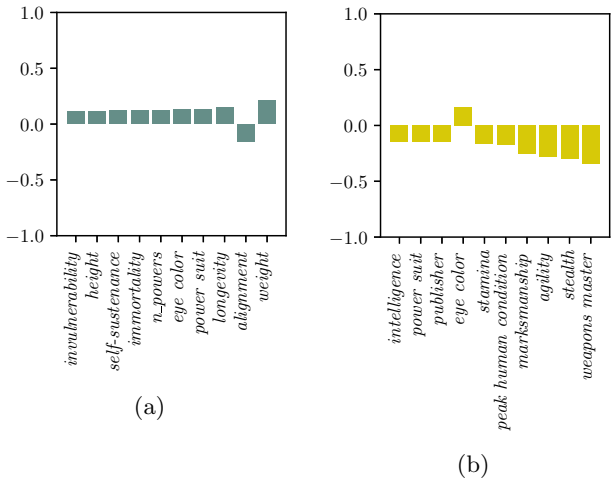


Fig. 2. Top ten ranked attributes in terms of their one-to-one correlation coefficient with (a) *gender* and (b) *race*. The ranking is computed taking the absolute value.

In what follows we discuss our experimental results. We begin by checking that the problem conditions are such that the two sensitive attributes can be safely removed from the synthetic dataset. On this basis, we evaluate copy performance in terms of defined metrics. Finally, we show how the removal of the protected variable results in a shift in the original decision boundary that mitigates the bias effect in the resulting decision behavior.

Hypothesis testing In many real scenarios, systematic bias results in individuals belonging to privileged and unprivileged groups not having access to the same resources, a reality that could very well be reflected in the data attributes of each group. Hence, before proceeding proposal, we ascertain the feasibility of our proposal: we verify that the removal of the two sensitive attributes will not result in any residual leakage of information into the copies. This could happen, for example, if the remaining variables encode information that could be traced back to each superhero's *gender* or *race*, even in the absence of this data. In order to ensure that the sensitive attributes can be safely removed in our case, we first check that no other variable is correlated with *gender* and *race*.

In Figure 2 we report the top ten ranked attributes in terms of their one-to-one correlation with these two variables. We show that at most, this correlation is equal to 0.18 in the case of variable *gender* and to 0.35 in the case of *race*. Thus, we conclude that there exist no residual information left in the synthetic dataset after the removal of this attributes and that we can therefore safely

remove them without incurring in any leakage of information. Note that for this particular check we use the original dataset.

Evaluating the copy performance Having established that our proposed approach is feasible, in Table 1 we report our results when replicating the original decision function using the variable-removed copies. The original network yields an accuracy, $\mathcal{A}_O$, of 0.65. The copy, in turn, obtains a copy accuracy, $\mathcal{A}_C$, of 0.65 ± 0.01 averaged over all runs. Notably, the loss in accuracy we incur when substituting the original with the copy in the original data space is close to zero in most cases. Thus, we incur in no effective loss when deploying the copy instead of the original to predict new data points.

Table 1. Performance metrics for original and copy models.

$\mathcal{A}_\mathcal{O}$	$R^{\boldsymbol{Z}}_{\mathcal{F}}$	$R^{\boldsymbol{X}}_{\mathcal{F}}$	$\mathcal{A}_\mathcal{C}$
0.65	0.059 ± 0.003	0.22 ± 0.01	0.66 ± 0.01

The empirical fidelity error measured over the synthetic dataset, $R^{\boldsymbol{Z}}_{\mathcal{F}}$, is equal to 0.031 ± 0.001. This value represents the residual error when learning the optimal copy model parameters to fit the original decision function encoded by the synthetic data points. Finally, the mean empirical fidelity error evaluated over the original test data, $R^{\boldsymbol{X}}_{\mathcal{F}}$, is 0.25 ± 0.01. This value corresponds to the level of agreement between original and copy when generalizing the prediction to new unobserved points in the original data environment. The value of this last error is specially relevant when understanding how the copy is able to replicate the original decision function in the absence of the sensitive information. Removal of the protected attributes from the synthetic dataset results in a certain shift in the learned decision function, with respect to the original. To better understand how this shift impacts the classification of individual data points, we further study the value of the reported performance metrics over the different groups.

Evaluating bias reduction The results above confirm a good fit of the copy to the original predictive performance. On this basis, we evaluate the difference in behavior derived from the removal of the sensitive data attributes in the copy. In Table 2 we report the mean accuracies by *gender* group for original and copy. We observe that there exist significant difference in the predictive accuracy of the original model across the different gender populations. In particular, *male* superheroes are more usually wrongly classified than does in the groups *female*. This is a clear sign of the presence of bias in the original classifier. Independently of whether the decision is dependent on the *gender* attributes it does affect the different groups in a disparate form. When compared to the results obtained by the copy, we observe that the disparity among *male* and *female* groups is notably reduced in the latter. In particular, the difference in accuracy among

the group goes from 0.09 for the original to 0.03 for the copy. As a result, the decisions output by the copy have a more balanced impact on individuals in both populations.

Table 2. Accuracy by *gender* groups for original and copy models.

	Original	**Copy**
female	0.73	0.69
male	0.64	0.66

To better characterize the results in the table above, we further provide the confusion matrices for the two gender groups. Thus, Figures 3(a), 3(b), 3(c) and 3(d) show the relation between true and false positives and negatives for data points in groups *male* and *female* for original and copy, respectively. We observe how in the case of the *male* group, the number of true positives increases for the copy, while the opposite effect is seen for the case of *female*. The net effect of this is the balancing of predictive accuracy between both groups.

Table 3. Accuracy by *race* group for original and copy models.

	Original	**Copy**
human	0.78	0.76
mutant	0.75	0.75
robot	0.67	0.5
extraterrestial	0.25	0.5
other	0.59	0.64

Importantly, these results are also observed for the case of the *race* attribute, albeit less strongly. As shown in Table 3, the mean accuracies by group tend to balance in the case of the copy. This is clearly observed in the two majority classes, namely *humans* and *mutants*. In the minority classes we also see the benefits of the proposal for the group *extraterrestial* which is more often incorrectly classified by the original.

We conclude that this simple approach, does result in a certain mitigation of the bias for the *gender* attribute. Moreover, it shows the potential of copies when used to tackle the issue of fair learning. These results pave the way for more complex treatments of the fairness problem by means of copies. Following our approach, one could, for example, endow copies with fairness metrics such as equity of learning or equality of odds, so that the resulting classifier retain the original accuracy while at the same time optimizing for this new measures.

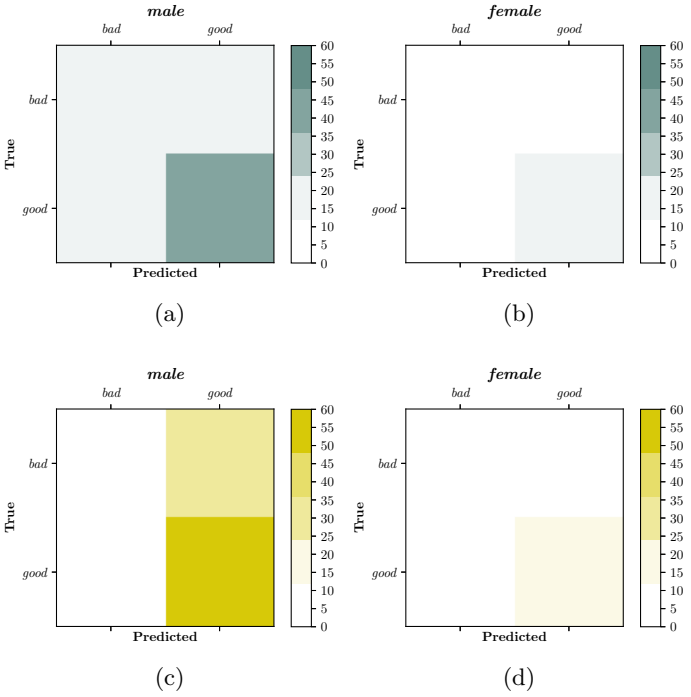


Fig. 3. Confusion matrices for *male* and *female* gender groups for (a) and (b) the original model and (c) and (d) the copy.

6 Conclusions and future work

In this paper we propose and validate a methodology to reduce the unfairness in machine learning classifiers by removing sensitive attributes from the training data. We present a case study using data from the SuperHero database. We train a classifier to learn the superhero alignment using all the data attributes. We then build a copy of this model by removing all the sensitive information. Our results demonstrate that this process can be performed without loss of accuracy and that it can be further exploited to mitigate biases of the original classifier with respect to sensitive attributes such as *gender* or *race*. Our proposed method allows us to redefine the learned decision function to get rid of the sensitive information without risk of leakage.

We here purposely use a fictitious dataset to conduct our experiments. Future work should focus on extending this technique to real datasets where correcting discriminative biases towards sensitive attributes may be crucial to ensure a fair classification. Importantly, this article shows the potential of copies in the study of fairness. Future research should move on from this approach to explore more complex methods for bias reduction in the presence of constraints such as the impossibility to re-train the models or access the original data.

Acknowledgment

This work has been partially funded by the Spanish project TIN2016-74946-P (MINECO/FEDER, UE), and by AGAUR of the Generalitat de Catalunya through the Industrial PhD grant 2017-DI-25. We gratefully acknowledge the support of BBVA Data & Analytics for sponsoring the Industrial PhD.

References

1. Super Heroes Dataset, Kaggle. https://www.kaggle.com/claudiodavi/superhero-set
2. Superhero Database. https://www.superherodb.com/
3. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks. ProPublica (2016)
4. Barocas, S., Selbst, A.D.: Big Data's Disparate Impact. California Law Review **104**(671), 671–732 (2016)
5. Bolukbasi, T., Chang, K.W., Zou, J., Saligrama, V., Kalai, A.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Proc. of Conf. on Neural Information Processing Systems (NIPS). pp. 4356–4364 (2016)
6. Buolamwini, J., Gebru, T.: Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification *. Proc. of Machine Learning Research **81**, 1–15 (2018)
7. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics Derived Automatically from Language Corpora Contain Human-like Biases. Science **356**(6334), 183–186 (2017)
8. Crawford, K.: The Hidden Biases in Big Data. Harvard Business Review (2013)
9. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness Through Awareness. In: Proc. of the 3rd Innovations in Theoretical Computer Science Conf. pp. 214–226 (2012)
10. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD). pp. 259–268 (2015)
11. Guha, S., Cheng, B., Francis, P.: Challenges in Measuring Online Advertising Systems. In: Proc. of ACM Int. Conf. on Data Communications (SIGCOMM). pp. 81–87 (2010)
12. Hardt, M.: How big data is unfair (2014), https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de
13. Hardt, M., Price, E., Srebro, N.: Equality of Opportunity in Supervised Learning. In: Proc. of Conf. on Neural Information Processing Systems (NIPS). pp. 3323–3331 (2016)
14. Klare, B.F., Burge, M.J., Klontz, J.C., Vorder Bruegge, R.W., Jain, A.K.: Face Recognition Performance: Role of Demographic Information. IEEE Trans. on Information Forensics and Security **7**(6), 1789–1801 (12 2012)
15. Popejoy, A.B., Fullerton, S.M.: Genomics is Failing on Diversity. Nature **538**, 161–164 (2016)
16. Unceta, I., Nin, J., Pujol, O.: Towards global explanations for credit risk scoring. In: eprint arXiv:1811.07698 (2018), http://arxiv.org/abs/1811.07698
17. Unceta, I., Nin, J., Pujol, O.: Copying machine learning classifiers. In: eprint arXiv:1903.01879 (2019)