BeFair: Addressing Fairness in the Banking Sector

Alessandro Castelnovo*, Riccardo Crupi*, Giulia Del Gamba‡, Greta Greco*,
Aisha Naseer†, Daniele Regoli*, Beatriz San Miguel Gonzalez†
*Data Science and Artificial Intelligence, Intesa Sanpaolo
‡European Regulatory and Public Affairs, Intesa Sanpaolo
Turin, Italy
Email: name.surname@intesasanpaolo.com
†Fujitsu Laboratories of Europe
London, UK
Email: name.surname@uk.fujitsu.com

2020 IEEE International Conference on Big Data (Big Data) | 978-1-7281-6251-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/BigData50022.2020.9377894

***Abstract*—Algorithmic bias mitigation has been one of the most difficult conundrums for the data science community and Machine Learning (ML) experts. Over several years, there have appeared enormous efforts in the field of fairness in ML. Despite the progress toward identifying biases and designing fair algorithms, translating them into the industry remains a major challenge. In this paper, we present the initial results of an industrial open innovation project in the banking sector: we propose a general roadmap for fairness in ML and the implementation of a toolkit called `BeFair` that helps to identify and mitigate bias. Results show that training a model without explicit constraints may lead to bias exacerbation in the predictions.**

***Index Terms*—machine learning, banking, fairness, bias, discrimination**

I. Introduction

The notion of fairness is still ambiguous and not uniquely defined, mainly because it is context-dependent with complex interdependencies among several attributes. The disparate nature of algorithmic bias and unfair discrimination suggests that fairness could neither be automated [1] nor it can be monolithic. Various Machine Learning (ML) techniques commonly exercise intuitively unfair behaviours, typically due to amplification of bias already encoded in the data or due to minimizing average error to fit majority populations [2], [3]. Anecdotal evidences suggest that a number of fairness metrics are being used to assess the inherent bias in data. Generally, fairness is treated at two different levels: group fairness and individual fairness [4], the first trying to protect vulnerable groups of people and the second focusing on the equality of treatment at the individual level. Nevertheless, it is still not clear whether these two notions are mutually compatible [5], [6]. The choice of which fairness level to use depends on the context of the use case or the concrete business problem at hand. However, most commonly fairness is applied at the group level in most domains.

Although research on fairness in ML has grown in both importance and volume over the past few years and, despite multiple metrics, approaches and methods to pursue fairness have been proposed, there is a lack of consensus on normative standards and industrial frameworks that can enable industry professionals (including data scientists and domain experts) to mitigate bias in ML models. Hence, it becomes imperative to anticipate potential sectoral requirements on fairness in Artificial Intelligence (AI) from a human-centric perspective.

Currently, policy-makers, companies and academic institutions are making efforts towards establishing guidelines and recommendations for Ethics in AI, which includes fairness in ML. One such initiative is AI4People[1], a multi-stakeholder forum in Europe with global activities around the promotion of a "good AI society". In this context, being AI4People members, Fujitsu Laboratories of Europe and Intesa Sanpaolo are collaborating on an open innovation project that reflects a proactive participation in the development and creation of a generic roadmap for Trustworthy AI.

In this paper, we present the initial results that are obtained from our collaboration, in particular, how to implement fairness in a specific banking use case of credit lending. We propose a generic roadmap for fairness in ML that can be applied to various banking use cases. Moreover, we present a toolkit called `BeFair` (Banking, explainability and Fairness) that implements multiple fairness mitigation strategies and demonstrates comparisons among them using different metrics for a specific use case.

The paper is organised as follows. Section II covers the existing background related to fairness metrics and mitigation techniques. Section III introduces our proposed generic roadmap to fairness in ML. Section IV elaborates our `BeFair` toolkit to monitor and assess bias and its mitigation presenting the description of findings in credit lending use case. Section V discusses our findings and results and finally section VI concludes the paper.

II. Background

The global prevalence of AI and its increasingly ubiquitous role in society and business, impacting on every aspect of our lives, has ushered an era where *trust* has never been so important before. Owing to the growing pervasiveness and adoption of AI systems and ML models, the need for achieving Trustworthy systems has become imperative.

Consequently, multiple initiatives from policy-makers, industries and academic institutions have established ethical

[1] AI4People, https://www.eismd.eu/ai4people/

978-1-7281-6251-5/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: UNIVERSITA BICOCCA MILANO. Downloaded on August 26,2021 at 12:40:51 UTC from IEEE Xplore. Restrictions apply.

principles, requirements and recommendations for AI systems where fairness is one of the fundamental principles (we refer to [7], [8] for a complete summary of these initiatives).

Diversity, non-discrimination and fairness are considered key requirements for AI systems according to these initiatives. The aim of the non-discrimination principle is to allow all individuals an equal and fair prospect to access opportunities available in a society. Individuals who are in similar situations should receive similar treatment and not be treated less favourably simply because of a particular "protected" characteristic[2] that they possess (e.g. sex, sexual orientation, disability, age, race, ethnic origin, national origin and religion or belief). Indirect discrimination is present when certain characteristic or factor occurs more frequently in the population groups against whom it is unlawful to discriminate. Since algorithmic decision-making systems may be based on correlations, there is a risk to perpetuate or exacerbate indirect discrimination through stereotyping, when differential treatment cannot be justified [9]. Financial data is prone to bias and imbalance [10] and a multitude of research conducted on specific AI use cases (e.g. credit loan screening applications) shows that putting into practice fairness principles in industrial processes is an open issue.

From a technological perspective, the research on fairness has been approached from two key dimensions: fairness definition (what is unfair discrimination) and bias mitigation (how unfair discrimination is reduced). Next, we describe the most relevant aspects related to each dimension.

A. Fairness definition

Habitually, fairness definitions are divided into two main categories depending on the purpose that is considered: individual and group fairness [4].

Individual fairness is embodied under the principle of "similar individuals should be treated similarly" [5]. Thus, this notion focuses on comparison of individuals and it consists in ensuring that any two individuals who are similar receive equal or similar outcomes. On the other hand, group fairness is focused on requiring that people belonging to protected groups receive on average the same treatment as the whole population, and are usually expressed as the equality of some statistical measure across groups [11]. Therefore, group fairness aims at providing equality of treatment for groups instead of specific individuals.

To assess fairness of a ML model, a precise definition is needed. In this sense, fairness has been mathematically formalised in multiple forms and there is not a clear agreement on which definition to apply in each situation. Moreover, some of the proposed definitions are mutually incompatible and exclusive under some conditions, while some of them are related in non-trivial ways [12], [13].

Afterwards, we describe the most common definitions of fairness with respect to groups identified by some protected attribute(s) (refer to [4], [11], [14] for more details):

- *Statistical Parity or Demographic Parity (DP)* is achieved when groups have the same probability of being assigned to the positive predicted class, i.e. when the decision is independent of the sensitive feature value.
- *Conditional Demographic Parity (CDP)* modifies DP by requiring the parity of outcomes to hold not unconditionally, but within groups given by the level of other variables (e.g. credit risk level).
- *Predictive Parity (PP), Equal Opportunity (EOpp) and Equality of Odds (EO)* not only consider the prediction of the model, but also the ground truth target. In particular, PP measures the precision or probability of a positive prediction to actually be in the positive class (positive predictive value) across groups. EOpp considers the probability of a subject in a positive class to have a positive prediction (true positive rate or recall). Finally, EO requires both true positive rate and false positive rate to be equal across groups.

The aforementioned metrics are based on statistics of observational data coming from the joint distribution of the protected ground(s), input features and target labels. Besides these statistical notions of fairness, there are also proposals of metrics focusing on the use of causal relations among variables, exploiting, together with observational information, domain and expert knowledge. See e.g. [15], [16].

B. Bias Mitigation

Bias mitigation refers to the process of addressing specific aspects of the ML pipeline in order to remove the effect of unfair bias. There are many techniques within the literature that can be roughly classified into the following categories [17], [18]:

- *Pre-processing methods* are based on the idea of removing potential unfair biases directly from the training dataset. Then, a "standard" classifier is learned on this cleaned dataset. In order to "clean the dataset" there are two possible families of methods.
  The first consists in performing a transformation of the the feature space such that the protected information is removed, while at the same time trying to preserve as much information as possible in order to efficiently estimate the target [19]–[22]. A very simple form of pre-processing is the straightforward suppression of the protected attribute(s) from the dataset: this is sometimes called Fairness Through Unawareness (FTU) and embodies a very intuitive notion of individual fairness, namely that two individuals identical in all features but the protected one(s) should be given the same decision [11].
  The second family of methods consists in transforming the dataset by working on observations, namely relabelling or resampling some of them in order to reach group fairness [22].

[2] In this paper we shall use the terms "protected" and "sensitive" interchangeably to indicate attributes to be taken into account when dealing with fairness and discrimination issues.

Authorized licensed use limited to: UNIVERSITA BICOCCA MILANO. Downloaded on August 26,2021 at 12:40:51 UTC from IEEE Xplore. Restrictions apply.

- *In-processing approach* consists in enforcing a model to produce fair outcomes by adding constraints or penalties to the optimization problem, thus imposing fairness at training time. This methods are highly tailored on specific underlying models, thus difficult to generalise. See e.g. [23], [24].
- *Post-processing strategies* are focused on mitigating unfair outcomes of an already trained ML model. The basic idea is to define a new classifier as a function of the (biased) outcomes of the unmitigated model, optimizing some cost function over false positives and false negatives subject to some fairness constraint. The main reference of this type of procedure is [25], while other approaches can be found in [26], [27].

Furthermore, there are mitigation strategies based on causality concepts. The majority of the existing work on these solutions has been focused on individual fairness and can be loosely considered as pre-processing methods, since they consist in training a ML model on a transformed dataset. For example, [15] proposes a method to produce counterfactually fair outcomes, where all protected information causally impacting the decision is removed from the dataset (i.e. an individual is given the same decision that she/he would have been given in the counterfactual world where sensitive features are different). [28] extends this counterfactual fairness, accounting for the fact that not all the causal impact of the sensitive information on the decision is in general unfair, thus mitigating only with respect to variables that are part of unfair causal paths. Another method inspired by causal reasoning but not involving counterfactuals can be found in [16].

Various companies and public institutions have made an effort to encompass fairness metrics and mitigation techniques through specific software tools, toolkits and checklists, such as IBM AI Fairness 360 [29], Google What-If Tool [30], Aequitas [31], and the research to co-design AI fairness checklists [32]. Despite the progress made, these solutions are usually context-agnostic. However, each industry and process have their particularities and it is needed to research and formalize ad-hoc solutions from institutions. Next sections describe the results of our research in the banking sector and in particular for a credit-lending use case.

III. Roadmap to Fairness

In this section, we present our proposal for a generic roadmap to enable fairness in ML. It encompasses five states: regulatory aspects, dataset assessment, choice of fairness metrics, bias mitigation and comparison/evaluations.

It is notable that the roadmap requires specific inputs from various expertise (such as legal and domain knowledge, expert knowledge, etc.). Indeed, pursuing fairness is a process far too intertwined with several ethical and social aspects to be treated as a purely technical issue.

As a final remark, this roadmap must be thought of as a flexible guideline, with steps back and iterations over specific points, in order to converge to the best possible solution between regulatory aspects, mathematical formulation, algorithmic performance and fairness optimization. The process is at least as important as metric optimization.

The roadmap is outlined in the following steps (see Fig. 1):

A. Regulatory aspects

[Domain knowledge, legal expertise] As we have seen in section II there is no single notion of fairness and its definition is highly dependent on specific aspects of the use case at hand and of its domain. In order to decide what is the potentially *sensitive information* and the *concept of fairness* to be pursued, it is necessary to take into account legal and regulatory aspects.

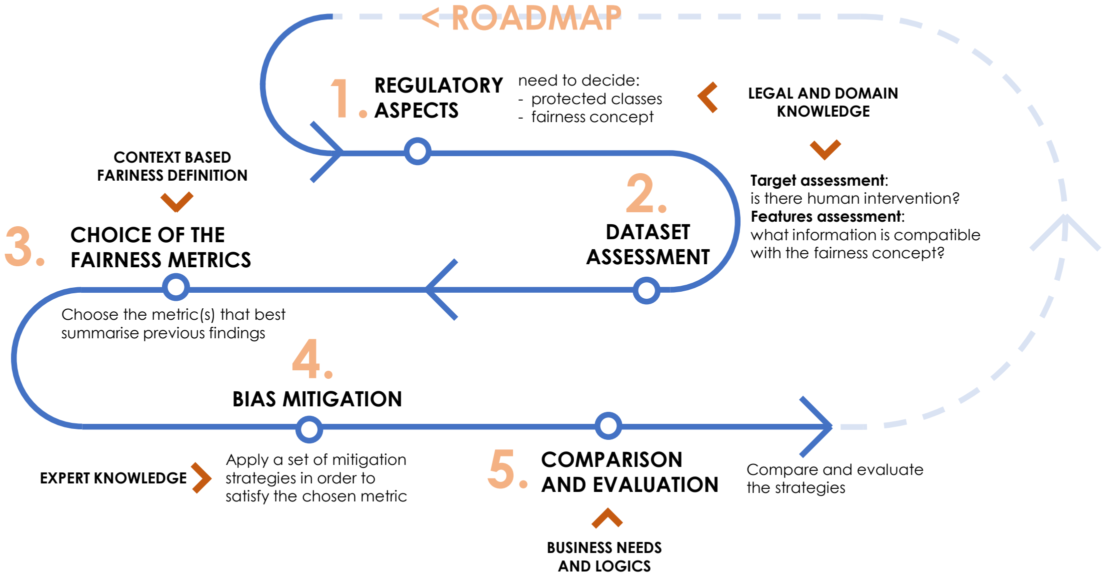
However, the fact that, sooner or later, there will be regulations to clearly prescribe what is and what is not fair in each situation is unlikely, unreasonable and in many ways undesirable. Thus, it is crucial that in each domain and even use case, people working on it, being them developers, scientists or domain experts, consider carefully the consequences of including or not potentially sensitive attributes and the variables correlated to them and what could mean, from the point of view of the final user, to be unfairly discriminated, of course taking into account all the regulatory aspects relevant to the specific situation. In any case, the advent of regulations is itself a (slow) process, during which companies and service providers can play a role by actively doing research and building their own policies and best practices.

B. Dataset assessment

[Domain knowledge, legal expertise] Once the concept of fairness and the sensitive variables have been defined, it is needed to assess what information in the dataset can be actually used to produce fair decisions. This can be done by trying to answer the following questions:

- *Target variable*: is the ground truth variable the result of some form of human judgement or is it based only on facts? This information is important to understand whether the target variable can or cannot be used to actually measure fairness. Metrics such as Equality of Odds heavily rely on the ground truth target to quantify fairness and cannot be used if the target itself is prone to some form of bias.
  Notice that, in the credit lending example, even a variable that may seem objective, as the actual repayment or not of a loan, is subject to a form of selection bias, since it is an information available only for people that were granted a loan in the first place and that are customers of a single bank [33].
- *Feature variables*: what information in the variables is compatible with the chosen concept of fairness? In the credit lending use case, e.g., it may be that a sensitive attribute like gender or citizenship is correlated with income, but it may be compatible with the chosen concept of fairness to use it nonetheless to make lending decisions, since it may be considered a fair way in which sensitive information impacts the final decision: income

Authorized licensed use limited to: UNIVERSITA BICOCCA MILANO. Downloaded on August 26,2021 at 12:40:51 UTC from IEEE Xplore. Restrictions apply.



< ROADMAP
1. REGULATORY ASPECTS
need to decide:
- protected classes
- fairness concept
LEGAL AND DOMAIN KNOWLEDGE
2. DATASET ASSESSMENT
Target assessment:
is there human intervention?
Features assessment:
what information is compatible with the fairness concept?
CONTEXT BASED FARINESS DEFINITION
3. CHOICE OF THE FAIRNESS METRICS
Choose the metric(s) that best summarise previous findings
4. BIAS MITIGATION
EXPERT KNOWLEDGE
Apply a set of mitigation strategies in order to satisfy the chosen metric
5. COMPARISON AND EVALUATION
Compare and evaluate the strategies
BUSINESS NEEDS AND LOGICS

Fig. 1. Schematic visualization of the proposed roadmap representing the process of pursuing fairness in ML projects.

is for sure a crucial variable to determine the probability of repayment.

On the other hand, one may consider that the differences in income (e.g. with respect to gender) are themselves due to historical bias and thus opt for the use of techniques to remove the dependence between income and the sensitive attributes, in order to mitigate this historical bias as well (see e.g. [33] and references therein for more insights).

Notice that, if the target variable is deemed to be objective, one may stick to the Equality of Odds fairness metric and in this case the information in the dataset that is compatible with the chosen fairness concept is precisely the one justified by the target itself.

This decision is of course closely interdependent with step 1, namely with the choice of the proper concept of fairness and of the sensitive attributes.

C. Choice of the fairness metric(s)

[Domain knowledge, legal expertise, data science expertise] Given the outcome of steps 1 and 2, it is possible to choose, among available fairness metrics, the one(s) that best embodies the chosen fairness concept, given the dataset assessment. In particular, some choices to be made are:

- target dependent / target independent;
- group / individual;
- observational / causal.

Depending on the use case, one may decide to monitor more than one metric, e.g. both a group and an individual notion of fairness.

In many respects individual and group fairness can be thought of as two extremes of a continuum of possible metrics, roughly depending on what kind of variables one is willing to accept on the basis of the chosen concept of fairness. Namely, one could condition over all the non-sensitive variables, thus enforcing a form of Conditional Demographic Parity which is equivalent to simply removing the sensitive feature only; or one could not condition at all, thus enforcing the group notion of Demographic Parity, e.g. by removing all the information of the sensitive feature present in the dataset (i.e. the variable itself and all its correlations with other variables). Intermediate forms of Conditional Demographic Parity lie between these two extremes.

The choice of the metric is a crucial step that summarises the knowledge and expertise of steps 1 and 2 and allows to translate them into a mathematical and algorithmic framework.

D. Bias mitigation

[Data science expertise] Given a specific fairness metric, different strategies can be implemented in order to fit a model by pursuing both algorithmic performance and an optimal value of the chosen metric. As discussed in II, these strategies are usually classified in pre-processing, in-processing, post-processing techniques, depending on the specific point of the algorithmic pipeline in which fairness optimization is implemented.

Notice that most of the literature on mitigation techniques is focused on group notions of fairness, while only a small fraction is devoted to mitigation of individual forms of bias. Moreover, individual fairness is still a somewhat ambiguous concept, being it usually defined loosely as “similar people are given similar decisions”, which can be interpreted in many different ways. For instance, the Fairness Through Unawareness is itself a possible form of individual fairness, since two individuals identical with respect to all features but different in the sensitive one(s) are given identical outcomes by design.

On the other hand, many have criticised group fairness, since it may happen that, in order to reach the desired metric value, two individuals similar in all features but in the sensitive

Authorized licensed use limited to: UNIVERSITA BICOCCA MILANO. Downloaded on August 26,2021 at 12:40:51 UTC from IEEE Xplore. Restrictions apply.

one, are given different outcome. Namely, protected individuals are effectively favoured with respect to other individuals. Thus, the demand for applications of individual fairness is constantly increasing, but, up to now, it has to face the problem of the difficulty of providing a clear and precise way to measure it.

E. Comparison of the strategies and evaluation

[Domain knowledge, data science expertise] Once a set of mitigation strategies has been implemented, an evaluation in terms of both algorithmic performance and fairness metric must be made in order to eventually choose the strategy to put in place.

IV. BeFair

As pointed out in section III, there is no such thing as the fairness metric and thus there is neither a single or definitive way to mitigate bias in all situations.

In order to support and guide data scientists through the steps toward the pursue of fairness, we have developed `BeFair`: a collection of tools that allows to implement the notional steps presented in the roadmap on a real use case. The toolkit includes functionalities to detect bias in the input dataset and in model outcomes, to mitigate bias using different strategies derived from the literature, to evaluate the performance of models according to the most relevant metrics, to compare different strategies with a performance/fairness trade-off rationale and to interpret specific feature relationships through a causal graph.

Our work is focused on technical aspects and does not explicitly incorporate the legal expertise prescribed in the roadmap.

In the reminder of this section we delve into the credit lending use case and subsequently introduce the toolkit and its application.

A. Data assessment on credit lending

Decisions concerning credit lending are indeed highly susceptible to unfairness. To evaluate each application, financial institutions often request a certain amount of personal information whose examination might potentially lead to, even unintentional, discrimination. The dataset we used for implementing `BeFair` comes from an anonymised portion of past loan granting applications, whose actual final outcome had no dependence in any way on a ML model.

The dataset consists of about 10^5 loan applications accompanied with a set of personal and financial information, including sensitive attributes, as well as information related to the application, such as the requested amount and duration. Throughout the following sections, we perform the assessment and apply the mitigation strategies over the feature *citizenship* as a mere example, to be able to point out that without intention a standard model can inject discrimination in the predictions even if no significant bias exists in input data.

While there is no specific numerical formula laid out by anti-discrimination laws, to quantitatively determine bias in data we leverage an instantiation of the U.S. Equal Employment Opportunity Commission (EEOC) [34]. In the document, they adopt the so-called 80% rule stating that the ratio between the percentage of subjects belonging to a certain protected group assigned the positive decision outcome and the percentage of subjects not belonging to that group also assigned the positive outcome should be no less than 80:100. In our running example the data satisfies the requirement since the value of the *disparate impact* measured on the attribute *citizenship* over the original target remains below the value 0.8, or equivalently Demographic Parity below 0.2.

B. Fairness mitigation techniques in credit lending

For the credit lending use case, we applied a set of fairness mitigation strategies that we briefly describe in the rest of the section and whose results are summarised in Table I and discussed in section V.

1) Pre-processing: We implement 3 different pre-processing techniques: suppression, massaging and sampling. In *Suppression* [22] the transformed dataset is derived by removing both the sensitive variable and the features with highest correlation with it. We removed all variables with correlation higher than 15%. *Massaging* [22], [35] consists in relabelling the target label of some observations in order to reach Demographic Parity for the “massaged” target. To choose which observations must be relabelled, an auxiliary classifier is trained (the literature proposes to use a Bayesian classifier but we have obtained better results using a Random Forest). Finally, *Sampling* [22] simply consists in over - or under - sampling observations in order to reach Demographic Parity.

Once the dataset has been transformed, we apply a standard Random Forest to get the mitigated outcomes.

Notice that these three techniques are all aimed at reaching Demographic Parity.

2) In-processing: We implement two different algorithms: Adversarial Debiasing and Reductions.

Adversarial Debiasing [23] is based on the simultaneous training of two competing classifiers (corresponding to a Generative Adversarial Network).

In the first one, the predictor P tries to accomplish the task of predicting the target variable Y given the input variables X by modifying its weights W to minimise some loss function $L_P(\hat{Y}, Y)$. The second one, the adversary A, tries to accomplish the task of predicting the sensitive variable, given $\hat{Y}$ by modifying its weights U to minimise some loss function $L_A(\hat{Z}, Z)$ and consequently backpropagates the error through the predictor P.

If the Adversary model is trying to estimate the sensitive attribute given only $\hat{Y}$ the result will satisfy Demographic Parity. Rather, if it is trying to estimate the sensitive attribute given $\hat{Y}$ and the true label Y, the result will satisfy Equality of Odds.

We rely on the Python module `AIF360` [36] to apply Adversarial Debiasing in order to either impose Demographic Parity (AdvDP), Equality of Odds (AdvEO) or Conditional

Authorized licensed use limited to: UNIVERSITA BICOCCA MILANO. Downloaded on August 26,2021 at 12:40:51 UTC from IEEE Xplore. Restrictions apply.

Demographic Parity (AdvCDP). The latest is achieved by training a different Adversarial model for each subgroup of the variable we condition on. In our use case, we condition on three different levels credit risk: high/medium/low.

Reductions approach [24] is based on the idea of finding a classifier that minimises the classification error subject to a specific fairness constrain, effectively reducing the problem to a sequence of cost-sensitive classification problems. More specifically, it takes an arbitrary ML model and trains it multiple times updating, at each iteration, the weights to be assigned to each observations in order to fulfill the chosen fairness constraint. It is actually an hybrid between an in-processing method, since it works by imposing fairness during training and a post-processing method, since it can be applied to any ML models, treating them as black-boxes.

We rely on `Fairlearn` Python module [37] for the actual implementation of the Reductions approach.

We train different models (starting from a logistic regression) using Demographic Parity metric. In particular, first, we train 50 models using the GridSearch algorithm (ReductionsGS) and select the best one. Second, we train 2 models using the ExponentiatedGradient (ReductionsEG) with 0.001 and 0.01 values for the constraints [24].

3) Post-processing: As mentioned in section II, these are a set of techniques that basically consists in computing the mitigated decisions as a function $\hat{Y} = f(R, A)$, where R is the outcome of any given (in general biased) classifier (in our case a Random Forest) and A the sensitive attribute, such that it satisfies a desired fairness metric.

We implement a simple algorithm to impose Demographic Parity (ThreshDP) and one to impose Equality of Opportunity (ThreshEopp) and we rely on `Fairlearn` Python module [37] to compute post-processing mitigation enforcing Equality of Odds (ThreshEO)[3]. Conditional Demographic Parity (ThreshCDP) is enforced by choosing a threshold not only group-dependent but also depending on the level of the variable we are conditioning on, that is chosen to be a 3-level credit risk, as for the Adversarial Debiasing case.

4) Counterfactual fairness through causality: `BeFair` includes the implementation and particularization of counterfactual fairness [15] based on the domain knowledge of our specific use case.

Counterfactual fairness requires to establish a causal graph that represents how the variables influences each other (including input features and outcome). This is usually represented by a Directed Acyclic Graph (DAG) in which each variable is represented by a node and arrows represent causal relationships.

We have defined our specific causal graph using multiple causal discovery algorithms. In particular, we employed the Python `Causal Discovery ToolBox` [38], including different graph modelling algorithms on observational data (i.e. SAM, PC) and the NOTEARS algorithm [39] included in the Python library `CausalNex`. Once the results were obtained and integrated, a manual revision has been performed to verify the validity of each relation detected (represented by arrows). A group of domain experts participated in the validation process to determine the final causal graph that is included in `BeFair` (Fig. 2).

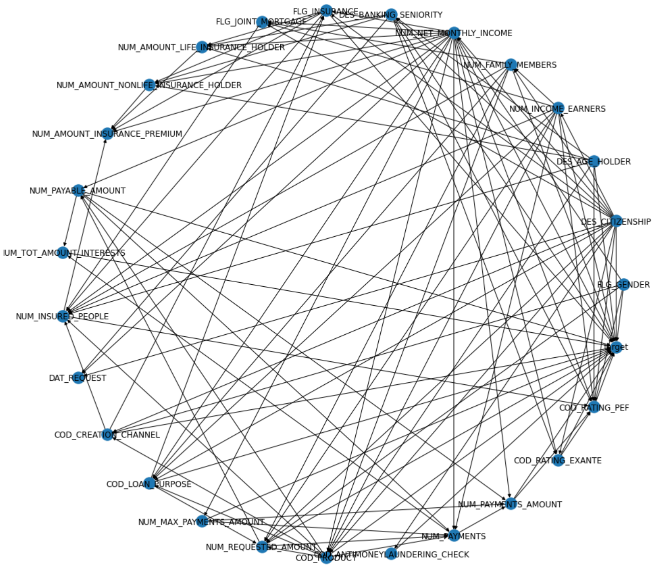


Fig. 2. Causal graph for the credit lending use case.

Given the defined causal graph, we have developed a Counterfactually Fair model (CFF) based on level 3 described in [15]: the main idea is to model the data using an additive error model with deterministic residuals (error terms of the model) and then fit the CFF-fair model on non-descendants of sensitive attribute(s) and the residuals. The process to obtain counterfactual outcomes entails to change the value of the sensitive attribute and to propagate the effect of the change throughout the causal graph to obtain the new values of the new input features. The model trained on the non-descendants of the sensitive feature(s) and the residuals is then counterfactually fair by design, meaning that an individual and its counterfactual version are given the same outcome.

C. Model comparison in credit lending

According to what mentioned in sections III and IV-B about the vast possibility of mitigation techniques and metrics, we propose two approaches that synthesise the trade-off between them and that might help to identify the most suitable model according to the specific domain:

- *Trade-off fairness-performance*. This indicator is inspired by F_β-score used in ML, whose value for $\beta = 1$ results in the harmonic mean of precision and recall:

$$(1+\beta^2)\frac{(1-|\phi|)\times\pi}{\beta^2\times(1-|\phi|)+\pi};$$

[3] Demographic Parity and Equality of Opportunity can be easily enforced by computing group-wise thresholds; while, in general, Equality of Odds requires some form of randomization, since it may not be possible for each group to select a singe threshold classifier reaching the same true positive rate and false positive rate. See [25] and `Fairlean` documentation [37] for more details.

Authorized licensed use limited to: UNIVERSITA BICOCCA MILANO. Downloaded on August 26,2021 at 12:40:51 UTC from IEEE Xplore. Restrictions apply.

TABLE I

Fairness and performance assessment of mitigation strategies discussed in section IV with respect to different metrics. Each panel is devoted to a particular family of mitigation techniques. Fairness metrics are expressed in terms of difference over sensitive groups, thus the lower in absolute value the better. Bold (underline) highlights the best (worst) value per column. Results are discussed in section V.

		fairness				**performance**		
family	**type**	**DP**	**EO**	**EOpp**	**PP**	**AUROC**	**Accuracy**	**F1**
no mitigation	Logistic	0.324	0.272	0.272	**0.032**	0.817	0.761	0.823
	Random forest	0.221	0.202	-0.104	0.068	**0.838**	0.804	0.875
	Neural network	0.219	0.198	0.104	0.072	0.830	0.811	**0.876**
pre-process	FTU	0.164	0.124	0.058	0.095	**0.838**	0.812	**0.876**
	Suppression	0.099	-0.053	0.065	0.152	0.753	0.748	0.840
	Massaging	-0.004	0.062	0.062	0.163	0.818	**0.868**	0.803
	Sampling	0.080	0.012	0.012	0.115	0.835	0.791	0.851
	CFF	0.218	0.192	0.104	0.070	0.832	0.810	0.874
in-process	AdvDP	-0.034	0.073	0.063	0.176	0.823	0.802	0.869
	AdvEO	0.102	0.029	-0.010	0.148	0.819	0.805	0.871
	AdvCDP	0.147	0.101	-0.050	0.112	0.830	0.807	0.872
	ReductionsGS	0.012	0.077	0.049	0.159	0.812	0.794	0.864
	ReductionsEG	0.007	0.084	0.051	0.161	–	0.794	0.864
post-process	ThreshDP	**0.003**	0.099	0.056	0.164	–	0.805	0.872
	ThreshEO	0.082	**0.006**	0.006	0.138	–	0.812	0.873
	ThreshEOpp	0.100	0.048	**0.005**	0.119	–	0.809	0.874
	ThreshCDP	0.186	0.159	0.072	0.083	–	0.810	0.875

where π and ϕ are the preferred performance and fairness metrics, respectively and β is the weight associated with the performance metric.

- *Constrained performance*. Once chosen an upper bound Φ for a desired fairness metric ϕ, this indicator corresponds to the highest possible performance given that fairness constraint

$$\max_{\phi \leq \Phi} \pi .$$

The optimal choice is then given by the model maximizing the selected indicator.

In our BeFair implementation we have considered the following parameters to compare models, although other metrics could be included easily. For the performance metric π we have considered accuracy, precision, recall and F1. For the fairness metric ϕ we have included Demographic Parity, Equal Opportunity, Predictive Parity and Equality of Odds.

BeFair allows users to configure the model comparison. Fig. 3 depicts an example of a graph generated by BeFair showing the trade-off among fairness (x-axis) and performance (y-axis). In particular, it shows the comparison using DP and F1 metrics for all the mitigation strategies implemented (blue dots in the graph) and models without mitigation (orange dots). The best strategy ($y\_prepro\_massaging$) is identified, in this example, using the constrained performance approach with $\Phi = 0.05$, reaching an F1 = 0.875.

V. Discussion of Results

Table I summarises the results of the mitigation strategies introduced in section IV relative to the sensitive attribute of *citizenship* for the credit lending use case.

Fairness quantification is done via the most used statistical fairness dimensions, namely Y-independent (DP), recall based (EO, EOpp) and precision based (PP); while performance is monitored via usual metrics such as Area Under the ROC (AUROC), accuracy and the F1 score (i.e. harmonic mean of precision and recall).

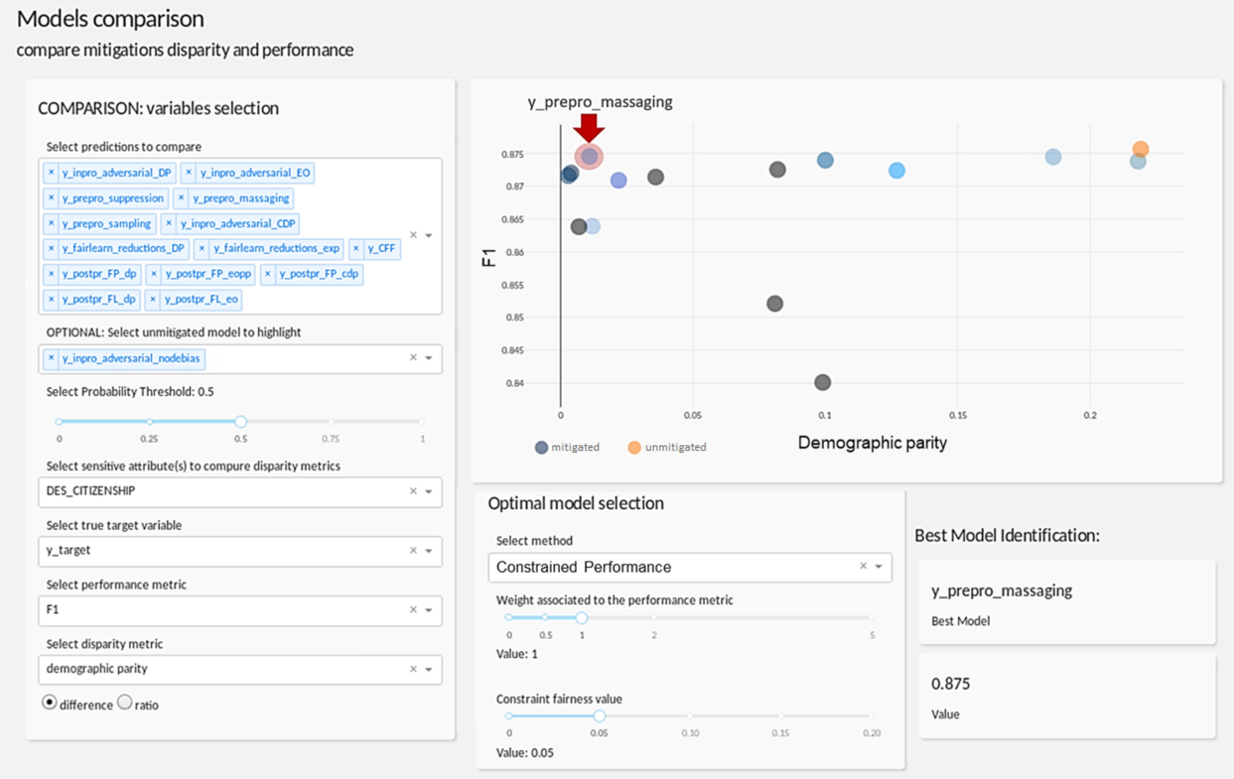
First of all, it is clear from the top panel that naively applying a common ML model to the entire dataset results in amplification of bias with respect to almost all the possible fairness dimensions. This simple fact is *per se* a sufficient reason to promote and foster the need of attention on fairness issues in ML models, in particular in its banking sector applications.

The intuitive and simplest pre-processing methodology of suppression seems not to be able to reach the same level of mitigation of the other strategies, moreover paying a higher price in terms of predictive performance. Thus, bias mitigation while preserving performance is not as simple as removing a bunch of variables from the training dataset. Better results come from massaging the dataset.

Fairness Through Unawareness (FTU)[4], as expected, is not able to mitigate bias with respect to group metrics, since

[4] The underlying classifier is a plain Random Forest.

Authorized licensed use limited to: UNIVERSITA BICOCCA MILANO. Downloaded on August 26,2021 at 12:40:51 UTC from IEEE Xplore. Restrictions apply.



Models comparison

compare mitigations disparity and performance

COMPARISON: variables selection

Select predictions to compare

× y_inpro_adversarial_DP × y_inpro_adversarial_EO × y_prepro_suppression × y_prepro_massaging × y_prepro_sampling × y_inpro_adversarial_CDP × y_fairlearn_reductions_DP × y_fairlearn_reductions_exp × y_CFF × y_postpr_FP_dp × y_postpr_FP_eopp × y_postpr_FP_cdp × y_postpr_FL_dp × y_postpr_FL_eo

OPTIONAL: Select unmitigated model to highlight

× y_inpro_adversarial_nodebias

Select Probability Threshold: 0.5

Select sensitive attribute(s) to compure disparity metrics

DES_CITIZENSHIP

Select true target variable

y_target

Select performance metric

F1

Select disparity metric

demographic parity

● difference ○ ratio

Optimal model selection

Select method

Constrained Performance

Weight associated to the performance metric

Value: 1

Constraint fairness value

Value: 0.05

Best Model Identification:

y_prepro_massaging

Best Model

0.875

Value

Fig. 3. Models comparison: snapshot of `BeFair`. Several mitigation strategies can be selected, together with the desired protected attribute and the chosen performance and fairness metrics to be used (left panel). The right panel is devoted to the performance-fairness plane (top) and to the selection of the optimal model (bottom) (see IV-C).

it does not take into account the sensitive information still present in the dataset via correlations to other variables.

The in-processing techniques are able to reduce DP and EO maintaining the same performance as the unmitigated models. The downside in using these methods is that they consist in training models specifically designed to meet some fairness constraint and they cannot be generalized to arbitrary models.

Post-processing techniques, selecting group-wise thresholds over a trained model, have binary 0/1 outcomes, thus for them it is not possible to compute the Area under the ROC. However, they seem to be overall pretty good in all fairness dimensions without loosing much in performance. On the other hand, they can be criticised for the fact that they, almost by design, consist in treating protected groups differently.

Further clarification must be made for techniques aimed at imposing Conditional Demographic Parity (AdvCDP, ThreshCDP). These techniques result in rather poor performance when assessed with respect to the fairness dimensions present in Table I. This is expected: imposing CDP means to go in the direction of a more individual notion of fairness, where equality is requested only for people having some common conditions (in this example the same level of credit risk). This, of course, does not mean that these strategies produce less fair outcomes, it only means that they are fair with respect to a different notion of fairness. This is in line with the discussions in section III about the paramount importance to carefully consider the notion of fairness appropriate for a specific domain and task. As mentioned above, this decision cannot be left entirely to data scientists and developers, since it involves both domain and legal expertise in complex ways. It is highly desirable that people with different expertise work together along the roadmap, especially in its first steps and their iterations, where crucial decisions are taken about the notion of fairness and the information that can be used safely.

A similar argument holds for the Counterfactual model (CFF), which indeed performs poorly in terms of group fairness dimensions, being it a method trying to enforce an individual notion of fairness. Moreover, it must be taken into account that counterfactual models are unfalsifiable by design, namely there are many alternative counterfactual realities compatible with the same causal graph and implying in general a different definition of what is fair, and no observation can be used to choose among them.

However, the process of building and validating a causal graph for a specific use case marks an important step in the comprehension of the network of interdependence among the variables involved. The more these connections and dependencies are known, the more it is possible to understand what it means, in that specific situation, to be unfairly discriminated. Thus, we believe it represents a valuable tool in the roadmap to fairness.

Notice that the fairness dimension related to equal precision

Authorized licensed use limited to: UNIVERSITA BICOCCA MILANO. Downloaded on August 26,2021 at 12:40:51 UTC from IEEE Xplore. Restrictions apply.

(PP) seems to be worsened by most mitigation techniques: this is coherent with the fact that most techniques aim at reducing DP or EO, thus “loosing ground” in a measure like PP which is related to precision and confirming. Once again, it is worth mentioning that it is not possible to be fair with respect to every possible dimension.

Even if the performance and fairness requirements for the specific case are well-defined, it still is needed to choose the model that best meets them simultaneously. For this purpose, we have proposed two different approaches: (i) use the *constrained performance* method when the fairness constraint is clearly defined as in case of the EEOC instances, otherwise (ii) use the *trade-off fairness-performance* method.

In general, Table I seems to suggest that in this credit-lending use case it is possible to enforce fairness along different dimensions almost without any deterioration of performance.

VI. Conclusion

Fairness has many dimensions. We have seen that several techniques are available both to assess and to mitigate them. However, it is still not clear which dimension should be pursued in each specific situation.

Our roadmap is an attempt to provide a general guideline to address fairness in ML projects and focuses on the fact that different expertise should work together along the process in order to properly take into account the context and the social impact of the technological service/product being developed. It is worth highlighting that our roadmap to fairness is independent of the credit lending dataset used for application in this paper, instead it has been conceived to be use case agnostic and generalised.

Our BeFair toolkit allows data scientists and developers to embed several bias mitigation techniques and assessment metrics within their ML projects and to compare these with a rationale of fairness/performance trade-offs. These can be used to eventually take the practical decisions with respect to the aforementioned roadmap.

We showed the use of BeFair to assess bias in real data from a credit lending use case and to compare different mitigation techniques. The results confirm that using ML models without taking measures to avoid unfair outcomes may lead to strong bias amplification and also that each mitigation methodology has its own strengths and limitations and the choice among them is strictly dependent on the fairness dimension one focuses on.

We believe that more research is needed on the ethical and legal side to disentangle and explicitly elaborate on various categorisations of bias that concur to form the overall discrimination in specific domains. It will facilitate in understanding the information that can be safely exploited in different situations.

On the technical side, effort is still needed to understand more clearly the relationship and trade-offs among different fairness metrics, in particular with respect to the group vs. individual dimension. Indeed, despite many attempts, a precise mathematical formulation of individual fairness remains an open challenge and will likely be subject of future research.

Disclaimer

The views and opinions expressed within this paper are those of the authors and do not necessarily reflect the official policy or position of Fujitsu Laboratories of Europe and Intesa Sanpaolo. Assumptions made in the analysis, assessments, methodologies, models and results are not reflective of the position of any entity other than the authors.

Furthermore, at the moment of writing, Intesa Sanpaolo does not have any ML algorithm in place to evaluate credit lending applications.

References

[1] S. Wachter, B. Mittelstadt, and C. Russell, “Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai,” *Available at SSRN*, 2020.

[2] A. Chouldechova and A. Roth, “The frontiers of fairness in machine learning,” *arXiv preprint arXiv:1810.08810*, 2018.

[3] ——, “A snapshot of the frontiers of fairness in machine learning,” *Communications of the ACM*, vol. 63, no. 5, pp. 82–89, 2020.

[4] B. Hutchinson and M. Mitchell, “50 years of test (un) fairness: Lessons for machine learning,” in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 49–58.

[5] 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*, 2012, pp. 214–226.

[6] R. Binns, “On the apparent conflict between individual and group fairness,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 514–524.

[7] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, “Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai,” *Berkman Klein Center Research Publication*, no. 2020-1, 2020.

[8] A. Jobin, M. Ienca, and E. Vayena, “The global landscape of ai ethics guidelines,” *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[9] Council of Europe, Committee of experts on Internet MSI-NET, “Study on the human rights dimensions of automated data processing techniques and possible regulatory implications,” 2017.

[10] Y. Zhang and L. Zhou, “Fairness assessment for artificial intelligence in financial industry,” *arXiv preprint arXiv:1912.07211*, 2019.

[11] S. Verma and J. Rubin, “Fairness definitions explained,” in *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 2018, pp. 1–7.

[12] J. Kleinberg, S. Mullainathan, and M. Raghavan, “Inherent trade-offs in the fair determination of risk scores,” *arXiv preprint arXiv:1609.05807*, 2016.

[13] 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.

[14] S. Mitchell, E. Potash, S. Barocas, A. D’Amour, and K. Lum, “Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions,” *arXiv preprint arXiv:1811.07867*, 2018.

[15] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, “Counterfactual fairness,” in *Advances in neural information processing systems*, 2017, pp. 4066–4076.

[16] N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, “Avoiding discrimination through causal reasoning,” in *Advances in Neural Information Processing Systems*, 2017, pp. 656–666.

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

[18] L. Oneto and S. Chiappa, “Fairness in machine learning,” in *Recent Trends in Learning From Data*. Springer, 2020, pp. 155–196.

[19] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, “Learning fair representations,” in *International Conference on Machine Learning*, 2013, pp. 325–333.

Authorized licensed use limited to: UNIVERSITA BICOCCA MILANO. Downloaded on August 26,2021 at 12:40:51 UTC from IEEE Xplore. Restrictions apply.

[20] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, "The variational fair autoencoder," *arXiv preprint arXiv:1511.00830*, 2015.

[21] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Advances in Neural Information Processing Systems*, 2017, pp. 3992–4001.

[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] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

[24] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," *arXiv preprint arXiv:1803.02453*, 2018.

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

[26] P. K. Lohia, K. N. Ramamurthy, M. Bhide, D. Saha, K. R. Varshney, and R. Puri, "Bias mitigation post-processing for individual and group fairness," in *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2019, pp. 2847–2851.

[27] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," in *Advances in Neural Information Processing Systems*, 2017, pp. 5680–5689.

[28] S. Chiappa, "Path-specific counterfactual fairness," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 7801–7808.

[29] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic *et al.*, "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *arXiv preprint arXiv:1810.01943*, 2018.

[30] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson, "The what-if tool: Interactive probing of machine learning models," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 56–65, 2019.

[31] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani, "Aequitas: A bias and fairness audit toolkit," *arXiv preprint arXiv:1811.05577*, 2018.

[32] M. A. Madaio, L. Stark, J. Wortman Vaughan, and H. Wallach, "Co-designing checklists to understand organizational challenges and opportunities around fairness in ai," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14.

[33] S. Yeom and M. C. Tschantz, "Discriminative but not discriminatory: A comparison of fairness definitions under different worldviews," *arXiv preprint arXiv:1808.08619*, 2018.

[34] 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*, 2015, pp. 259–268.

[35] F. Kamiran and T. Calders, "Classifying without discriminating," in *2009 2nd International Conference on Computer, Control and Communication*. IEEE, 2009, pp. 1–6.

[36] 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. 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," Oct. 2018. [Online]. Available: https://arxiv.org/abs/1810.01943

[37] 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/

[38] D. Kalainathan and O. Goudet, "Causal discovery toolbox: Uncover causal relationships in python," *arXiv preprint arXiv:1903.02278*, 2019.

[39] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, "Dags with no tears: Continuous optimization for structure learning," in *Advances in Neural Information Processing Systems*, 2018, pp. 9472–9483.

Authorized licensed use limited to: UNIVERSITA BICOCCA MILANO. Downloaded on August 26,2021 at 12:40:51 UTC from IEEE Xplore. Restrictions apply.