Bias on Demand: A Modelling Framework That Generates Synthetic Data With Bias

Joachim Baumann
University of Zurich
Zurich University of Applied Sciences
Zurich, Switzerland
baumann@ifi.uzh.ch

Alessandro Castelnovo*
Data Science & Artificial Intelligence, Intesa Sanpaolo S.p.A.
Dept. of Informatics, Systems and Communication, Univ. Milano Bicocca
Milan, Italy
alessandro.castelnovo@intesasanpaolo.com

Riccardo Crupi*
Data Science & Artificial Intelligence, Intesa Sanpaolo S.p.A.
Turin, Italy
riccardo.crupi@intesasanpaolo.com

Nicole Inverardi*
Data Science & Artificial Intelligence, Intesa Sanpaolo S.p.A.
Milan, Italy
nicole.inverardi@intesasanpaolo.com

Daniele Regoli*
Data Science & Artificial Intelligence, Intesa Sanpaolo S.p.A.
Milan, Italy
daniele.regoli@intesasanpaolo.com

ABSTRACT

Nowadays, Machine Learning (ML) systems are widely used in various businesses and are increasingly being adopted to make decisions that can significantly impact people's lives. However, these decision-making systems rely on data-driven learning, which poses a risk of propagating the bias embedded in the data. Despite various attempts by the algorithmic fairness community to outline different types of bias in data and algorithms, there is still a limited understanding of how these biases relate to the fairness of ML-based decision-making systems. In addition, efforts to mitigate bias and unfairness are often agnostic to the specific type(s) of bias present in the data. This paper explores the nature of fundamental types of bias, discussing their relationship to moral and technical frameworks. To prevent harmful consequences, it is essential to comprehend how and where bias is introduced throughout the entire modelling pipeline and possibly how to mitigate it. Our primary contribution is a framework for generating synthetic datasets with different forms of biases. We use our proposed synthetic data generator to perform experiments on different scenarios to showcase the interconnection between biases and their effect on performance and fairness evaluations. Furthermore, we provide initial insights into mitigating specific types of bias through post-processing techniques. The implementation of the synthetic data generator and experiments can be found at https://github.com/rcrupiISP/BiasOnDemand.

*The views and opinions expressed are those of the authors and do not necessarily reflect the views of Intesa Sanpaolo, its affiliates or its employees.

This work is licensed under a Creative Commons Attribution International 4.0 License.

FAccT '23, June 12–15, 2023, Chicago, IL, USA

© 2023 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0192-4/23/06.
https://doi.org/10.1145/3593013.3594058

CCS CONCEPTS

• **Computing methodologies** → **Philosophical/theoretical foundations of artificial intelligence**; **Machine learning**; • **General and reference** → *Metrics*.

KEYWORDS

bias, fairness, synthetic data, moral worldviews

ACM Reference Format:
Joachim Baumann, Alessandro Castelnovo, Riccardo Crupi, Nicole Inverardi, and Daniele Regoli. 2023. Bias on Demand: A Modelling Framework That Generates Synthetic Data With Bias. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3593013.3594058

1 INTRODUCTION

The increasing digitisation of society has led to a surge of available data, driving the widespread adoption of Machine Learning (ML) in various businesses, governments, and organisations. In many domains, more and more ML-based decision-making systems are used and produce outcomes that affect people's lives. However, algorithms, like humans, are susceptible to biases that might lead to unfair outcomes [2]. Bias is not a recent problem; rather, it is ingrained in human society and, as a result, it is reflected in data [49]. The risk is that the adoption of ML algorithms could amplify or introduce biases that will recur in society in a perpetual cycle [14, 47, 51]. To prevent harmful consequences, it is essential to comprehend how and where bias is introduced and how to mitigate it.

Recent developments show the importance of trustworthy and fair development of AI-based solutions. Many countries have started debating the opportunity of introducing explicit regulation for AI-based automated solutions, one of the risks being precisely that of uncontrolled bias exacerbation and unfair discrimination.

The most important example in this respect is that of the European Union (EU). The European Commission has put out, in April

2021, a Proposal for a "Regulation laying down harmonised rules on artificial intelligence" (Artificial Intelligence Act) [20]. This Proposal – currently in advanced stages of negotiation – is based on EU values and fundamental rights, and seeks to foster trust in AI-based solutions among users[1] while encouraging their development by businesses. Not surprisingly, much attention has been paid to addressing the issue of bias since the very first version of the Proposal. In the US, the White House Office of Science and Technology Policy has recently published a Blueprint for an AI Bill of Rights with 5 overarching guiding principles for AI systems design, development and use, one of which is titled *"Algorithmic Discrimination Protection"* [50]. Alongside, binding legislation has been proposed, now under discussion [48] – the *Algorithmic Accountability Act*. The UK government published in July 2022 an AI Regulation Policy paper where one of the pillars is *"Embed considerations of fairness into AI"* [25], while more advanced and binding legislative proposals are still to come.

Both academia and industry have recently launched many initiatives and projects with the ambitious goal of fostering the development of bias-aware ML models. Following [49], we divide these works into three main categories: *understanding bias*, which includes approaches that help to understand how bias is generated in society and manifests in data [59]; *accounting for bias*, which includes approaches discussing how to manage bias depending on the context, regulation, vision and strategy on fairness [14, 16]; *mitigating bias*, which includes technical approaches aimed at reducing bias throughout the ML development pipeline [6, 17, 22, 32].

One common approach to investigate algorithmic developments is through synthetically generated data [40, 55]. The benefits of this strategy include the possibility of examining circumstances not available with real-world data but that may occur, and – even when real-world data is available – to precisely control and understand the data generation mechanism. Moreover, it is indisputable that making data, and related challenges, accessible to the research community for analysis could be of help for the development of sound policy decisions and thus benefit society [55]. However, to the best of our knowledge, a structured approach to generate synthetic data including (various types of) bias is currently still missing.

1.1 Contributions

In this work, we aim to contribute to understanding, accounting for, and mitigating bias by introducing a model framework for generating synthetic data with specific types of bias. Our formalisation of these various types of bias is based on the theoretical classifications present in the relevant literature, such as the surveys on bias in ML by Mehrabi et al. [47], Ntoutsi et al. [49], and Suresh and Guttag [59]. We provide an explicit mathematical representation of the fundamental types of bias and link it with the stream of literature that investigates their relation with moral worldviews. In particular, following [30, 33, 35], we analyse some biases that our framework is able to generate, considering their fundamental relation with the worldview assumed. We leverage our framework to generate different scenarios characterised by the presence of various types of bias. These scenarios highlight initial empirical insights relating the effects of the presence of specific types of bias on the fairness of ML-based decision-making systems on a group level and also on the mitigation strategies that can be applied.

Our findings confirm previous theoretical results and intuitions: We find that post-processing predicted scores can effectively mitigate different types of bias unless there is measurement bias on the target variable and that this does not necessarily cause a decrease in accuracy. This confirms the findings of Baumann et al. [6], Baumann and Heitz [7], Rodolfa et al. [57]. However, many group fairness criteria are mutually exclusive, i.e. satisfying some fairness criterion comes at the cost of other notions of fairness, which is a logical consequence of the impossibility theorems provided by Chouldechova [19], Kleinberg et al. [39]. Interestingly, the accuracy costs of enforcing the fairness criterion *demographic parity* are higher throughout many of the experiments compared to other notions of fairness. Furthermore, our experimental results show that in the case of measurement bias on some features, *blinding* a classifier (i.e., removing the protected attribute during training and prediction) may introduce unfairness since using information regarding the group membership would enable the classifier to cope with the bias, which confirms the findings of Lipton et al. [41].

Our experiments do not cover the full range of possible scenarios and applications in the field of fairness-aware ML that our system can generate. Through an open-source implementation of the proposed model framework, we aim to allow the research community to exploit our synthetic data generator to create *ad hoc* scenarios that are difficult to find in benchmark datasets available online. This work aims to draw attention to the issue of bias in AI systems and its potential impact on fundamental rights and legal compliance. The objective is to raise awareness and promote the development of equitable AI systems, aligning them with a shared set of ethical principles.

The paper is structured as follows: in Section 2 we briefly discuss related work on synthetic data in ML; Section 3 is devoted to background and related work on the bias landscape. In addition, we discuss different types of bias, definitions of fairness, and bias mitigation techniques. Then, in Section 4, we provide a mathematical formulation for these biases, which is implemented in our synthetic data generator. Section 5 is devoted to presenting a series of experiments where we make use of the synthetic data generator to simulate different scenarios. We discuss some relevant findings from the experiments in Section 6, together with concluding remarks.

2 SYNTHETIC DATA GENERATION

Synthetic data generation is a relevant practice for both businesses and the scientific community. Two main directions in the research on synthetic data are: the *emulation* of certain key information in real dataset while preserving privacy [3, 55], and the *generation* of different testing scenarios for evaluating phenomena not covered by available data [40]. According with Assefa et al. [3], synthetic representations should possess several desirable properties, including *human readability*, *compactness* and *privacy preservation*. Notice that synthetic data generation may also be a valid alternative to data anonymisation as a means of preserving privacy in data to be published or shared [46]. Indeed, synthetic data are typically newly

[1] According to the AI Act, "users" are entities that employ AI systems after self-development or purchase from the market. We shall use the same meaning throughout the paper.

generated data (thus different, by design, from real observations), subject to constraints to protect sensitive personal information while still allowing valid inferences [55].

Synthetic data generation can be approached in several ways, depending mainly on the objective – see [28] for a detailed overview of the techniques for generating synthetic data. When there is (enough) real data available and the main goal is to emulate the "structure" of that data, synthetic samples can be drawn from a probability distribution learned from the real data. This is achieved through distribution fitting approaches, such as Gaussian Mixture Models or Hidden Markov Models, as well as modern Deep Learning-based approaches, ranging from Autoencoders to Generative Adversarial Networks, Diffusion models and Language Models, which are collectively referred to as Generative AI. If the objective is to create benchmark scenarios that comply with specific properties, a possible strategy is to simulate instances using a set of (stochastic) equations that represent the desired relationships among variables. This approach is aligned with the method we propose in the following.

Researchers in the field of algorithmic fairness acknowledge the difficulty in finding suitable datasets for their experiments, relying heavily on a handful of benchmark datasets, e.g. for studying and developing bias mitigation strategies [26]. To overcome this limitation, it is not uncommon to use synthetic datasets to demonstrate specific properties of a novel discrimination-aware method, as highlighted in algorithmic fairness dataset surveys, such as [29, 40]. They show that some works, such as [23], use well-known benchmark synthetic datasets to validate fair representation learning, whereas other studies, such as [6, 15, 24, 42, 56, 63], generate *ad hoc* toy datasets for their testing scenarios. Reddy et al. [56], e.g. evaluate different fairness methods trained with deep neural networks on synthetic dataset: different imbalances and correlations are embedded in the data to verify the limits of the current models and better understand under which setups they are subject to failure.

3 BIAS LANDSCAPE IN ML

There is little consensus in the literature regarding bias classification and taxonomy. Indeed, the very notion of bias depends on deep philosophical considerations, and ethical issues are rarely resolved in a definitive and univocal way. Different understandings of bias and fairness depend on the assumption of a belief system beforehand. Friedler et al. [30] and Hertweck et al. [35] talk about *worldviews*. In particular, Friedler et al. [30] outline two extreme cases, referred to as *What You See Is What You Get* (WYSIWYG) and *We are All Equal* (WAE).

Starting from the definition of three different metric spaces, these two perspectives differ because of the way they consider the relations in between. The first space is the *Construct Space* (CS) and represents all the unobservable realised characteristics of an individual, such as intelligence, skills, determination or commitment. The second space is the *Observable Space* (OS) and contains all the measurable properties that aim to quantify the unobservable features, think e.g. of IQ or aptitude tests. The last space is the *Decision Space* (DS), representing the set of choices made by the algorithm on the basis of the measurements available in OS.

According to WYSIWYG, CS and OS are essentially equal, and any distortion between the two is altogether irrelevant to the fairness of the decision resulting in DS. Contrarily, WAE does not make assumptions about the similarity of OS and CS, and moreover, assumes that we are all equal in CS, i.e. that any difference between CS and OS is due to a biased observation process that results in an unfair mapping between CS and OS. If WYSIWYG is assumed, non-discrimination is guaranteed as soon as the mapping between OS and DS is fair, since CS $\approx$ OS. On the other hand, according to WAE the mapping between CS and OS is distorted by some bias whenever a difference among individuals emerges (this difference is named Measurement Bias in [35]); therefore, to obtain a fair mapping between CS and DS those biases should be mitigated properly.

Building on [30], Hertweck et al. [35] describe a more nuanced scenario by introducing the notion of *Potential Space* (PS): individuals belonging to different groups may indeed have different realised talents (i.e. they actually differ in CS), and these may be accurately measured by resumes (i.e. CS $\approx$ OS), but, if we assume that these groups have the same *potential* talents (i.e. they are equal in PS), then the realised difference must be due to some form of unfair treatment of one group, that is referred to as *life bias*. Hertweck et al. [35] call this view *We Are All Equal in Potential Space* (WAEPS).

With a different perspective, Suresh and Guttag [59] argue that bias can also be seen as a source of harm that arises during different stages of the ML life cycle. Indeed, the entire ML life cycle, from data collection to model deployment, involves a series of decisions and actions that can lead to unintended consequences. Even if detected, it is difficult to establish the proper mitigation method for dealing with biases. A first step in this direction is to understand the different types of bias, their sources and consequences. Figures 1 and 2 exemplify the representation of the fundamental biases from a philosophical (Figure 1) and technical (Figure 2) point of view.

It is important to distinguish between biases that arise during the data collection (affecting the data generation) and biases that arise during the development and deployment of the model (affecting the system's outcome). Namely, because in real cases, the former typically depend on context and are inherent in the data without the user being able to eliminate them during data collection, whereas the latter depend on user's decisions in handling the data. The proper mitigation strategy depends on the comprehension of the biases that affect the data generation and should be determined through both technical and philosophical considerations.

3.1 Fundamental Types of Bias

In what follows we focus on what we consider the core building blocks of most types of bias involved in data generation, namely: *historical bias*, *measurement bias*, *representation bias*, and *omitted variable bias*.

User to Data. Biases going *from user to data* impact the phenomenon to be studied and thus the dataset [47].

Historical bias – sometimes referred to as *social bias*, *life bias*, or *structural bias* [35, 47, 49] – occurs whenever a variable of the dataset relevant to some specific goal or task is dependent on some sensitive characteristic of individuals, *but in principle it should not*. An example is the different average income among men and women,

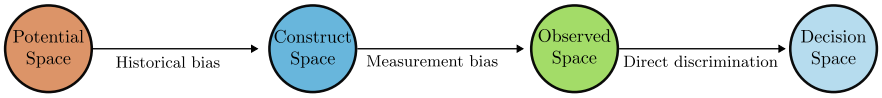


Figure 1: Schematic representation of biases in terms of abstract spaces, as introduced in [30] and extended in [35].

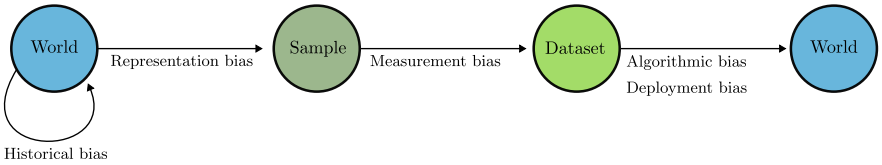


Figure 2: Schematic representation of biases in the ML modelling pipeline, as introduced in [59].

due to long-lasting social barriers and not reflecting intrinsic differences among genders. A similar situation may arise when a dependence on sensitive individual characteristics is present with respect to the variable that we are trying to estimate or predict. These are the cases in which the target of model estimation is itself prone to some form of bias, e.g. because it is the outcome of some human decision. Note that the actual presence of historical bias is conditioned by the previous assumption of the WAEPS worldview. Indeed, arguing that, in principle, there should be no dependence on some sensitive features only makes sense if a moral belief of substantial equity is required in the first place. Otherwise, according to WYSIWYG, CS is fairly reported in OS, and therefore structural differences between individuals are legitimate sources of inequality.

Data to Algorithm. Biases going *from data to algorithm* impact the dataset but not the phenomenon itself [47].

Measurement bias occurs when a proxy of some variable relevant to a specific goal or target is employed, and that proxy happens to be dependent on some sensitive characteristics. For instance, one may argue that IQ is not a "fair" approximation of actual "intelligence", and it might systematically favour/disfavour specific groups of individuals. Statistically speaking, this type of bias is not very different from historical bias – since it results in employing a variable correlated with sensitive attributes – but the underlying mechanism is nevertheless different, and in this case the bias needs not to be present in the phenomenon itself, but rather it may be a consequence of the means chosen to translate unobservable properties into OS. This is an example of bias *from data to algorithm* in the taxonomy of [47], i.e. a bias due to data availability, choice and collection. Note, incidentally, that this form of bias might as well occur with the target variable (i.e. the label). In this situation, it is the quantity that we are trying to estimate/predict that is somehow "flawed". Further, notice that the WYSIWYG worldview assumes that CS ≈ OS, i.e. that there is no measurement bias. On the other hand, the WAE worldview assumes equality among groups only in the CS, which allows for measurement bias (i.e. $CS \neq OS$).

Representation bias occurs when, for some reason, data are not representative of the world population. For example, one subgroup of individuals, identified by a sensitive characteristic such as ethnicity, age, etc., may be heavily underrepresented. This may occur in different ways. It may be at random, i.e. the subgroup is less numerous than it should be, but without any particular skewness in the other characteristics: in this scenario, this single mechanism is not sufficient to create disparities, but it may exacerbate existing ones. Alternatively, the under-represented subgroup might contain individuals with disproportionate characteristics with respect to their corresponding world population, e.g. only low-income individuals or only low-education individuals. In the latter case, representation bias may be sufficient to create inequalities in decision-making processes based on that data.

Omitted variable bias may occur when the collected dataset omits a variable relevant to some specific goal or task. If the variables that are present in the dataset have some dependence on sensitive characteristics of individuals, an ML model trained on such a dataset will learn those dependencies, thus producing outcomes with spurious dependence on sensitive attributes. Notice that the omission of a relevant variable alone cannot, in general, be a source of disparities and bias in the data, but it can amplify and exacerbate other biases already present (e.g. historical biases).

The above list of biases should be seen as the set of the most important mechanisms through which unfairness can be introduced to ML-based decision-making systems due to the used dataset. However, biases can also occur during the development of the ML algorithm (*algorithm to user* biases) or when the system is deployed (*user to world* biases). We provide an overview of these types of bias in Section S1 of the Supplementary Material. For the remainder of this paper, we focus on the biases introduced above, which affect the dataset.

In terms of consequences on the data, it may well be that different types of bias result in very similar effects. For example, representation bias may create in the dataset spurious correlations among sensitive characteristics of individuals and other characteristics relevant to the problem at hand, a situation very similar to the correlations present as a consequence of historical bias. This reminds us that, in general, we are not aware of the type of bias (or biases) affecting the data and that their interpretation depends on former assumptions.

3.2 Fairness Metrics and Bias Mitigation Techniques

The complex nature of biases, as well as the corresponding moral and technical perspectives, results in a large number of possible

fairness metrics [58, 60]. According to recent literature, fairness definitions can be broadly categorised into three groups: *disparate impact* (*DI*), *disparate mistreatment* (*DM*), and *disparate treatment* (*DT*) [13, 62]. Table 1 provides mathematical definitions for all *DI* and *DM* criteria expressed using a binary label Y, binary decisions D, and binary sensitive attribute A (i.e. $Y, D, A \in \{0, 1\}$).

DI is a "group" fairness notion closely related to the concept of *Independence* [4]. A decision-making process suffers from *DI* if it grants a disproportionately large fraction of beneficial outcomes to certain sensitive attributes. The most popular metrics used to measure independence are *demographic parity* (*DP*, also called *statistical parity*) and *conditional demographic parity (CDP)* – both of which are unconditional on the decision D and the outcome Y. *CDP* is slightly weaker than DP as it only requires equal decision rates across subgroups of A that are equal in their value $L = l$, which denotes so-called "legitimate" attributes.

A decision-making process suffers from *DM* if its accuracy (or error rate) is different for different subgroups based on sensitive features. The concept of *DM* can be further divided into *separation* (sometimes referred to as *equalised odds*) and *sufficiency* (also called *conditional use accuracy equality* or *calibration by groups* if enforced over the entire range of predicted scores) [4]. *Separation* prescribe a conditioning on the outcome Y and requires *true positive rate* (*TPR*) *parity* (also known as *equality of opportunity*, *false negative error rate balance*, or *equal recall*) and *false positive rate* (*FPR*) *parity* (also known as *false positive error rate balance* or *predictive equality*). Compared to *DP*, *TPR parity* and *FPR parity* require equal decision rates across all subgroups of A that have the same label Y. *Sufficiency* conditions on the decision D and requires *positive predictive value (PPV) parity* (also known as *predictive parity*, the *outcome test*, or *equal precision*) and *false omission rate (FOR) parity*. *PPV parity* requires individuals that are assigned a positive decision $D = 1$ (a negative decision $D = 0$ in the case of *FOR parity*) to be equally likely to belong to the positive class $Y = 1$ across A.

DT, also known as *individual fairness*, is based on the following principle: *similar individuals should be given similar decisions* [27]. The simplest way to represent *DT* is to define similar individuals as couples belonging to different groups with respect to sensitive features but with the same values for all the other features. In this approach, the outcome for each observation is required not to change when the sensitive attribute is flipped. This concept is usually referred to as *Fairness Through Unawareness (FTU)* or *blindness* [61], and is expressed as the *requirement to avoid explicitly employing protected attributes when making decisions* – though, alternative conceptualisations of *individual fairness* exist.

Bias mitigation techniques are usually divided into *pre-processing* [36, 45], *in-processing* [1, 13, 38, 63] and *post-processing* [6, 22, 32, 43]. *Pre-processing* methods are based on the idea of directly removing potentially unfair biases from the training dataset. A standard classifier is then trained to learn on this cleaned dataset. However, these methods do not guarantee the mitigation of *DM*. *In-processing* approaches consist of forcing a model to produce fair outcomes by adding constraints or penalties to the optimisation problem, thus imposing fairness at the training stage. This method is highly tailored to the specific underlying model and is, therefore, difficult to generalise. In this work, we specifically focus on *post-processing* techniques as they can be easily used for any ML model, only requiring access to the model's outputs and the sensitive attribute information [17].

The algorithmic fairness community has provided formal proofs and implementations for optimal *post-processing* solutions satisfying existing notions of group fairness, as described in the last column of Table 1 [6, 22, 32]. These include finding a so-called decision rule, which transforms the ML prediction into a final decision. Hardt et al. [32] and Corbett-Davies et al. [22] showed that among rules satisfying *DP*, *CSP*, *TPR parity*, or *FPR parity*, the optimum always takes the form of group-specific thresholds.[2] Furthermore, Baumann et al. [6] showed that among rules satisfying *PPV parity* or *FOR parity*, the optimum always takes the form of group-specific upper- or lower-bound thresholds. This means that, in certain situations, it can be optimal to assign a positive decision to the 'worst' individuals of one group (i.e. those with the lowest predicted scores) and omit the most promising ones. This can happen if, for a utility-maximising decision-maker, it is overall better to 'sacrifice' the 'best' individuals of the smaller group in favour of 'keeping' the 'best' individuals from the larger group. For the fairness notions that combine two parity constraints (i.e. *separation* and *sufficiency*), some randomisation is needed to satisfy both constraints at the same time. Among rules satisfying *separation* or *sufficiency*, the optimal decision rules always take the form of randomised group-specific upper- or lower-bound thresholds (see Hardt et al. [32] for *separation* and Baumann et al. [6] for *sufficiency*).

4 DATASET GENERATION

We propose a simple modelling framework able to simulate the bias-generating mechanisms described in Section 3.1.

The rationale behind the model is that of being at the same time sufficiently flexible to accommodate all the main forms of bias *while* maintaining a structure as simple and intuitive as possible to facilitate *human readability* and ensure *compactness* avoiding unnecessary complexities that might hide the relevant patterns.

As noted in Section 3.1, following [47], we can distinguish between *from user to data* and *from data to algorithm* biases. Namely, between biases that impact the phenomenon to be studied and thus the dataset and biases that directly impact the dataset but not the phenomenon itself.[3] Formally, we model the relevant quantities describing a phenomenon as random variables, in particular, we label Y the *target* variable, namely the quantity to be estimated or predicted on the basis of other *feature* variables, that we collectively call X. As usual, we assume that the underlying phenomenon is described by the formula

$$Y = f(X) + \epsilon, \tag{1}$$

where f represents the actual relationship between features and target variables, modulated by some idiosyncratic noise ϵ. Oftentimes, what we observe in the OS is not equivalent to the construct we would like to grasp (in the CS). Formally, this refers to how features and labels are generated and collected:

$$\widetilde{X} = g(X), \quad \widetilde{Y} = h(Y); \tag{2}$$

[2] For the fairness criterion *CSP*, the group-specific thresholds additionally depend on the "legitimate" attributes L.

[3] Bias *From algorithm to user* (impacting the predictor) and *from user to world* (impacting the final decisions) are described in Section S1 of the Supplementary Material.

Table 1: Group fairness criteria

Conditioning on Y, D	Group fairness criterion	Mathematical representation	Post-processing bias mitigation
Unconditional	Demographic parity	$P(D=1 \mid A=0) = P(D=1 \mid A=1)$	Corbett-Davies et al. [22]
	Conditional demographic parity	$P(D=1 \mid L=l, A=0) = P(D=1 \mid L=l, A=1)$	Corbett-Davies et al. [22]
Conditioned on Y	Separation	$P(D=1 \mid Y=i, A=0) = P(D=1 \mid Y=i, A=1), i \in \{0,1\}$	Hardt et al. [32]
	TPR parity	$P(D=1 \mid Y=1, A=0) = P(D=1 \mid Y=1, A=1)$	Corbett-Davies et al. [22],
	FPR parity	$P(D=1 \mid Y=0, A=0) = P(D=1 \mid Y=0, A=1)$	Hardt et al. [32]
Conditioned on D	Sufficiency	$P(Y=1 \mid D=j, A=0) = P(Y=1 \mid D=j, A=1), j \in \{0,1\}$	Baumann et al. [6]
	PPV parity	$P(Y=1 \mid D=1, A=0) = P(Y=1 \mid D=1, A=1)$	Baumann et al. [6]
	FOR parity	$P(Y=1 \mid D=0, A=0) = P(Y=1 \mid D=0, A=1)$	Baumann et al. [6]

where g and h represent the collection and measurement of relevant individual attributes and outcomes. The use of $(\widetilde{X}, \widetilde{Y})$ instead of (X, Y) describes the fact that the set of variables and outcomes employed to make inferences about a phenomenon may not coincide with the actual variables that are relevant to that phenomenon. This is precisely what happens in some biases, such as measurement bias or omitted variable bias, but also in case of representation issues.

A data-driven decision-maker infers from a (training) set of samples $\{(\widetilde{x_i}, \widetilde{y_i})\}_{i=1}^{N}$, an estimate for f that we label $\hat{f}$, thus producing its best estimate for Y, namely

$$\hat{Y} = \hat{f}(\widetilde{X}). \tag{3}$$

The prediction $\hat{Y}$ is then used to inform a final decision D. Thereby, the decision rule, which we denote by r, specifies how a decision is taken based on the individual prediction.

$$D = r(\hat{Y}). \tag{4}$$

In its simplest, fully automated, form without any fairness constraints, optimal decision rules r^* usually take the form of a uniform threshold, i.e. all individuals with a prediction that lies above a certain value τ (i.e. $\hat{Y} > \tau$) are assigned a positive decision $D = 1$, all others are assigned a negative decision $D = 0$. As described in Section 3.2, *post-processing* techniques to ensure a certain fairness constraint act on the decision rule r and take the form of group-specific (upper- or lower-bound) thresholds, i.e. $D = r(\hat{Y}, A)$. However, in many real-world scenarios, decisions are not fully automated but are taken by human decision-makers who take a decision (potentially) based on the predicted outcome. In this case, decisions are not necessarily just based on $\hat{Y}$ (on $\hat{Y}$ and A if a group-specific *post-processing* is applied), i.e. it can depend on any other environmental information Z ($r : \hat{Y}, A, Z \rightarrow D$). If the decision rule r applied by (human or machine) decision-makers introduces unexpected behaviour resulting in disparities between the decision received by individuals from different groups *deployment bias* can arise [59].

Notice that *user to data* types of bias impact directly Equation (1), *data to algorithm* biases affect the data observation process described in Equation (2), *algorithm to user* biases (i.e. *algorithmic bias*) occur at the level of Equation (3), and *user to world* biases (i.e. *deployment bias*) is linked to the decision rule formalised in Equation (4).[4]

[4]See Section S1 of the Supplementary Material for a more detailed description of *algorithmic bias* and *deployment bias*.

Our framework is very much in line with the discussions outlined by Suresh and Guttag [59]. In particular, we refer to Figure 2 in [59] and the corresponding discussion. Incidentally, notice that while Suresh and Guttag [59] make explicit reference to the sampling process, i.e. the act of drawing specific observations from the target population, we embed this aspect directly in the measurement Equations (2). What we propose in the following is a simple and explicit mathematical formalisation of the framework.

First, it is useful to illustratively represents the building blocks of biases as discussed in Section 3.1 via Directed Acyclic Graphs (similar to [52–54]). In general, in order to provide an intuitive grasp on interesting mechanisms and patterns, we shall use the following notation: R are variables representing *resources* of individuals – be them economic resources, or personal talents and skills – which are relevant for the problem, i.e. they directly impact the target Y; A denote variables indicating sensitive attributes, such as ethnicity, gender, etc.; P_R stand for proxy variables that we have access to instead of the original variable R; Q denote additional variables, that may or may not be relevant for the problem (i.e. impacting Y) and that may or may not be impacted either by R or A, e.g. the neighbourhood one lives in.

In particular, Figure 3 shows four minimal graph representations of historical, omitted variable and measurement biases that make use of the notation just introduced. Historical bias occurs when the relevant variable R is somehow impacted by sensitive feature A. Omitted variable bias occurs when, for some reason, we omit the relevant variable R from our dataset and we employ another variable which happens to be impacted by A. Measurement bias occurs when the relevant variable R is, in general, free of bias, but we cannot access it. Therefore, we employ a proxy P_R (which is typically strongly dependent on R) that *is* impacted by sensitive characteristic A. Measurement bias can also occur on the target variable Y when we only have access to a (biased) proxy P_Y of the phenomenon we want to predict.

The system of Equations (5) formalises the relationships between variables used to simulate specific forms of biases. Notice that the variables $N.$ and $B.$ denote independent random variables, either continuous-valued ($N.$) or integer-valued ($B.$). Intuitively, they represent the sources of variability in the generated dataset, while the structure of the equations imposes the (desired) dependence among the relevant variables. The continuous variable R could represent, e.g., salary and the discrete variable Q – which can take $K+1$ different values – could represent a zone in a city. Indeed, Q is distributed

as a binomial variable in $\{0, \dots, K\}$, with Bernoulli marginal probability p_Q dependent on R and A via a simple logistic function. The binary sensitive variable (A) is distributed as a Bernoulli $\{0, 1\}$ variable, with p_A proportion. Variable S is an auxiliary variable used to effectively generate a binary target Y by thresholding S.[5]

$$A = B_A, \quad B_A \sim \mathcal{B}er(p_A); \tag{5a}$$

$$R = -\beta_h^R A + N_R, \quad N_R \sim Gamma(k_R, \theta_R); \tag{5b}$$

$$Q = B_Q, \quad B_Q \mid (R, A) \sim \mathcal{B}in(K, p_Q(R, A)),$$

$$p_Q(R, A) = \text{sigmoid}\left(-(\alpha_{RQ} R - \beta_h^Q A)\right); \tag{5c}$$

$$S = \alpha_R R - \alpha_Q Q - \beta_h^Y A + N_S, \quad N_S \sim \mathcal{N}(0, \sigma_S^2); \tag{5d}$$

$$Y = \mathbf{1}_{\{S > \overline{P_S}\}}. \tag{5e}$$

When simulating measurement bias, either on resources R or on target Y,[6] we are going to use the following *proxies* as noisy (and biased) substitutes for the actual variables:

$$P_R = R - \beta_m^R A + N_{P_R}, \quad N_{P_R} \sim \mathcal{N}(0, \sigma_{P_R}^2); \tag{6a}$$

$$P_S = S - \beta_m^Y A + N_{P_S}, \quad N_{P_S} \sim \mathcal{N}(0, \sigma_{P_S}^2); \tag{6b}$$

$$P_Y = \mathbf{1}_{\{P_S > \overline{P_S}\}}. \tag{6c}$$

We denote with β's the parameters governing the presence and strength of each form of bias, while we use α's for parameters that regulate the relationships among variables not directly involving bias introduction. By varying the values of the parameters, we are able to generate different aspects of biases as follows:

- β_h^j determines the presence and amplitude of the *historical bias* on the variable $j \in \{R, Q, Y\}$;
- β_m^j, when the proxy P_j is used instead of the original variable j, governs the intensity of *measurement bias* on $j \in \{R, Y\}$;
- α_R, α_Q control the linear impact on (S and thus) Y of R and Q, respectively; α_{RQ} represents the intensity of the dependence of Q on R.

Additionally, in order to account for *representation bias*, we undersample the group $A = 1$. The amount of undersampling is governed by the parameter p_u defined as the proportion of the underrepresented group $A = 1$ with respect to the majority group $A = 0$. We draw the undersampling *conditioned* on R by selecting the $A = 1$ individuals with lower values for R. Finally, simulating *omission bias* is as simple as dropping the variable R from the set of features the model uses to estimate Y.

We want to make clear that what we propose is by no means the more general modelling framework to generate any form of bias: we just propose one possibility to formalise different types of bias mathematically, guided by two principles: *simplicity* and *exhaustiveness* with respect to bias types. One can easily think of many variations (some of which are also included in the code implementation) of the modelling framework generating the same bias types in a different way. For example, one could use other distributions for N_R, N_{P_R}, N_S, and N_{P_S}. Other alternatives lie in the functional forms relating the variables, which are here assumed mostly linear for sake of simplicity. Moreover, in some cases, even the mechanism underlying the biases can be more complicated than the simple shift in the expected values: e.g. historical bias could be due to a different variance of R among sensitive groups, or, in general, to the fact that the distribution of $R \mid A$ varies with the specific value of A. Further, note that we understand bias as systematic differences across groups, which is in line with [5]. Thus, as can be seen in Equations (5) and (6), we multiply the bias parameters β with the sensitive attribute A and do not make any explicit assumptions on the underlying causal paths.

5 EXPERIMENTS

As previously stated, the main goal of this work is to provide a simple generative model able to reproduce datasets with (combinations of) fundamental types of bias. Such datasets can be useful to illustrate how various biases may occur in data and to investigate them. In particular, we shall make reference to the two specific examples of biased features in *college admissions* and biased labels in *financial lending* to showcase the effect of measurement bias and historical bias on features and labels.[7] Similar to Binns [12] we use these two specific examples to show that considerations w.r.t. fairness primarily depends on the ethical and social assumptions about the underlying phenomenon. In line with his consideration, we emphasise that the assumptions about worldviews also determine the understanding of the type of bias present in the data, with significant implications on the performance and fairness of ML-based decisions and on the effectiveness of different bias mitigation strategies.

We first generate datasets simulating toy scenarios with different assumptions about the presence and magnitude of relevant biases.[8] Then, each generated dataset is used to train and test a supervised ML classifier[9] that aims to maximise performance by utilising all available variables. Alongside the unmitigated ML model, we train the same model *blinding* the sensitive variable A (i.e. implementing the *FTU* approach), and we use *post-processing* mitigation strategies to enforce *DP* and *TPR parity*.[10] We evaluate the outcomes both in terms of predictive performance (we shall use the overall *accuracy*) and of fairness (through the group differences with respect to the metrics introduced in Section 3.2.

5.1 Example 1: Biased Features in College Admissions

For the first example, we focus on the decision-making context of college admissions, where the task is to determine which candidates are more suitable for a degree program. Let us assume that

[5] A more detailed description of all parameters is provided in Section S2.1 of the Supplementary Material.

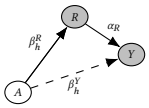
[6] We use the distribution mean of P_S, denoted by $\overline{P_S}$, to derive binary values for Y and its proxy P_Y to avoid predominantly positive or negative labels for one of the groups in the dataset.

[7] In Section S2.3 of the Supplementary Material, we provide additional results for experiments that showcase other types of biases.

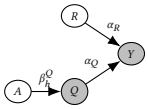
[8] Parameters not directly related to bias are fixed as by Tables S1 and S2 in the Supplementary Material.

[9] Specifically, for the experiments presented here we use a *Random Forest* [10], but any other supervised ML classifier could be used as well. Also, notice that all bias examples presented here result in a disadvantage for individuals of the group $A = 1$ if left unmitigated.

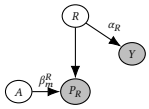
[10] In the Supplementary Material we provide additional results for the remaining *post-processing* bias mitigation techniques enforcing the following group fairness criteria: *FPR parity*, *separation*, *PPV parity*, and *FOR parity* (see Table 1).



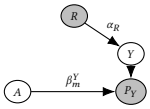
(a) Historical bias on feature and target (dashed) variable



(b) Omitted variable bias



(c) Measurement bias on R



(d) Measurement bias on Y

Figure 3: Illustrative representation of biases. Grey-filled circles represent variables employed by the model $\hat{f}$.

the committee responsible for the admission decision heavily relies on SAT scores of applicants and that these are not independent of individual sensitive characteristics. We shall consider two alternative assumptions: *a) SAT scores are a faithful representation of applicants' skills and competencies*, vs. *b) SAT scores do not faithfully represent applicants' skills and competencies*. A general underlying assumption is that, absent any bias, skills and competencies should be uniform across sensitive groups. Case *a)* is an example of the assumption CS ≈ OS. Thus, as argued in Section 3, SAT score disparities should be a consequence of some form of historical bias impacting the actual skills and competencies of applicants. On the other hand, case *b)* is the result of a measurement bias, where SAT scores are not the proper way to assess skills and competencies, and this creates the undesired disparities. Referring to our notation in Section 4, case *a)* represents a form of *historical bias on R*, and case *b)* represents a form of *measurement bias on R*.

Case a): historical bias on R. The generative model in this scenario is $Y = f(R, Q) + \epsilon,\ R = R(A),\ Q \perp\!\!\!\perp A$, with Y depending on A through R. Figure 4a shows the effect of different magnitudes of historical bias on R (denoted by the parameter β_h^R). In the unconstrained case, all group fairness criteria are violated, and the group disparities are proportional to the size of β_h^R. Interestingly, blinding the model w.r.t. the sensitive attribute (i.e. *FTU*) has no effect since the dependence on A is embedded in R. However, all other bias mitigation techniques manage to ensure the associated group fairness criteria. Post-processing the ML model to achieve *DP* (requiring equal acceptance rates across groups) is the only mitigation technique that is unconditional on Y. As a result, group-level differences in SAT scores are not reflected in the admission decisions, reducing the accuracy with increasing historical bias. However, as can be seen in Figure 4a, the between-group differences of other group fairness metrics (*TPR*, *FOR* and *PPV differences*) increase. Other bias mitigation techniques do not reduce the overall accuracy but also come at the cost of other fairness criteria, empirically confirming their theoretical incompatibility [19, 39]: for example, enforcing *TPR parity* increases *PPV* and *FOR differences* (even though it also brings the groups' acceptance rates closer together and thus has a positive effect on *DP* and *FPR differences*).

Case b): measurement bias on R. In contrast to case *a)*, Y and R do not depend on A, with the only dependence on A being in the proxy of R (P_R). Figure 4b (and Figure S3b in the Supplementary Material, which contains the full results) shows that, in general, the models can cope with the measurement bias on R by leveraging the sensitive attribute A (with a slight accuracy reduction), without any increase in unfairness. This is not the case when the model is blinded w.r.t. A (*FTU*), i.e. the ML model can only cope with measurement bias on SAT scores as long as it is aware of the group memberships A. As can be seen in Figure 4b, *FTU* further reduces the accuracy and generates unfairness w.r.t. to all of the considered fairness metrics. This shows that *FTU* is not a suitable technique to deal with measurement bias on the features. In Section S2.3 of the Supplementary Material, we show the results for an experiment that combines both cases *a)* and *b)* of this example, i.e. using a dataset that contains different magnitudes of historical bias on R and measurement bias on R (see Figure S11).

5.2 Example 2: Biased Labels in Financial Lending

For a second example, we focus on the scenario in which a bank uses an ML model to determine whether loan applications should be approved or denied. Let us assume that the bank notices that the labels are biased, i.e. the rate of repayment is not uniform with respect to gender. As in the first example, we are again considering two distinct scenarios, corresponding to the following two alternative assumptions: *a) historical bias on Y, i.e. the repayment rate disparity reflects a real mismatch in creditworthiness between men and women*; and *b) measurement bias on Y, i.e. the observed repayments are a skewed measure of real creditworthiness.*

Case a): historical bias on Y. Analogously to case *a)* in the first example (Section 5.1), the observed disparity represents a historical bias on Y and is a consequence of a structural discrimination, e.g. via factors like income disparities. As can be seen in Figure 5a, the resulting effects on the fairness and accuracy of the outcomes are very similar to the ones with historical bias on the labels R (shown in Figure 4a). Only the bias mitigation technique *FTU* produces very different results: in contrast to historically biased features, the *FTU* approach is able to reach equality of acceptance rates in the case of historically biased features. Indeed, the generative equation reads $Y = f(R, Q, A) + \epsilon,\ R, Q \perp\!\!\!\perp A$, which is why blinding A is enough to achieve *DP*.

Case b): measurement bias on Y. For this scenario, the observed proxy P_Y of the true outcomes Y is increasingly biased with larger values of β_m^Y.[11] The ML model is trained on the biased label P_Y, and also all bias mitigation techniques are conducted on P_Y. However, the final results are measured w.r.t. the real Y, which is unobservable in reality. Hence, Figure 5b shows that with increasing measurement bias on the labels, the accuracy continuously decreases, as the trained model is unaware of the bias in observed proxy P_Y for the true label Y (see Figure S4b in the Supplementary Material for the full results).

[11] We show in Figure S5 in the Supplementary Material that in the case of measurement bias on Y, A is correlated with the observed Y and $\hat{Y}$ but not with the real Y.

In this case, we assume that instead of measuring the actual creditworthiness of applicants, the repayment rate results are skewed in favour of men, for whom conditions are easier.[12] In this case, the underlying phenomenon reads $Y = f(R, Q) + \epsilon,\ R, Q \perp\!\!\!\perp A$, and the observed dependence on the sensitive attribute comes entirely from the proxy P_Y. The classifier is trained on this proxy, which is why it is calibrated against P_Y but increasingly miscalibrated against the real Y for group $A = 1$ the larger the measurement bias β_m^Y.[13] Consequently, the produced outcomes are unfair w.r.t. the true Y, as visualised in Figure 5b. More precisely, measurement bias on the labels shifts the calibration curve of the ML model against Y upwards, i.e. predicted scores underestimate the ratio of positives.

Most *post-processing* bias mitigation techniques fail to achieve any group fairness criteria (see Figure S4b in the Supplementary Material for the full results). However, similarly to the decreasing accuracy, this is due to the fact that, in Figure 5b, accuracy and fairness are measured w.r.t. Y, and the "distance" between Y and P_Y grows with β_m^Y. Only *FTU* and enforcing *DP* through *post-processing* are effective in mitigating measurement bias on the labels. Both methods behave very similarly since they do not depend on the observed outcome Y (and because the feature R is free of any bias, i.e. it does not depend on A, see Figure S5 in the Supplementary Material) – in contrast to other group fairness criteria as shown in Table 1. Notice that those two techniques manage to fully mitigate any measurement bias on Y. For *FTU*, this effect occurs since using the unbiased feature R without being aware of the group membership A makes it impossible for the ML model to capture the group-level disparities in P_Y. In contrast, for *DP*, this is due to the linear implementations of the measurement bias and of the effect of R on Y (through S) using the parameters β_m^Y and α_Y, respectively (see Equations (6) and (5)). Namely, this shifts the distribution of P_S (and, thus, its mean $\overline{P_S}$), which is equivalent to flipping the label from $Y = 1$ to $P_Y = 0$ for those individuals of group $A = 1$ that have $\overline{S} > P_S > \overline{P_S}$. In a non-linear implementation of the measurement bias, where the label flipping of the individuals in group $A = 1$ depends on other variables, the application of group-specific thresholds would not be as effective as it is in the simple scenario presented here. Notice that in the lending scenario, we are considering, S could represent an individual's true probability of repaying on due time. Thus, the linear shift of S makes sense to replicate the lower leniency of bank clerks towards women (denoted by $A = 1$) when it comes to the repayment deadline. However, in other scenarios, *non-linear* implementations of bias might be more realistic. See Section S2.4 in the Supplementary Material for an experiment showing the results of a non-linear implementation of measurement bias on Y.

In the Supplementary Material, we show the results of an experiment that combines both cases, i.e. using a dataset that contains different magnitudes of historical bias on Y and measurement bias on Y (see Figure S12).

[12] Binns [12] mentions that such discrimination against women can be the result of bank clerks being systematically more lenient with loan repayment deadlines for men. This means that men (m) are more likely to end up repaying their loan despite not being more creditworthy compared to women (w), i.e. $\mathbb{E}(P_Y \mid A = m) > \mathbb{E}(P_Y \mid A = w)$ but $\mathbb{E}(Y \mid A = m) = \mathbb{E}(Y \mid A = w)$.

[13] This is visualised in Figure S2 in the Supplementary Material.

6 DISCUSSION AND CONCLUSION

Connecting worldviews and bias mitigation techniques. As outlined in both examples of Section 5, the type of bias present in a dataset may depend crucially on assumptions about moral worldviews and, ultimately, about the data generation mechanism. This is even more important in light of the fact that mitigation strategies behave differently when facing different types of bias: measurement and historical bias have very similar patterns on data observable statistics but quite different consequences on the choice of the most appropriate bias mitigation strategy, as exemplified in the difference between Figures 4a and 4b, and between Figures 5a and 5b.

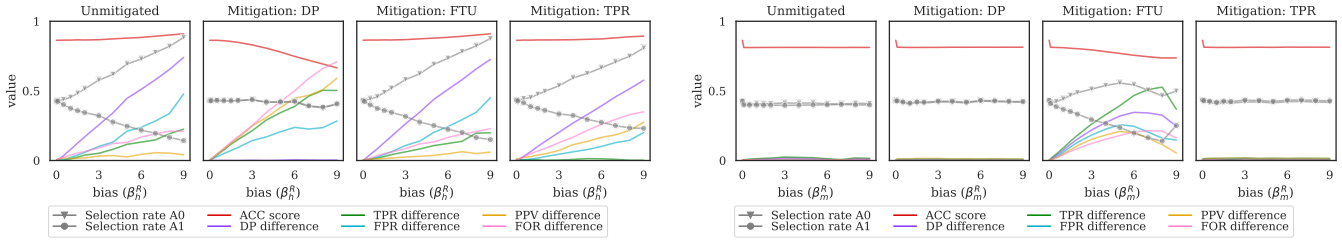
The biased label problem. Our findings show that the solutions proposed by Baumann et al. [6], Corbett-Davies et al. [22], Hardt et al. [32] to post-process predicted scores effectively manage to mitigate biases as long as there is no measurement bias on the label.

As Figure 5b shows, the case of measurement bias on labels is particularly subtle: having access only to the (biased) proxy P_Y, it is only possible to control the bias when imposing fairness criteria that do not make use of target variable, namely *DP* and *FTU* in our experiments. For all other criteria (*TPR*, *FPR*, *PPV*, and *FOR parity* or combinations of these), one would mitigate the group differences of errors with respect to P_Y (and not to Y), and thus would be erroneously induced to judge the model as fair and accurate when it is not.

On the trade-offs of ML-based decision-making systems. Our findings empirically confirm the existence of different trade-offs emerging when enforcing fairness in ML-based decision-making systems [31]. There is a trade-off between fairness and performance (e.g., measured by the accuracy), as well as between different notions of fairness.

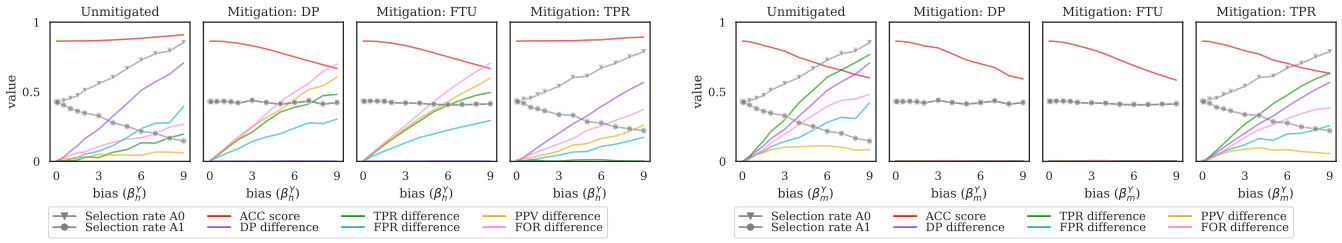
Figures 4-5 show that in certain cases, a biased dataset results in lower overall performance compared to an unbiased dataset. Furthermore, the application of bias mitigation techniques generally comes at an increasing cost, in terms of performance, as the bias increases. However, this is not necessarily the case in all situations, meaning that the performance-fairness trade-off may sometimes be negligible, which is in line with the findings of Rodolfa et al. [57]. As our experiments show, depending on the bias present in the dataset, certain bias mitigation techniques are 'cheaper' in terms of accuracy. At the same time, not all post-processing bias mitigation techniques result in equal outcomes for the affected individuals. Assessing whether the system is fair for the affected individuals requires normative choices w.r.t. what constitutes a just outcome. Several works emphasise how the moral appropriateness of certain notions of fairness heavily depends on the context, such as Baumann and Loi [9], who provide an ethical argument in favour of the *sufficiency* criterion in the context of personalised insurance premiums or Binns [11], who argues that in other applications (such as the selection of candidates from a pool of job applicants) *equality of opportunity* might be more appropriate. However, in this paper, our main focus is on the connection between different biases and fairness metrics.[14]

[14] We refer the interested reader to [7, 33, 44], who provide a framework to choose a morally appropriate group fairness criterion, and to [8, 34] for a unification, extension, and interpretation of group fairness metrics.



(a) Historical bias on R **(b) Measurement bias on R**

Figure 4: Accuracy and fairness metrics for biased features R in the college admission example.



(a) Historical bias on Y **(b) Measurement bias on Y**

Figure 5: Accuracy and fairness metrics for biased labels Y in the financial lending example. Notice that all metrics in (b) are computed with respect to the "true" target Y.

Our findings empirically confirm that there is a trade-off between different notions of fairness: enforcing some fairness criteria may come at the cost of others. This is a logical consequence of the mathematical incompatibility between certain fairness criteria [19, 39]. As can be seen in Figures 4-5, group-level differences in *PPV*s are usually relatively small compared to acceptance rate differences, and enforcing *TPR* or *FPR parity* oftentimes brings those acceptance rates closer together. Furthermore, our experiments show that *DP* mitigates every group fairness criteria in the case of measurement bias on the label Y or on the feature R. However, as explained in Section 5.2, this is due to the linear implementations for those biases. Alternatively, if there is historical bias, clear trade-offs between the different bias mitigation techniques and w.r.t. accuracy emerge.

On the effect of blinding. Despite its simplicity, *FTU* should be applied with particular care, as other works have pointed out [18, 21, 27, 37, 41]. Even though the reason for applying *FTU* is not primarily to achieve group fairness, we believe that it is still relevant to take its effects on different group fairness criteria into account. In fact, in some very particular cases, *FTU* can be effective even w.r.t. group fairness metrics, e.g. when the observed proxy for the target variable P_Y is wrongly assumed to be free of bias (see Figure 5b). However, in general, it is not an effective bias mitigation technique w.r.t. the fairness of the produced outcome for the affected individuals, as shown for the example of historically biased features R, where *FTU* has no effect whatsoever since the information on group membership is redundantly encoded in R (see Figure 4a). And, in fact, there are even cases in which the application of *FTU* leads to biased results and performance deterioration even when the unconstrained model does not (see Figure 4b).

Conclusion. In this work, we contribute to investigating bias in ML-based decision-making systems by introducing a modelling framework to generate synthetic data, including specific types of bias. We present an explicit mathematical representation of the fundamental types of bias discussed by the algorithmic fairness community. Furthermore, we show that the assumptions on different worldviews influence the interpretation of biases that could be present in data. We showcase our framework by simulating different plausible scenarios with various types of bias. Thereby, we observe the effects of employing ML models on biased datasets as well as the behaviour of several bias mitigation techniques. In real-world scenarios, data is typically observed without a clear knowledge of the underlying generation mechanism. We argue that the assumptions on the data generation mechanism are crucial to shaping the interpretation of bias present in the use case under consideration.

This work aims to raise awareness of bias in AI systems and its potential impacts on individuals and society, promoting the development of bias-free AI systems that are consistent with the universal ethical principle of non-discrimination. Through the open-source implementation of the presented framework, we hope to encourage the research community to conduct further studies using synthetic datasets where real-world datasets are missing, by exploiting our synthetic data generator.

ACKNOWLEDGMENTS

The authors are grateful to Andrea Cosentini, Ilaria Penco and to all the members of the Artificial Intelligence & Data Science Dept.

of Intesa Sanpaolo S.p.A. for their support in Trustworthy AI initiatives and research. We would like to thank Michele Loi and the members of the Social Computing Group at the University of Zurich (Anikó Hannák, Nicolò Pagan, Corinna Hertweck, Stefania Ionescu, Aleksandra Urman, Azza Bouleimen, Leonore Röseler, and Desheng Hu) for their helpful feedback on an earlier draft of this manuscript. We would also like to thank the anonymous reviewers for their valuable comments and helpful suggestions. Joachim Baumann was supported by the National Research Programme "Digital Transformation" (NRP 77) of the Swiss National Science Foundation (SNSF) – grant number 187473 – and by Innosuisse – grant number 44692.1 IP-SBM.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*. PMLR, 60–69.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals, and it's biased against blacks. *ProPublica* (2016).

[3] Samuel A Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E Tillman, Prashant Reddy, and Manuela Veloso. 2020. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance*. 1–8.

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

[5] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[6] Joachim Baumann, Anikó Hannák, and Christoph Heitz. 2022. Enforcing Group Fairness in Algorithmic Decision Making: Utility Maximization Under Sufficiency. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2315–2326. https://doi.org/10.1145/3531146.3534645

[7] Joachim Baumann and Christoph Heitz. 2022. Group Fairness in Prediction-Based Decision Making: From Moral Assessment to Implementation. In *2022 9th Swiss Conference on Data Science (SDS)*. 19–25. https://doi.org/10.1109/SDS54800.2022.00011

[8] Joachim Baumann, Corinna Hertweck, Michele Loi, and Christoph Heitz. 2023. Distributive Justice as the Foundational Premise of Fair ML: Unification, Extension, and Interpretation of Group Fairness Metrics. (2023). arXiv:2206.02897 http://arxiv.org/abs/2206.02897

[9] Joachim Baumann and Michele Loi. 2023. Fairness and Risk: An Ethical Argument for a Group Fairness Definition Insurers Can Use. *Philosophy & Technology* (2023). https://doi.org/10.1007/s13347-023-00624-9

[10] Gérard Biau and Erwan Scornet. 2016. A random forest guided tour. *Test* 25, 2 (2016), 197–227.

[11] Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 149–159. https://proceedings.mlr.press/v81/binns18a.html

[12] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 514–524.

[13] Alessandro Castelnovo, Andrea Cosentini, Lorenzo Malandri, Fabio Mercorio, and Mario Mezzanzanica. 2022. FFTree: A flexible tree to handle multiple fairness criteria. *Information Processing & Management* 59, 6 (2022), 103099.

[14] Alessandro Castelnovo, Riccardo Crupi, Giulia Del Gamba, Greta Greco, Aisha Naseer, Daniele Regoli, and Beatriz San Miguel Gonzalez. 2020. BeFair: Addressing Fairness in the Banking Sector. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 3652–3661.

[15] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. 2022. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports* 12, 1 (2022), 1–21.

[16] Alessandro Castelnovo, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Andrea Cosentini. 2021. Towards Fairness Through Time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 647–663.

[17] Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053* (2020).

[18] Irene Y Chen, Fredrik D Johansson, and David Sontag. 2018. Why is My Classifier Discriminatory?. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 3543–3554.

[19] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[20] The European Commission. 2021. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence.

[21] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. arXiv:1808.00023 [cs.CY] https://arxiv.org/abs/1808.00023

[22] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. Association for Computing Machinery, New York, NY, USA, 797–806. https://doi.org/10.1145/3097983.3098095

[23] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. 2019. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*. PMLR, 1436–1445.

[24] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, David Sculley, and Yoni Halpern. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 525–534.

[25] Media Department for Digital, Culture and Sport. 2021. Establishing a pro-innovation approach to regulating AI An overview of the UK's emerging approach. https://www.gov.uk/government/publications/establishing-a-pro-innovation-approach-to-regulating-ai.

[26] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. *Advances in Neural Information Processing Systems* 34 (2021).

[27] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[28] Khaled El Emam, Lucy Mosquera, and Richard Hoptroff. 2020. *Practical synthetic data generation: balancing privacy and the broad availability of data*. O'Reilly Media.

[29] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery* 36 (09 2022), 1–79. https://doi.org/10.1007/s10618-022-00854-z

[30] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM* 64, 4 (2021), 136–143.

[31] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A Comparative Study of Fairness-Enhancing Interventions in Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 329–338. https://doi.org/10.1145/3287560.3287589

[32] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[33] Hoda Heidari, Michele Loi, Krishna P Gummadi, and Andreas Krause. 2019. A moral framework for understanding fair ml through economic models of equality of opportunity. In *Proceedings of the conference on fairness, accountability, and transparency*. 181–190.

[34] Corinna Hertweck, Joachim Baumann, Michele Loi, Eleonora Viganò, and Christoph Heitz. 2023. A Justice-Based Framework for the Analysis of Algorithmic Fairness-Utility Trade-Offs. (2023). arXiv:2206.02891 http://arxiv.org/abs/2206.02891

[35] Corinna Hertweck, Christoph Heitz, and Michele Loi. 2021. On the moral justification of statistical parity. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 747–757.

[36] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*. IEEE, 1–6.

[37] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33.

[38] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Machine Learning and Knowledge Discovery in Databases*, Peter A Flach, Tijl De Bie, and Nello Cristianini (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 35–50.

[39] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*

(2016).
[40] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. 2022. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (2022), e1452.
[41] Zachary C. Lipton, Alexandra Chouldechova, and Julian McAuley. 2018. Does mitigating ML's impact disparity require treatment disparity?. In *Advances in Neural Information Processing Systems*, Vol. 31. Curran Associates, Inc., 8136–8146. arXiv:1711.07076 https://proceedings.neurips.cc/paper/2018/file/8e0384779e58ce2af40eb365b318cc32-Paper.pdf
[42] Wei-Yin Loh, Luxi Cao, and Peigen Zhou. 2019. Subgroup identification for precision medicine: A comparative review of 13 methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9, 5 (2019), e1326.
[43] Pranay K Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R Varshney, and Ruchir Puri. 2019. Bias mitigation post-processing for individual and group fairness. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2847–2851.
[44] Michele Loi, Anders Herlitz, and Hoda Heidari. 2019. A Philosophical Theory of Fairness for Prediction-Based Decisions. *Available at SSRN 3450300* (2019).
[45] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. 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*. 1–14.
[46] Abdul Majeed and Sungchang Lee. 2020. Anonymization techniques for privacy preserving data publishing: A comprehensive survey. *IEEE access* 9 (2020), 8512–8545.
[47] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.
[48] Jakob Mökander, Prathm Juneja, David S Watson, and Luciano Floridi. 2022. The US Algorithmic Accountability Act of 2022 vs. The EU Artificial Intelligence Act: what can they learn from each other? *Minds and Machines* (2022), 1–8.
[49] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 3 (2020), e1356.
[50] White House Office of Science and Technology Policy. 2022. Blueprint for an AI Bill of Rights: Making Automated Systems Work For The American People. https://www.whitehouse.gov/ostp/ai-bill-of-rights/.
[51] Nicolò Pagan, Joachim Baumann, Ezzat Elokda, Giulia De Pasquale, Saverio Bolognani, and Anikó Hannák. 2023. A Classification of Feedback Loops and Their Relation to Biases in Automated Decision-Making Systems. (2023). arXiv:2305.06055 http://arxiv.org/abs/2305.06055
[52] Judea Pearl. 2009. *Causality*. Cambridge university press.
[53] Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic Books.
[54] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. 2014. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research* 15, 58 (2014).
[55] Trivellore E Raghunathan. 2021. Synthetic data. *Annual Review of Statistics and Its Application* 8 (2021), 129–140.
[56] Charan Reddy, Deepak Sharma, Soroush Mehri, Adriana Romero-Soriano, Samira Shabanian, and Sina Honari. 2021. Benchmarking bias mitigation algorithms in representation learning through fairness metrics. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.
[57] Kit T Rodolfa, Hemank Lamba, and Rayid Ghani. 2021. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence* 3, 10 (2021), 896–904. https://doi.org/10.1038/s42256-021-00396-x
[58] Nripsuta Ani Saxena. 2019. Perceptions of Fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 537–538.
[59] Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and access in algorithms, mechanisms, and optimization*. 1–9.
[60] Carmit T Tadmor, Ying-yi Hong, Melody M Chao, Fon Wiruchnipawan, and Wei Wang. 2012. Multicultural experiences reduce intergroup bias through epistemic unfreezing. *Journal of personality and social psychology* 103, 5 (2012), 750.
[61] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 1–7.
[62] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. 2019. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research* 20, 1 (2019), 2737–2778.
[63] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*. PMLR, 962–970.