Bias On Demand: Investigating Bias with a Synthetic Data Generator

Joachim Baumann[1,2] , **Alessandro Castelnovo**[3,4] , **Andrea Cosentini**[3] , **Riccardo Crupi**[3] , **Nicole Inverardi**[3] and **Daniele Regoli**[3]
[1]Department of Informatics, University of Zurich, Switzerland
[2]Institute of Data Analysis and Process Design, Zurich University of Applied Sciences, Switzerland
[3]Data Science and Artificial Intelligence, Intesa Sanpaolo S.p.A., Italy
[4]Dept. of Informatics, Systems and Communication, Univ. of Milan-Bicocca, Italy
baumann@ifi.uzh.ch,
{alessandro.castelnovo,riccardo.crupi,nicole.inverardi,daniele.regoli}@intesasanpaolo.com

Abstract

Machine Learning (ML) systems are increasingly being adopted to make decisions that might have a significant impact on people's lives. Because these decision-making systems rely on data-driven learning, the risk is that they will systematically propagate the bias embedded in the data. To prevent harmful consequences, it is essential to comprehend how and where bias is introduced and possibly how to mitigate it. We demonstrate `Bias on Demand`, a framework to generate synthetic datasets with different types of bias, which is available as an open-source toolkit and as a pip package. We include a demo of our proposed synthetic data generator, in which we illustrate experiments on different scenarios to showcase the interconnection between biases and their effect on performance and fairness evaluations. We encourage readers to explore the full paper for a more detailed analysis.

1 Introduction and Motivation

The increasing digitisation of society has led to a surge in available data, driving the widespread adoption of ML. However, algorithms, like humans, are susceptible to biases that might lead to unfair outcomes [Angwin *et al.*, 2016]. Bias is not a recent problem: it is ingrained in human society and, as a result, it is reflected in data [Ntoutsi *et al.*, 2020; Castelnovo *et al.*, 2022a]. The risk is that the adoption of ML algorithms could amplify or introduce biases that will recur in society in a perpetual cycle [Mehrabi *et al.*, 2021; Castelnovo *et al.*, 2020; Pagan *et al.*, 2023]. 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 [Hutchinson and Mitchell, 2019]. 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 [Ntoutsi *et al.*, 2020], 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 [Suresh and Guttag, 2021]; *accounting for bias*, which includes approaches discussing how to manage bias depending on the context, regulation, vision and strategy on fairness [Hu *et al.*, 2019; Castelnovo *et al.*, 2021; Crupi *et al.*, 2022]; *mitigating bias*, which includes technical approaches aimed at reducing bias throughout the ML development pipeline [Caton and Haas, 2020]. One common approach to investigate algorithmic developments is through synthetically generated data [Le Quy *et al.*, 2022; Howe *et al.*, 2017; Gujar *et al.*, 2022].

In this work, we demonstrate a way to investigate bias by exploiting `Bias on Demand` [Baumann *et al.*, 2023], our modeling framework for generating synthetic data with specific types of bias[1]. The formalisation of various types of bias is based on the theoretical classifications in relevant surveys on bias in ML [Mehrabi *et al.*, 2021; Ntoutsi *et al.*, 2020; Suresh and Guttag, 2021]. 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 contribute to sound policy decisions that benefit society [Raghunathan, 2021]. We leverage the framework to generate different scenarios characterised by the presence of various types of bias.[2] 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 bring attention to the issue of bias and promote the development of *free of bias* AI systems, aligning them with the sustainable development goals.

2 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 and ethical considerations. Different understandings of bias and fairness depend on the as-

[1] Please refer to the full paper [Baumann *et al.*, 2023] for further details on the synthetic data generator, as well as a set of examples.

[2] Demo available at tinyurl.com/biasondemand. For more experiments and the code, see github.com/rcrupiISP/BiasOnDemand.

sumption of a belief system beforehand. [Friedler *et al.*, 2021] and [Hertweck *et al.*, 2021] talk about *worldviews*. In particular, [Friedler *et al.*, 2021] 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 or skills. 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. If WYSIWYG is assumed, non-discrimination is guaranteed as soon as the mapping between OS and DS is fair, since WYSIWIG assumes CS $\approx$ OS. In contrast, according to WAE, the mapping between CS and OS is distorted by some bias whenever an observable difference among groups emerges (this difference is called measurement bias in [Hertweck *et al.*, 2021]); therefore, to obtain a fair mapping between CS and DS those biases should be mitigated properly. In their paper, [Hertweck *et al.*, 2021] build upon the work of [Friedler *et al.*, 2021] and present a more detailed scenario by introducing the concept 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*.

With a different perspective, [Suresh and Guttag, 2021] 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. 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). Indeed, 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, while the latter depend on users' decisions in handling the data. Proper mitigation relies on the comprehension of the biases that affect the data generation and should be determined through both technical and philosophical considerations.

2.1 Fundamental Types of Bias

We now introduce what we consider the core building blocks of most types of bias involved in data generation, namely: *historical bias*, *measurement bias*, and *representation bias*.

Biases going *from User to Data* (UtD) impact the phenomenon to be studied and thus the dataset, instead biases going *from Data to Algorithm* (DtA) impact the dataset but not the phenomenon itself [Mehrabi *et al.*, 2021].

Historical bias (UtD). 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 due to long-lasting social barriers and not reflecting intrinsic differences among genders. A similar situation may arise when dependence on sensitive individual characteristics is present with respect to the variable that we are trying to 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.

Measurement bias (DtA). Occurs when a proxy of some variable relevant to a specific goal or target is employed, and that proxy is 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. Incidentally, 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" in the data.

Representation bias (DtA). Occurs when data are not representative of the 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. In the latter case, representation bias may be sufficient to create inequalities in decision-making processes based on that data.

The above list of biases should be seen as the set of the most important mechanisms through which unfairness can leak into ML-based decision-making systems due to the used 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 Dataset Generation

We propose a simple modelling framework to simulate the bias described in Section 2.1. 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 2.1, following [Mehrabi *et al.*, 2021], we can distinguish between *from user to data* and *from data to algorithm* biases. 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:

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

where g and h represent the collection and measurement of relevant individual attributes and outcomes. The use of $(\widetilde{X}, \widetilde{Y})$ rather than (X, Y) describes the fact that the set of variables employed to make inferences about a phenomenon may not coincide with the actual variables that play a role in that phenomenon. This is precisely what happens in some forms of bias. Notice that UtD types of bias impact directly Equation (1), while DtA biases affect the data observation process described in Equation (2).

Our framework is in line with that proposed by [Suresh and Guttag, 2021]. In the following, we propose a simple and explicit mathematical formalisation of the framework, using the following notation: R are variables representing *resources* of individuals 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.

Historical bias occurs when the relevant variable R is somehow impacted by sensitive feature 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* impacted by A. Measurement bias could also occur on the target variable Y when we can only access a (biased) proxy P_Y of the phenomenon we want to predict.

The following system of Equations formalises the relationships between variables used to simulate specific forms of biases. Notice that the independent random variables $N_\cdot$ and $B_\cdot$ are continuous-valued and integer-valued, respectively. 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 $\{0, \ldots, K\}$, 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 generate a binary target Y by thresholding S. The magnitude of the historical bias (on the features or labels) is denoted by the variable β_h^i for $i \in \{R, Q, Y\}$.

$$A = B_A, \quad B_A \sim \mathcal{B}er(p_A); \tag{3a}$$

$$R = -\beta_h^R A + N_R, \quad N_R \sim Gamma(k_R, \theta_R); \tag{3b}$$

$$Q = B_Q, \quad B_Q \mid (R, A) \sim \mathcal{B}in(K, p_Q(R, A)), \tag{3c}$$

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

$$S = \alpha_R R - \alpha_Q Q - \beta_h^Y A + N_S, \quad N_S \sim \mathcal{N}(0, \sigma_S^2); \tag{3d}$$

$$Y = \mathbf{1}_{\{S > \overline{P_S}\}}. \tag{3e}$$

When simulating measurement bias (denoted by $\beta_m^j > 0$ for $j \in \{R, Y\}$), either on resources R or on target Y,[3] we are going to use the following *proxies* as noisy (and biased) substitutes for the actual variables:

$$P_R = R - \beta_m^R A + N_{P_R}; \quad N_{P_R} \sim \mathcal{N}(0, \sigma_{P_R}^2); \tag{4a}$$

$$P_S = S - \beta_m^Y A + N_{P_S}; \quad N_{P_S} \sim \mathcal{N}(0, \sigma_{P_S}^2); \tag{4b}$$

$$P_Y = \mathbf{1}_{\{P_S > \overline{P_S}\}}. \tag{4c}$$

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. Additionally, in order to account for *representation bias*, we undersample the group $A = 1$ conditioned on R by selecting the $A = 1$ individuals with lower values for R (governed by the parameter p_u).

4 Demo and Call for Further Work

The `Bias on Demand` demonstration is available at tinyurl.com/biasondemand and consists of a set of different synthetic datasets that are coupled with bias mitigation techniques (such as the ones proposed by [Hardt *et al.*, 2016; Corbett-Davies *et al.*, 2017; Baumann *et al.*, 2022]). This allows us to investigate the effects of different types of bias on the outcomes of ML-based decision making systems (measured through standard performance and fairness metrics proposed by the algorithmic fairness community [Castelnovo *et al.*, 2022b; Verma and Rubin, 2018]).

This work aims to raise awareness of bias in AI-based systems and its potential impacts on individuals and society, promoting the development of systems that are consistent with the universal ethical principle of non-discrimination.

A large set of experiments, as well as the code to create new ones, is publicly available at `BiasOnDemand`. The package can be installed via `pip` and used to generate synthetic datasets with various types of bias in just a few lines of code.

We lay the groundwork for developing novel tools and strategies, such as systems for detecting and identifying different types of bias, as well as implementing specific bias mitigation techniques. We hope that our toolkit will encourage the research community to undertake further studies using synthetic datasets where real-world datasets are lacking.

[3] Notice that the labels Y are a binary realisation of S (of its proxy P_S for P_Y, respectively). We use the distribution mean of P_S, denoted by $\overline{P_S}$, to derive binary values for Y and its proxy P_Y.

References

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

[Baumann *et al.*, 2022] Joachim Baumann, Anikó Hannák, and Christoph Heitz. 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, pages 2315–2326, New York, NY, USA, 2022. Association for Computing Machinery.

[Baumann *et al.*, 2023] Joachim Baumann, Alessandro Castelnovo, Riccardo Crupi, Nicole Inverardi, and Daniele Regoli. Bias on Demand: A Modelling Framework That Generates Synthetic Data With Bias. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, New York, NY, USA, 2023. Association for Computing Machinery.

[Castelnovo *et al.*, 2020] Alessandro Castelnovo, Riccardo Crupi, Giulia Del Gamba, Greta Greco, Aisha Naseer, Daniele Regoli, and Beatriz San Miguel Gonzalez. Befair: Addressing fairness in the banking sector. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3652–3661. IEEE, 2020.

[Castelnovo *et al.*, 2021] Alessandro Castelnovo, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Andrea Cosentini. Towards fairness through time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 647–663. Springer, 2021.

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

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

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

[Corbett-Davies *et al.*, 2017] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 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, pages 797–806, New York, NY, USA, 2017. Association for Computing Machinery.

[Crupi *et al.*, 2022] Riccardo Crupi, Alessandro Castelnovo, Daniele Regoli, and Beatriz San Miguel Gonzalez. Counterfactual explanations as interventions in latent space. *Data Mining and Knowledge Discovery*, pages 1–37, 2022.

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

[Gujar *et al.*, 2022] Shubham Gujar, Tanishka Shah, Dewen Honawale, Vedant Bhosale, Faizan Khan, Devika Verma, and Rakesh Ranjan. Genethos: A synthetic data generation system with bias detection and mitigation. In *2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)*, pages 1–6. IEEE, 2022.

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

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

[Howe *et al.*, 2017] Bill Howe, Julia Stoyanovich, Haoyue Ping, Bernease Herman, and Matt Gee. Synthetic data for social good. *arXiv preprint arXiv:1710.08874*, 2017.

[Hu *et al.*, 2019] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 259–268, 2019.

[Hutchinson and Mitchell, 2019] Ben Hutchinson and Margaret Mitchell. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 49–58, 2019.

[Le Quy *et al.*, 2022] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, page e1452, 2022.

[Mehrabi *et al.*, 2021] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[Ntoutsi *et al.*, 2020] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1356, 2020.

[Pagan *et al.*, 2023] Nicolò Pagan, Joachim Baumann, Ezzat Elokda, Giulia De Pasquale, Saverio Bolognani, and Anikó Hannák. A Classification of Feedback Loops and Their Relation to Biases in Automated Decision-Making Systems. *arXiv preprint arXiv:2305.06055*, 2023.

[Raghunathan, 2021] Trivellore E Raghunathan. Synthetic data. *Annual Review of Statistics and Its Application*, 8:129–140, 2021.

[Suresh and Guttag, 2021] Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, New York, NY, USA, 2021. Association for Computing Machinery.

[Verma and Rubin, 2018] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.