An Open-Source Toolkit to Generate Biased Datasets

Joachim Baumann[1,2,*], Alessandro Castelnovo[3,4,†], Riccardo Crupi[4,†],
Nicole Inverardi[4,†] and Daniele Regoli[4,†]

[1] *University of Zurich, Zurich, Switzerland*
[2] *Zurich University of Applied Sciences, Zurich, Switzerland*
[3] *Dept. of Informatics, Systems and Communication, Univ. Milano Bicocca, Milan, Italy*
[4] *Data Science & Artificial Intelligence, Intesa Sanpaolo, Milan, Italy*

Abstract

Many different types of bias are discussed in the algorithmic fairness community. A clear understanding of those biases and their relation to fairness metrics and mitigation techniques is still missing. We introduce `Bias on Demand`: a modelling framework to generate synthetic datasets that contain various types of bias. Furthermore, we clarify the effect of those biases on the accuracy and fairness of ML systems and provide insights into the trade-offs that emerge when trying to mitigate them. We believe that our open-source package will enable researchers and practitioners to better understand and mitigate different types of biases throughout the ML pipeline. The package can be installed via `pip` and the experiments are available at https://github.com/rcrupiISP/BiasOnDemand. We encourage readers to consult the full paper [1].

Keywords

bias, fairness, synthetic data, bias mitigation

Problem statement Systems based on Machine Learning (ML) are increasingly being adopted to make decisions that might have a significant impact on people's lives [2–5]. Because these decision-making systems rely on data-driven learning, the risk is that they will systematically propagate the bias that can be introduced at different steps throughout the ML pipeline [6–8]. The risk is that the adoption of ML algorithms could amplify or introduce biases that will recur in society in a perpetual cycle [9, 10]. 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 [11].

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland

[*] Corresponding author.

[†] 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.

baumann@ifi.uzh.ch (J. Baumann); alessandro.castelnovo@intesasanpaolo.com (A. Castelnovo); riccardo.crupi@intesasanpaolo.com (R. Crupi); nicole.inverardi@intesasanpaolo.com (N. Inverardi); daniele.regoli@intesasanpaolo.com (D. Regoli)
https://www.ifi.uzh.ch/en/scg/people/Baumann.html (J. Baumann)
0000-0003-2019-4829 (J. Baumann); 0000-0001-5234-1155 (A. Castelnovo); 0009-0005-6714-5161 (R. Crupi); 0009-0006-0048-7455 (N. Inverardi); 0000-0003-2711-8343 (D. Regoli)

CEUR Workshop Proceedings (CEUR-WS.org)

Contributions We build on computer science and philosophical literature from the field of algorithmic fairness to explore fundamental types of bias. We provide a modelling framework to generate synthetic datasets that can include those biases. We use our proposed framework to investigate the interconnection between biases and their effect on performance and fairness evaluations. Furthermore, we provide some initial insights into mitigating specific types of bias through post-processing techniques [12–15].

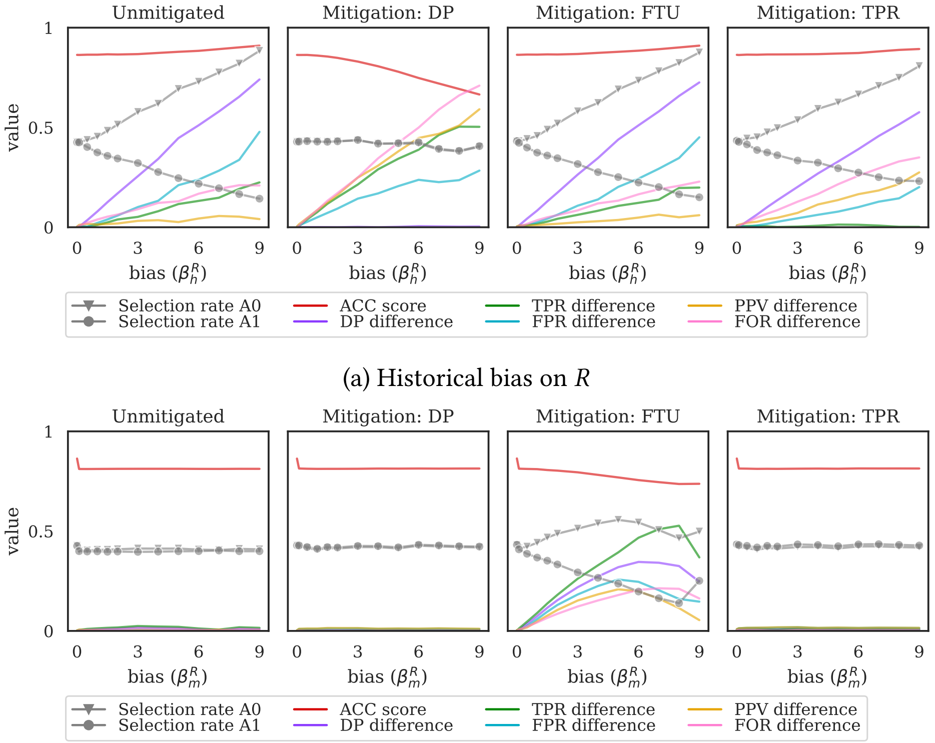
We provide a mathematical representation of the following types of bias: i) *Historical bias* – sometimes referred to as *social bias*, *life bias*, or *structural bias* [6, 7, 16] – occurs whenever a variable of the dataset relevant to some specific goal or task is dependent on some sensitive characteristic of individuals, *but in principle it should not.* ii) *Measurement bias* occurs when a proxy of some variable relevant to a specific goal or target is employed, and that proxy depends on some sensitive characteristics. iii) *Representation bias* occurs when, for some reason, data are not representative of the world population. iv) *Omitted variable bias* may occur when the collected dataset omits a variable relevant to some specific goal or task. v) *Algorithmic bias* may occur whenever the algorithmic outcomes affect the behaviour of users. i.e. the bias is generated purely by the algorithm using unbiased data.[1] vi) *Deployment bias* arises if the process followed to take decisions based on the algorithm's prediction results in harmful downstream consequences. Further details, the full set of experiments and discussions about our data generation framework can be found in the full paper [1].

Showcase As a simple example, we generate datasets for different magnitudes of historical bias and measurement bias on the features R, denoted by β_h^R and β_m^R. Figure 1 shows the effects of those biases w.r.t. different performance and fairness metrics and for applying various bias mitigation techniques: *DP*, *FTU*, and *TPR parity*. In line with [17–20], we find that *FTU* should be applied with particular care, despite its simplicity. *FTU* has no effect whatsoever since the information on group membership is redundantly encoded in R (see Figure 1a). However, 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 1b). Similarly, Figure 2 shows the results for different magnitudes of historical bias and measurement bias on the labels Y, denoted by β_h^Y and β_m^Y. The results of historically biased Y are comparable to the ones of historically biased features, except for *FTU*, which results in *DP* (see Figure 2a). As Figure 2b shows, the case of measurement bias on Y is particularly subtle: having access only to a (biased) proxy of Y, it is only possible to control the bias when imposing fairness criteria that do not use the target variable, namely *DP* and *FTU* in our experiments. All examples show that there are trade-offs between fairness and accuracy as well as between different fairness criteria.[2]

Outlook This work aims to raise awareness of bias in artificial intelligence (AI) systems and its potential impacts on individuals and society, promoting the development of *bias-free* AI systems. This is in line with the European Union' Proposal for a regulation laying down harmonised rules on AI (AI Act) [21]. By exploiting our toolkit, we hope to encourage the research community to conduct further studies using synthetic datasets where real-world datasets are missing.

[1] Notice that we use *algorithmic bias* as an umbrella term for *aggregation bias*, *learning bias*, and *evaluation bias* as they are all associated with the ML model development [8].

[2] The entire and reproducible set of experiments and the code to develop new ones are available in open-source.



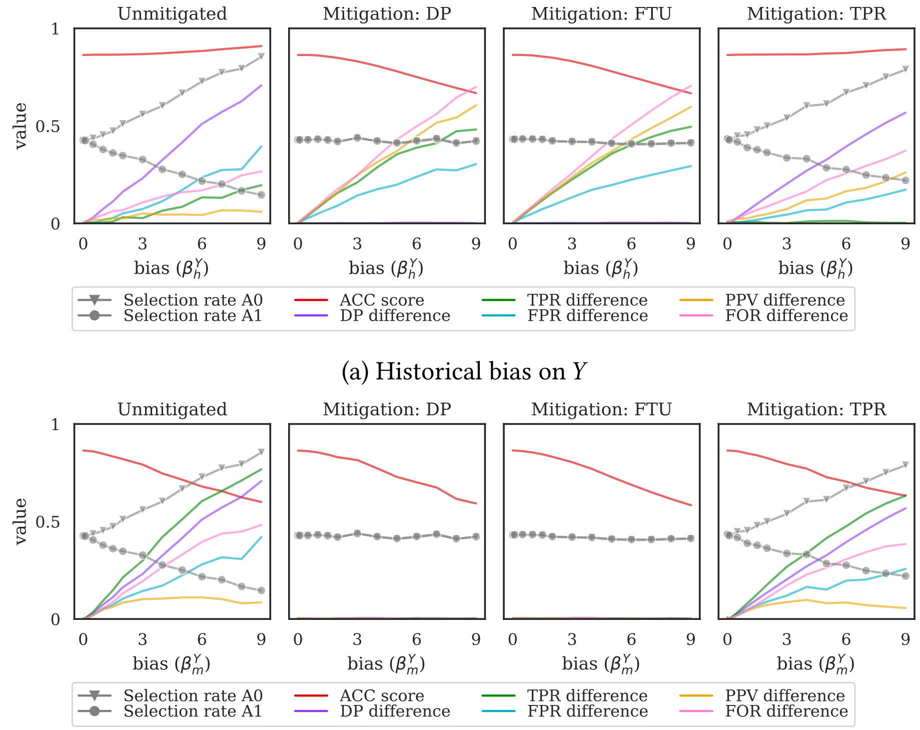
(a) Historical bias on R

(b) Measurement bias on R

Figure 1: Accuracy (ACC) and fairness metrics for biased features R. The acronyms stand for demographic parity (DP), fairness through unawareness (FTU), true positive rate (TPR), false positive rate (FPR), positive predictive value (PPV), and false omission rate (FOR).

Acknowledgments

The authors wish to thank Andrea Cosentini, Ilaria Penco and all the members of the Artificial Intelligence & Data Science Department of Intesa Sanpaolo for their commitment to supporting responsible AI initiatives. We also 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 version 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.



(a) Historical bias on Y

(b) Measurement bias on Y

Figure 2: Accuracy and fairness metrics for biased labels Y. Notice that all metrics in (b) are computed with respect to the “true”, unbiased target Y, which is usually not known in practice.

References

[1] J. Baumann, A. Castelnovo, R. Crupi, N. Inverardi, D. Regoli, Bias on Demand: A Modelling Framework That Generates Synthetic Data With Bias, in: 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’23, Association for Computing Machinery, New York, NY, USA, 2023. doi:10.1145/3593013.3594058.

[2] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning, fairmlbook.org, 2019. http://www.fairmlbook.org.

[3] M. Kearns, A. Roth, The Ethical Algorithm: The Science of Socially Aware Algorithm Design, Oxford University Press, Inc., USA, 2019.

[4] S. Mitchell, E. Potash, S. Barocas, A. D’Amour, K. Lum, Algorithmic fairness: Choices, assumptions, and definitions, Annual Review of Statistics and Its Application 8 (2021) 141–163.

[5] S. Barocas, A. D. Selbst, Big data’s disparate impact, Calif. L. Rev. 104 (2016) 671.

[6] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Computing Surveys (CSUR) 54 (2021) 1–35.

[7] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, et al., Bias in data-driven artificial intelligence systems—an introductory survey, Wiley Interdisciplinary Reviews: Data Mining and Knowledge

Discovery 10 (2020) e1356.

[8] H. Suresh, J. Guttag, A framework for understanding sources of harm throughout the machine learning life cycle, in: Equity and access in algorithms, mechanisms, and optimization, 2021, pp. 1–9.

[9] N. Pagan, J. Baumann, E. Elokda, G. De Pasquale, S. Bolognani, A. Hannák, A Classification of Feedback Loops and Their Relation to Biases in Automated Decision-Making Systems (2023). arXiv:2305.06055.

[10] A. Castelnovo, R. Crupi, G. Del Gamba, G. Greco, A. Naseer, D. Regoli, B. S. M. Gonzalez, Befair: Addressing fairness in the banking sector, in: 2020 IEEE International Conference on Big Data (Big Data), IEEE, 2020, pp. 3652–3661.

[11] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, I. G. Penco, A. C. Cosentini, A clarification of the nuances in the fairness metrics landscape, Scientific Reports 12 (2022) 1–21.

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

[13] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. 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, Association for Computing Machinery, New York, NY, USA, 2017, pp. 797–806. doi:10.1145/3097983.3098095.

[14] J. Baumann, A. Hannák, C. 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, Association for Computing Machinery, New York, NY, USA, 2022, pp. 2315–2326. doi:https://doi.org/10.1145/3531146.3534645.

[15] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, K. Walker, Fairlearn: A toolkit for assessing and improving fairness in AI, Technical Report MSR-TR-2020-32, Microsoft, 2020. URL: https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/.

[16] C. Hertweck, C. Heitz, M. Loi, On the moral justification of statistical parity, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 747–757. doi:10.1145/3442188.3445936.

[17] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: Proceedings of the 3rd innovations in theoretical computer science conference, 2012, pp. 214–226.

[18] F. Kamiran, T. Calders, Data preprocessing techniques for classification without discrimination, Knowledge and Information Systems 33 (2012) 1–33.

[19] I. Y. Chen, F. D. Johansson, D. Sontag, 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, 2018, pp. 3543–3554.

[20] S. Corbett-Davies, S. Goel, The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning, 2018. arXiv:1808.00023.

[21] The European Commission, 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, 2021. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence.