What If? Using Multiverse Analysis to Evaluate the Influence of Model Design Decisions on Algorithmic Fairness

Jan Simson[1], Florian Pfisterer[1] and Christoph Kern[1]

[1]Institute of Statistics, Ludwig-Maximilian University of München, Ludwigstr. 33, 80809 München, Germany

Abstract

A vast number of systems in Europe and beyond currently use algorithmic decision making (ADM) to (partially) automate decisions that have previously been done by humans. When designed well, these systems promise both more accurate and more efficient decisions all the while saving large amounts of resources and freeing up human time. When ADM systems are not designed well, however, they can lead to unfair algorithms which discriminate against parts of the population under the guise of objectivity and legitimacy. Many examples of both fair and helpful as well as discriminatory algorithms exist in the wild to date. The group they fall into typically depends on the decisions made during their design. It is therefore clearly important to properly understand the decisions that go into the design of ADM systems and how these decisions affect the fairness of the resulting system. To study this, we introduce the method of multiverse analysis for algorithmic fairness.

During the creation and design of an ADM system one needs to make a multitude of different decisions. Many of these decisions are made implicitly without knowing exactly how they will impact the final system and whether or not it will lead to fair decisions. In our proposed adaptation of multiverse analysis for ADM we plan to turn these implicit decisions made during the design of an ADM system into explicit ones. While many of these decisions apply to any machine-learning system, there are also a large number of domain- or problem-specific decisions to be made. Using the resulting decision space, we create a grid of all possible "universes" of decision-combinations. For each of these universes, the fairness of the ADM system is computed. Using the resulting dataset of possible decisions and fairness one can see how and which decisions impact fairness.

We demonstrate how multiverse analyses can be used to better understand variability and robustness of algorithmic fairness using an exemplary case study of predicting public health coverage. We show how small decisions during the design of an ADM system can have surprising effects on its fairness and how to detect them using multiverse analysis.

Keywords
multiverse analysis, algorithmic fairness, automated decision making, robustness, reliable machine learning

1. Extended Abstract

Across the world, more and more decisions are being made with the support of algorithms, so called algorithmic decision making (ADM). Examples of such systems can be found in finance, the labour market, criminal justice system and beyond. While these systems are very promising when designed well, raising hopes of more accurate, just and fair decisions, their impact can be quite the opposite when designed wrongly. Ample examples exist of unfair ADM systems discriminating against people in the wild, with the Dutch childcare benefits being an especially prominent and recent example [1].

While these fairness problems are often due to biases in the underlying data, gathering perfectly fair data is usually not an option, so the only way of making sure that the algorithm doesn't reinforce these biases is via the design of the ADM system. With the promise and peril of ADM systems depending so much on their proper design, it is of clear importance to properly understand the decisions that go into their design and how these decisions affect algorithmic fairness. To enable this we introduce the method of multiverse analysis for algorithmic fairness. Multiverse analyses were introduced in Psychology [2] to improve reproducibility and to combat p-hacking and cherry-picking of results. This makes them particularly useful to assess the susceptibility of ADM systems with respect to their fairness implications.

In the proposed adaptation of multiverse analysis for ADM one starts by making the many implicit decisions, also referred to as researcher degrees of freedom, during the design of an ADM system explicit. One of the differences in the present analysis compared to a classic multiverse analysis is, that we will evaluate machine learning systems in the end, whereas classical multiverse analyses will typically culminate in a null-hypothesis-significance-test (NHST). While many of the decision points apply to any machine-learning system (e.g. choice of algorithm, how to preprocess certain variables, cross-validation splits), many of them are also domain specific (e.g. coding of certain variables, how to set classification thresholds, how fairness is operationalized). While we vary certain decisions related to the training of machine learning models, our focus will not be on hyperparameter-selection or optimization. In particular we focus on decisions made during the pre-processing of data and in the translation of predictions into possible decisions. Using all possible unique combinations of these decisions we create a grid of possible *universes of decisions*. For each of these universes, we compute the fairness of the ADM system and collect it as a data point. The resulting dataset of possible decisions and resulting fairness is treated as our source data for further analysis where we evaluate how individual decisions relate back to fairness.

Existing articles in the literature have focused on specific pre-processing or modeling decisions in isolation, such as the influence of different imputation methods [3] or of model architecture and hyperparameters [4] on fairness in different contexts. Multiverse analyses have also been used to try and model the performance distribution in hyperparameter-space [5], yet not fairness. Besides multiverse analyses a highly related type of analysis emerged around the same time in the specification curve analysis [6], yet multiverse analysis seems to be the more common approach in the literature to date.

Here we present a generalizable approach of using multiverse analysis to estimate the effect of decisions during the design of an ADM system on its algorithmic fairness. We demonstrate the feasibility of this approach using a case study of predicting public health coverage in US

census data. We use the ACSPublicCoverage benchmark problem [7] of predicting public health insurance coverage, as other well-established examples have been shown to have non-trivial quality issues [7, 8, 9].

We will present preliminary results from the case study, demonstrating how plausible and seemingly small design decisions of the ADM system can have significant effects on its algorithmic fairness. We would welcome the discussion of other use cases and possible case studies, especially within the European context.

References

[1] Amnesty International, Xenophobic Machines, Technical Report, 2021. URL: https://www.amnesty.org/en/wp-content/uploads/2021/10/EUR3546862021ENGLISH.pdf.

[2] S. Steegen, F. Tuerlinckx, A. Gelman, W. Vanpaemel, Increasing transparency through a multiverse analysis, Perspectives on Psychological Science 11 (2016) 702–712. URL: https://doi.org/10.1177/1745691616658637. doi:10.1177/1745691616658637, publisher: SAGE Publications Inc.

[3] S. Caton, S. Malisetty, C. Haas, Impact of imputation strategies on fairness in machine learning, Journal of Artificial Intelligence Research 74 (2022). URL: https://doi.org/10.1613/jair.1.13197. doi:10.1613/jair.1.13197.

[4] R. Sukthanker, S. Dooley, J. P. Dickerson, C. White, F. Hutter, M. Goldblum, On the importance of architectures and hyperparameters for fairness in face recognition (2022). doi:10.48550/arXiv.2210.09943.

[5] S. J. Bell, O. P. Kampman, J. Dodge, N. D. Lawrence, Modeling the machine learning multiverse (2022). URL: https://arxiv.org/abs/2206.05985. doi:10.48550/ARXIV.2206.05985.

[6] U. Simonsohn, J. P. Simmons, L. D. Nelson, Specification curve analysis, Nature Human Behaviour 4 (2020) 1208–1214. URL: https://www.nature.com/articles/s41562-020-0912-z. doi:10.1038/s41562-020-0912-z, number: 11 Publisher: Nature Publishing Group.

[7] F. Ding, M. Hardt, J. Miller, L. Schmidt, Retiring adult: New datasets for fair machine learning (2021). URL: https://arxiv.org/abs/2108.04884. doi:10.48550/ARXIV.2108.04884.

[8] A. Fabris, S. Messina, G. Silvello, G. A. Susto, Algorithmic fairness datasets: the story so far, Data Mining and Knowledge Discovery (2022). URL: https://doi.org/10.1007/s10618-022-00854-z. doi:10.1007/s10618-022-00854-z.

[9] M. Bao, A. Zhou, S. Zottola, B. Brubach, S. Desmarais, A. Horowitz, K. Lum, S. Venkatasubramanian, It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks (2022). doi:10.48550/arXiv.2106.05498.