*HHAI 2023: Augmenting Human Intellect*
*P. Lukowicz et al. (Eds.)*

*doi:10.3233/FAIA230104*

# Using Multiverse Analysis to Evaluate the Influence of Model Design Decisions on Algorithmic Fairness

Jan SIMSON [a,1], Florian PFISTERER [a] and Christoph KERN [a]

[a] *Department of Statistics, LMU Munich*
ORCiD ID: Jan Simson https://orcid.org/0000-0002-9406-7761, Florian Pfisterer https://orcid.org/0000-0001-8867-762X, Christoph Kern https://orcid.org/0000-0001-7363-4299

**Abstract.** A vast number of systems across the world currently use algorithmic decision making (ADM) to augment human decision making or even 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 societal groups under the guise of objectivity and legitimacy. Whether systems are ultimately fair or not typically depends on the decisions made during the systems' design. It is therefore 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 they will lead to fair outcomes. 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. Using the resulting decision space, we create a grid of all possible "universes" of decision-combinations. For each of these universes, a fairness metric 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 preliminary results illustrating 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] Corresponding Author: Jan Simson, jan.simson@lmu.de.

## 1. Introduction

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 labor market, the 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. There are many examples of unfair ADM systems discriminating against people in the wild, with the Dutch childcare benefits providing an especially prominent and recent example [1].

While these fairness problems often occur because algorithms replicate biases in the underlying training data, gathering perfectly fair data is usually not feasible in practice. Biases can also originate or increase in other parts of a typical machine-learning pipeline. As a result, preventing algorithms from introducing new or reinforcing existing biases requires careful study and evaluation of the, often implicit, decisions made while designing ADM systems. To facilitate this, we introduce the method of multiverse analysis for algorithmic fairness.

Multiverse analyses originate from the field of Psychology [2] to improve reproducibility and create more robust research. This makes them particularly useful to assess the susceptibility of ADM systems with respect to their fairness implications. We adapt this methodology to work in the context of machine learning with a focus on evaluating metrics of algorithmic fairness.

## 2. Methodology

In our proposed adaptation of multiverse analysis for algorithmic fairness one starts by making the many implicit decisions required 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, 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). 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 resulting fairness metric of the ADM system and collect it as a data point. The resulting dataset of decision universes and resulting fairness is treated as our source data for further analysis where we evaluate how individual decisions relate back to metrics of fairness.

Existing work has focused on specific pre-processing or modeling decisions in isolation, such as the influence of different imputation methods [3] or of the 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], but not yet for analysing algorithmic fairness. Besides multiverse analyses a highly related type of analysis emerged around the same time in the specification curve analysis [6].

## 3. Contribution

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* dataset [7] predicting public health insurance coverage, as other well-established datasets 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 sometimes have significant effects on algorithmic fairness metrics. Results from the analysis can increase transparency and robustness of ADM systems and may be used to inform human decision makers in hybrid decision making contexts. We welcome the discussion of other use cases and possible case studies.

## References

[1] Amnesty International. Xenophobic Machines; 2021. Available from: https://www.amnesty.org/en/wp-content/uploads/2021/10/EUR3546862021ENGLISH.pdf.

[2] Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing Transparency Through a Multiverse Analysis. Perspectives on Psychological Science. 2016 09;11(5):702-12. Publisher: SAGE Publications Inc. Available from: https://doi.org/10.1177/1745691616658637.

[3] Caton S, Malisetty S, Haas C. Impact of Imputation Strategies on Fairness in Machine Learning. Journal of Artificial Intelligence Research. 2022 09;74. Available from: https://doi.org/10.1613/jair.1.13197.

[4] Sukthanker R, Dooley S, Dickerson JP, White C, Hutter F, Goldblum M. On the Importance of Architectures and Hyperparameters for Fairness in Face Recognition; 2022. Available from: https://doi.org/10.48550/arXiv.2210.09943.

[5] Bell SJ, Kampman OP, Dodge J, Lawrence ND. Modeling the Machine Learning Multiverse; 2022. Available from: https://arxiv.org/abs/2206.05985.

[6] Simonsohn U, Simmons JP, Nelson LD. Specification curve analysis. Nature Human Behaviour. 2020 11;4(11):1208-14. Number: 11 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41562-020-0912-z.

[7] Ding F, Hardt M, Miller J, Schmidt L. Retiring Adult: New Datasets for Fair Machine Learning. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Vaughan JW, editors. Advances in Neural Information Processing Systems. vol. 34. Curran Associates, Inc.; 2021. p. 6478-90. Available from: https://proceedings.neurips.cc/paper_files/paper/2021/file/32e54441e6382a7fbacbbbaf3c450059-Paper.pdf.

[8] Fabris A, Messina S, Silvello G, Susto GA. Algorithmic fairness datasets: the story so far. Data Mining and Knowledge Discovery. 2022 09. Available from: https://doi.org/10.1007/s10618-022-00854-z.

[9] Bao M, Zhou A, Zottola S, Brubach B, Desmarais S, Horowitz A, et al.. It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks; 2022. Available from: https://doi.org/10.48550/arXiv.2106.05498.