# Everything, Everywhere All in One Evaluation: Using Multiverse Analysis to Evaluate the Influence of Model Design Decisions on Algorithmic Fairness

Jan Simson Florian Pfisterer Christoph Kern

A vast number of systems across the world use algorithmic decision making (ADM) to (partially) automate decisions that have previously been made by humans. When designed well, these systems promise more objective decisions while saving large amounts of resources and freeing up human time. However, when ADM systems are not designed well, they can lead to unfair decisions which discriminate against societal groups. The downstream effects of ADMs critically depend on the decisions made during the systems' design and implementation, as biases in data can be mitigated or reinforced along the modeling pipeline. Many of these design decisions are made implicitly, without knowing exactly how they will influence the final system. It is therefore important to make explicit the decisions made during the design of ADM systems and understand how these decisions affect the fairness of the resulting system.

To study this issue, we draw on insights from the field of psychology and introduce the method of multiverse analysis for algorithmic fairness. In our proposed method, we turn implicit design decisions into explicit ones and demonstrate their fairness implications. By combining decisions, we create a grid of all possible "universes" of decision combinations. For each of these universes, we compute metrics of fairness and performance. Using the resulting dataset, 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 of vulnerable populations for potential interventions. Our results illustrate how decisions during the design of a machine learning system can have surprising effects on its fairness and how to detect these effects using multiverse analysis.

# 1 Introduction

Across the world, more and more decisions are being made with the support of machine learning (ML) and algorithms; so called algorithmic decision making (ADM). Examples of such systems can be found in finance for loan approvals (Mukerjee et al. 2002), the labor market for hiring decisions or filtering resumes (Faliagka, Ramantas, and Tzimas 2012) and the criminal justice system to assess risks of recidivism (Angwin et al. 2016). While these systems are very promising when designed well, raising hopes of more accurate and objective decisions, their impact can be quite the opposite when designed wrongly. There are many examples of ADM systems discriminating against people (Mehrabi et al. 2021). One prominent example was the *robodebt* system, where the Australian government used an algorithm to detect potential social security overpayments. Due to serious flaws in the design of the system it often overestimated debts and put the burden on the accused to prove the contrary (Henriques-Gomes 2023). Other examples include the Dutch childcare benefits system using an ADM system that was much more likely to accuse immigrants of having committed fraud (International 2021).

These fairness problems often occur because algorithms replicate biases in the underlying training data. However, biases can be amplified throughout the machine learning pipeline depending on how exactly data is processed and turned into outputs (Kern et al. 2021; Rodolfa, Saleiro, and Ghani 2020). Unfortunately, no single straightforward method exists to prevent biases in the machine learning pipeline (Agrawal et al. 2021). Understanding how modeling decisions interact with fairness is therefore a prerequisite for effectively mitigating unintended outcomes in practice. A systematic mapping of design decisions to fairness outcomes can critically guide the model selection process as multiple models may achieve similar accuracy, but can considerably differ in their fairness properties (Black, Raghavan, and Barocas 2022). 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 a machine learning system. To facilitate this objective in a systematic and efficient way, we introduce the method of multiverse analysis for algorithmic fairness. Multiverse analyses were introduced to psychology with the intent to improve reproducibility and create more robust research (Steegen et al. 2016). We adapt this methodology across domains to work in the context of machine learning with a focus on evaluating metrics of algorithmic fairness.

In the following, we present a generalizable approach of using multiverse analysis to estimate the effect of decisions during the design of a machine learning or ADM system on fairness outcomes. We demonstrate the feasibility of this approach using a case study of predicting public health coverage in US census data. We provide modular source code to allow streamlined adaptation of the proposed method in other use cases and contexts.

## 1.1 Multiverse Analysis

Multiverse analyses were first introduced in psychology by Steegen et al. (2016) in response to the reproducibility crisis affecting the field (OPEN SCIENCE COLLABORATION 2015).

The goal of this analysis type is to investigate the invariance of results to researchers' analysis decisions. Specifically, when analyzing a dataset, researchers make many implicit and explicit choices (Simmons, Nelson, and Simonsohn 2011), often without the option of confirming whether a choice is correct or incorrect. This leads to many plausible scenarios when analyzing data, as one traverses a *garden of forking paths* (Gelman and Loken 2014), where each fork corresponds to a decision. The multitude of these scenarios becomes especially evident when multiple researchers analyze the same data, coming to staggeringly different results (Breznau et al. 2022).

Multiverse analysis focuses on the pre-processing steps applied to a dataset: Steps such as selecting the observations and predictor variables to include in a dataset or scaling and binning their values. Based on the different decisions made and paths taken when pre-processing a dataset, analysts will end up with one of many possible datasets for the actual analysis. In a multiverse analysis, the goal is to make this variation explicit by using the complete grid of decisions and their options to generate all plausible datasets. Using all potential datasets, a multiverse analysis re-runs the analysis on each of them to receive the distribution of results instead of a single result point (Figure 1, Steps 1 - 3). We adapt this methodology for the machine learning context with a special focus on using it to generate insights on metrics of algorithmic fairness.

Besides multiverse analyses, a highly related type of analysis emerged around the same time in the specification curve analysis (Simonsohn 2018). Similarly to a multiverse analysis, a specification curve analysis uses the complete grid of possible decision combinations to estimate the variability of research findings. This data is used to create a specification curve, a graph displaying the distribution of the effect size or coefficient of interest using a single curve. All decision combinations are ordered by the metric creating a single, monotonically increasing curve. The actual decisions are one-hot encoded and displayed as a binary rug below the curve, clearly identifying all options that produced a certain value on the curve. Both multiverse analysis and specification curve analysis focus on bringing the diversity in pre-processing (and beyond) to light and we incorporated the essential ideas from both approaches.

## 1.2 Multiverse Analysis for Algorithmic Fairness

In our proposed adaptation of multiverse analysis for algorithmic fairness, one starts by compiling a list of all potentially relevant decisions that are being made during the design of a particular system. We differentiate between different kinds of decisions in this context: (1) decisions which are already made explicitly with a consideration of their different options e.g. choice of model and its hyperparameters, and (2) decisions which are made explicitly, but without any consideration for alternatives e.g. log-transforming an income column because it is common practice. In a multiverse analysis, the goal is to turn both types of decisions into completely explicitly made decisions and evaluate their impacts. There are also decisions which may initially not even be considered as such e.g. modifying classification cutoffs post-hoc due to external constraints. Conducting a multiverse analysis invites reflection on the

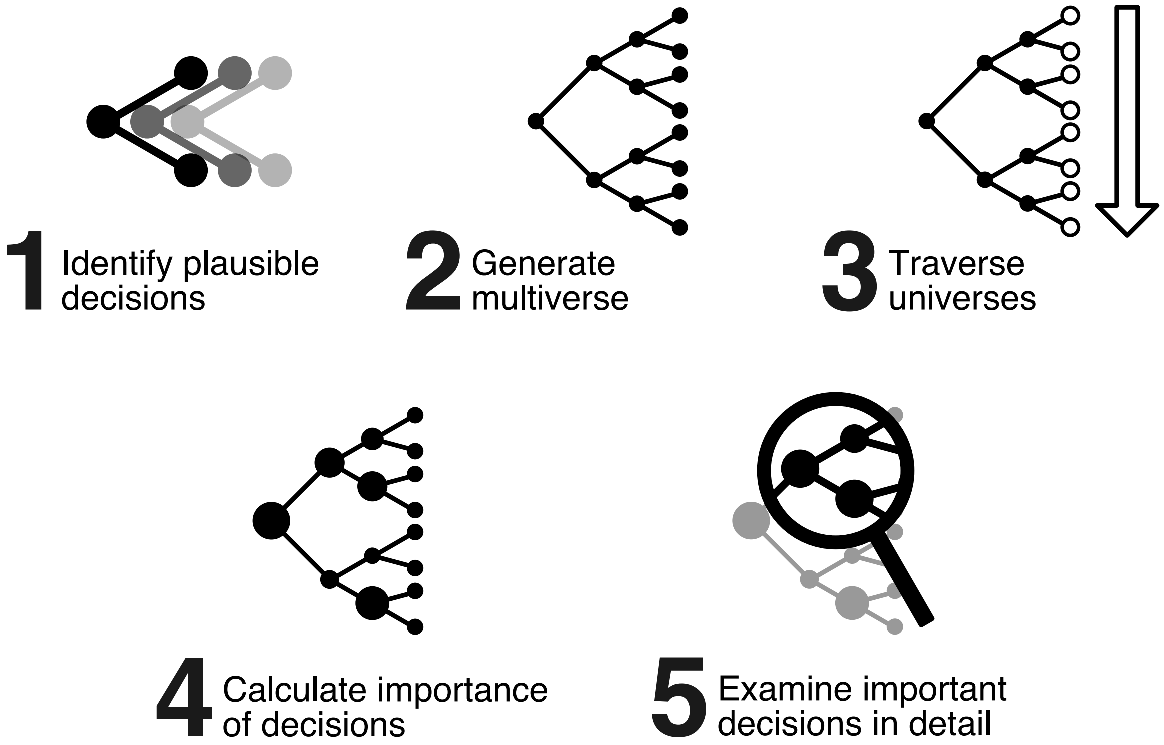
1 Identify plausible decisions
2 Generate multiverse
3 Traverse universes
4 Calculate importance of decisions
5 Examine important decisions in detail

Figure 1: **Steps to conduct a multiverse analysis for algorithmic fairness.** Steps 1 - 3 apply to multiverse analyses in general, whereas steps 4 - 5 are unique to multiverse analyses for algorithmic fairness.

modeling pipeline such that implicit decisions may surface and are turned into explicit ones. One of the key 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 evaluate the outcomes of null-hypothesis-significance-tests (NHST) across analysis choices. 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). We focus on decisions made during the pre-processing of data, in line with the original approach of multiverse analysis (Steegen et al. 2016). We extend this approach to incorporate decisions relevant to algorithmic fairness, particularly with regard to protected attributes and the translation of predictions into real-world actions or interventions. Similarly to a classical multiverse analysis, we use the resulting *garden of forking paths* to generate a grid of all possible universes of decisions combinations, the multiverse. For each of these universes, we compute the resulting fairness metric of the machine learning system and collect it as a data point. Based on the resulting dataset of decision universes and corresponding fairness scores, we evaluate how individual decisions influence the fairness metric and explore the most important decisions in more detail (Figure 1).

### 1.2.1 Related Research

Existing work has described the effects of specific pre-processing or modeling decisions in isolation, such as the influence of different imputation methods (Caton, Malisetty, and Haas 2022), of the model architecture and hyperparameters (Sukthanker et al. 2022) on fairness in different contexts. Multiverse analyses have also been used to model the performance distribution in hyperparameter-space (Bell et al. 2022), but not yet for analyzing algorithmic fairness.

The field of hyperparameter-optimization (HPO) (Feurer and Hutter 2019; Bischl et al. 2023), tries to optimize the process of tuning machine learning model hyperparameters. This field typically focuses on optimizing algorithm performance by employing efficient search strategies. These search strategies allow achieving higher performance without requiring exploration of the complete hyperparameter space. However, such search patterns usually focus on finding the optimal configuration and usually yield non-i.i.d. optimization traces. This makes them unsuitable for assessing the influence and robustness of any particular decisions. While algorithmic fairness and the importance of pre-processing are also explored (Perrone et al. 2021), they are typically only examined on the sidelines. Here, we draw on insights and methodology from the field of HPO, in particular the functional analysis of variance (FANOVA) (Hooker 2007; Hutter, Hoos, and Leyton-Brown 2014) to allow a more interpretable and efficient analysis of the results from the multiverse analysis. Our focus, however, is on uncovering and systematically exploring variation induced by the different decisions instead of finding the setting that optimizes fairness metrics.

## 1.3 Case Study

We illustrate how multiverse analysis can enrich the machine learning fairness toolkit using a case study of predicting public health insurance coverage. Accurate and fair prediction of public health insurance coverage in the United States is an important issue as access to healthcare is quite expensive in the US, with the country spending almost 16% of its gross domestic product per capita on healthcare (Ortiz-Ospina and Roser 2017). Whether or not someone is covered by health insurance has large effects on their health and financial situation: People with insurance have better self-reported health, have more preventative doctors appointments, improved depression outcomes, and fewer personal bankruptcies (Sommers, Gawande, and Baicker 2017).

Given the complexity of US public healthcare system it is easy for people to fall through its cracks, missing the chance for coverage due to not understanding how it works or by not fulfilling all necessary requirements. To combat this, one might want to set up either financial or informational interventions targeting low-income individuals who are at risk of not being covered by public health insurance. Determining who and who not to target with such an intervention could be facilitated with an ADM system. However, given the vulnerability of the target population and ethical implications of distributing interventions, fair and well-calibrated predictions are of the utmost importance in such a scenario.

We implement our case study using the ACSPublicCoverage dataset (Ding et al. 2021). We use this particular dataset as it is rich enough for us to implement a wide range of design decisions and because many other well-established datasets used in the fairness literature suffer from non-trivial quality issues (Ding et al. 2021; Fabris et al. 2022; Bao et al. 2022): UCI Adult (Kohavi and Becker 1996), the most popular dataset in the fairness literature (Fabris et al. 2022), uses an arbitrary threshold of 50,000 USD to create a binary task of income prediction. This threshold has been shown to greatly influence the accuracy of predictions in certain groups, biasing measures of algorithmic fairness and threatening external validity (Ding et al. 2021). The ACSPublicCoverage dataset is one of the datasets which have been specifically developed in response to the issues in UCI Adult.

Here, we operationalize having public insurance coverage as being covered by either Medicare, Medicaid, Medical Assistance (or any kind of government-assistance plan for those with low incomes or a disability) or Veterans Affairs Health Care, following the official Guidance for Health Insurance Data Users from the US Census Bureau (Bureau 2021). In line with the original task setup by Ding et al. (2021), only individuals with an age below 65 years and a yearly income of less than $30,000 are examined. Low-income households are also more likely to rely on public health insurance (Keisler-Starkey and Bunch 2022).

As there are no clear guidelines on how to set up an ADM system within this context (as would be the case in heavily regulated contexts such as credit scoring) one faces a multitude of decisions when designing a solution for this task, each of which can govern how bias is fed into the final system. A multiverse analysis for algorithmic fairness requires developers to make these design decisions explicit and shows their fairness implications in the present context.

# 2 Methodology

## 2.1 Fairness Metric

While our proposed analysis works with multiple different fairness metrics, it requires one to choose a primary metric for analysis. For the present case study we used *equalized odds difference* (Agarwal et al. 2018; Hardt et al. 2016) as the primary fairness metric, as it quantifies the degree to which a system's predictions are equally good across different groups defined by a protected attribute. Equalized odds require both the *true positive rate* (TPR) and the *false positive rate* (FPR) of a system's predictions to be equal across all groups of the protected attribute. Values of the *equalized odds difference* can range from 0 to 1. A value of 0 corresponds to a perfectly fair model according to the metric, whereas a value of 1 corresponds to a completely unfair model. We use the implementation from the fairlearn package (Bird et al. 2020) to calculate the metric, where the differences in both the *true positive rate* and the *false positive rate* are calculated and the larger of the two is used as the metric. We consider *race* as the protected attribute in our case study given the persisting racial disparities in various domains, including health outcomes, in the US (Obermeyer et al. 2019).

## 2.2 Decision Space

When conducting a multiverse analysis, the first step is the identification of relevant and plausible decisions to be made. Based on the literature on data science and machine learning workflows (Kuhn and Johnson 2020; Le Quy et al. 2022) we identified five distinct categories to structure and guide the identification of decisions: Data Selection, Preprocessing, Modeling, Evaluation and Post-Hoc decisions (Table 1).

For this case study, we considered 10 distinct and orthogonal design decisions. Each of these decisions has two to five unique choice options, leading to a total of $N = 122880$ combinations of decisions or universes. An overview of the decisions and their respective options can be seen in Table 1, and a detailed description of each is provided below. We consider decisions roughly in the order they would be made during a typical analysis and sort them under the list of typical decision categories. As there is a potentially infinite list of possible decisions to consider, the present list is not intended to be exhaustive, but rather to highlight the most common and important categories of decisions one may typically encounter when designing a machine learning or ADM system. We also deliberately set the focus on decisions where alternative options are typically not considered or ones that are not identified as decisions at all. When adapting the methodology to a new system, this list can serve as an inspiration, however, one must also consider the domain-specific decisions unique to each applied problem.

Table 1: Overview of the typical decision categories, the actual decisions examined in the case study and their respective options used to construct the multiverse.

| | *Decisions and Options Examined in Case Study* | |
|---|---|---|
| Category | Decision | Options |
| **Data Selection** | Exclude Features | (1) none; (2) race; (3) sex; (4) race-sex |
| | Exclude Subgroups | (1) keep-all; (2) drop-smallest-1; (3) drop-smallest-2; (4) keep-largest-2; (5) drop-other |
| **Preprocessing** | Scale | (1) do-not-scale; (2) scale |
| | Preprocess Age | (1) none; (2) bins-10; (3) quantiles-3; (4) quantiles-4 |
| | Preprocess Income | (1) none; (2) bins-10000; (3) quantiles-3; (4) quantiles-4 |
| | Encode Categorical | (1) one-hot; (2) ordinal |
| **Modeling** | Model | (1) logreg; (2) rf; (3) gbm; (4) elasticnet |
| **Evaluation** | Stratify Split | (1) none; (2) target; (3) protected-attribute; (4) both |
| | Fairness Grouping | (1) majority-minority; (2) separate |
| **Post-Hoc** | Cutoff | (1) raw-0.5; (2) quantile-0.1; (3) quantile-0.25 |

### 2.2.1 Data Selection

#### 2.2.1.1 Excluding Variables as Predictors (Exclude Features)

Selecting features to train a model on presents a critical design decision. In the ADM context, it can be required to exclude certain protected features (such as sex/gender, race, ethnicity) as predictors due to legal constraints when designing a machine-learning system. However, as prominently shown in various studies this does not necessarily lead to increased fairness, as the protected attribute is often correlated with other ("legitimate") features (Weerts 2021).

We implement the following options for this decision in our case study: (1) use all features as predictors (incl. protected ones), (2) exclude race, the protected attribute in the case study, (3) exclude sex, a sensitive attribute and (4) exclude both race and sex as protected / sensitive attributes.

#### 2.2.1.2 Excluding Subgroups of the Protected Attribute (Exclude Subgroups)

When working with variables with an uneven distribution or very rare categories one may focus only on the most common groups, dropping data for smaller ones. This can be done to preserve the privacy of small groups or out of convenience to allow for an easier model interpretation downstream. However, the exclusion of subgroups of the population can potentially be harmful, with discriminatory differences in downstream model predictions. While we decided to include this practice as a decision in our analysis to (1) raise awareness of the issue and (2) represent the effects of the practice in our analysis, this should not be taken as an endorsement of this practice.

We try to capture the implications of this practice via the attribute race. We therefore chose to include a decision of dropping certain groups from the training data based on their prevalence. To accurately compute the fairness metric, groups were *not* dropped from the test data used for evaluation.

We include six options for this decision, with the fraction of discarded data in brackets[1]: (1) to keep all groups (0.00%), (2) to drop the smallest group (0.01%), (3) to drop the two smallest groups (0.33%), (4) to keep the two largest groups (27.45%) and (5) to drop the category "Some Other Race alone" specifically (15.81%).

### 2.2.2 Preprocessing

#### 2.2.2.1 Scaling of Continuous Variables (Scale)

It is common to scale continuous variables during preprocessing, centering them on a mean of $\mu = 0$ and standard deviation of $\sigma = 1$ (also referred to as z-scaling). Scaling may be particularly advisable if kernel-based learners are used as it typically leads to improved performance for such models.

We include two options for this decision: (1) to keep continuous variables as they are and (2) to scale continuous variables.

#### 2.2.2.2 Binning of Continuous Variables (Preprocess Age, Preprocess Income)

Another common practice is binning continuous variables, i.e., turning continuous variables into ordinal variables with discrete categories. The reasons to do this are plentiful: To deal with outliers, to address privacy concerns, or for a more tangible interpretation to name a few.

We provide two distinct and orthogonal decisions here on whether or how to bin the variables *age* and *income*. We include four options for the variable *age*: (1) perform no binning, (2) bin into bins of size 10, (3) bin into three evenly sized quantiles, (4) bin into four evenly sized quantiles. Likewise, we include four options for the variable *income*: (1) perform no binning, (2) bin into bins of size $10,000$, (3) bin into three evenly sized quantiles, (4) bin into four evenly sized quantiles.

[1] Fractions of discarded training data are only reported for a non-stratified train-test split, as there are only *very slight* differences in the fraction of discarded data based on stratification strategy.

#### 2.2.2.3 Encoding of Categorical Variables (Encode Categorical)

Another common pre-processing step includes transforming categorical variables into a numerical format. When doing this one typically has two options: (1) One-hot (or dummy) coding each variable with $K$ categories into $K$ (or $K-1$) new binary variables or (2) ordinally encoding each variable by assigning an integer value from 1 to $K$ for each category. Ordinal encoding is only applicable, however, for variables with a natural ordering.

For all ordinal variables (including continuous variables that have been binned), we include both options. Any variables without a natural ordering are always one-hot coded.

### 2.2.3 Modeling

#### 2.2.3.1 Model Type (Model)

A major choice when designing any statistical or machine learning system is which model type one decides to use. While there is a large number of potential models to explore here, we focused on the most commonly used ones in the context of ADM in the literature. We note that hyperparameter selection has shown to have an impact on fairness, but choose to focus on the simple case, as HPO has already been studied elsewhere (Perrone et al. 2021).

We therefore support the following model types as options for this decision: (1) logistic regression (Cox 1958), (2) random forest (Ho 1995), (3) gradient boosted machine (Friedman 2001), and (4) elastic net (Zou and Hastie 2005).

### 2.2.4 Evaluation

#### 2.2.4.1 Stratification of Train-Test Split (Stratify Split)

Training and test sets are often created by simple random splitting of the full dataset. It can be beneficial, however, to perform this split conditional on certain groupings to ensure equal representation of all labels within both the train and test sets.

We include four options for this decision: (1) to not stratify at all, using a completely random split instead, (2) to stratify using the target variable (*public coverage*), (3) to stratify using the protected attribute (*race*) and (4) to stratify using a combination of both variables.

#### 2.2.4.2 Grouping of Protected Attribute (Fairness Grouping)

When working with a fairness metric, it is necessary to specify for which groups of the protected attribute it is calculated. The present case study uses *race* as the protected attribute. For protected attributes with more than two categories, however, multiple comparisons can be computed. Depending on the application context one may, e.g., simplify these groups into the largest group (*majority*) and all other groups (*minority*)[2].

An important note regarding this decision is that it changes how the fairness metric is calculated: with two groups, the difference between those two groups is calculated, however, with more than two groups all possible differences between group-pairs are calculated and the largest difference between them is used (the standard procedure for this metric). Naturally, this has a strong influence on the fairness metric. We therefore conduct many of our later analyses separately for each of its options.

We include two options for this decision: (1) The fairness metric is computed between the *majority* group and *minority* group and (2) the fairness metric is computed as the maximum of the metric as computed between all groups of the protected attribute (*race*)[3].

### 2.2.5 Post-Hoc

#### 2.2.5.1 Cutoff for Final Classification (Cutoff)

At the end of the ML pipeline, the prediction models' (risk) scores can be used to classify new observations based on a pre-specified classification threshold. By default a threshold of 0.5 would be used with every score equal or above classified as 1 (*having coverage*) and everything below as 0 (*not having coverage*). Actual interventions, however, are often based on the ranked list of scores such that (costly) interventions are targeted at the top $X$ percent with the highest risk. With real-world scenarios often coming with resource-bound restrictions, one may for example only be able to provide an intervention for, say, 10% or 25% of the most in-need in the population. These real-world restrictions are typically not taken into account in fairness evaluations, despite having potentially devastating implications.

We therefore also consider different cutoff values for the final predictions of the system. We support the following options for this decision: (1) use the default raw cutoff value of 0.5, (2) only treat the lowest 0.1 quantile as *not having coverage*, (2) only treat the lowest 0.25 quantile as *not having coverage*.

## 2.3 Analysis

We examined the overall variation of the fairness metric and the degree to which it is explained by the outlined decisions. To estimate the importance of individual decisions and their interactions,

[2] **Majority group**: 'White alone'; **Minority group**: 'Asian alone', 'Two or More Races', 'Some Other Race alone', 'Black or African American alone', 'American Indian alone', 'Native Hawaiian and Other Pacific Islander alone', 'American Indian and Alaska Native tribes specified; or American Indian or Alaska Native, not specified and no other races' and 'Alaska Native alone'.

[3] This corresponds to the default behavior in the fairlearn library.

we used a functional analysis of variance (FANOVA) (Hooker 2007; Hutter, Hoos, and Leyton-Brown 2014). Based on the results of the FANOVA, we examined the most important decisions and interactions in detail.

As it can be costly to iterate over the complete multiverse grid, we also examined the feasibility of running our analyses on smaller subsets of the data and comparing results with analysis performed on the full dataset. The implementation for running a FANOVA by Hutter, Hoos, and Leyton-Brown (2014) used here has also been demonstrated to work well for assessing hyperparameter importance with only a subset of data from the hyperparameter grid available.

### 2.4 Technology

Analyses were conducted using Python Version 3.8 (Van Rossum and Drake 2009) and pipenv (P. M. Team 2017) for reproducibility. The Python package scikit-learn (Pedregosa et al. 2011) was used for pre-processing and fitting of models, pandas (team 2020) for loading and modification of data, folktables (Ding et al. 2021) for retrieval of data, fairlearn (Bird et al. 2020) for computation of fairness metrics, fANOVA (Hutter, Hoos, and Leyton-Brown 2014) for calculation of variable importance and papermill (contributors 2017) for parameterized computation of decision universes. This reproducible document was generated using quarto (Allaire et al. 2022), R (R. C. Team 2022) Version 4.2, the R packages from the tidyverse (Wickham et al. 2019) and ggpubr (Kassambara 2023) for generation of figures. The source code of the analyses and this publication is available at https://github.com/reliable-ai/fairml-multiverse.

## 3 Results

### 3.1 Distribution of Metrics

The multiverse analysis in our case study produced a total of $N = 122880$ values of the fairness metric across all possible analysis choices. When examining the distribution of the fairness metric across the multiverse of decisions, the large variation of the fairness metric becomes apparent, with values spanning the entire possible range of the metric from 0 to 1 (Figure 2). Overall performance of the resulting models was moderate with raw accuracies between 0.419 and 0.722. Performance and the fairness metric were almost uncorrelated with a Pearson correlation of $r = 0.082$. Raw performance varied largely based on the decision *Cutoff*, with three large clusters of similar performance (Figure 3). Within these three clusters of almost equal performance there was a large variance of the fairness metric, highlighting the opportunity to optimize algorithmic fairness without sacrificing performance.

The fairness metric is calculated differently depending on how the groupings of the protected attribute are made: for the *majority-minority* grouping only a single comparison can be made,

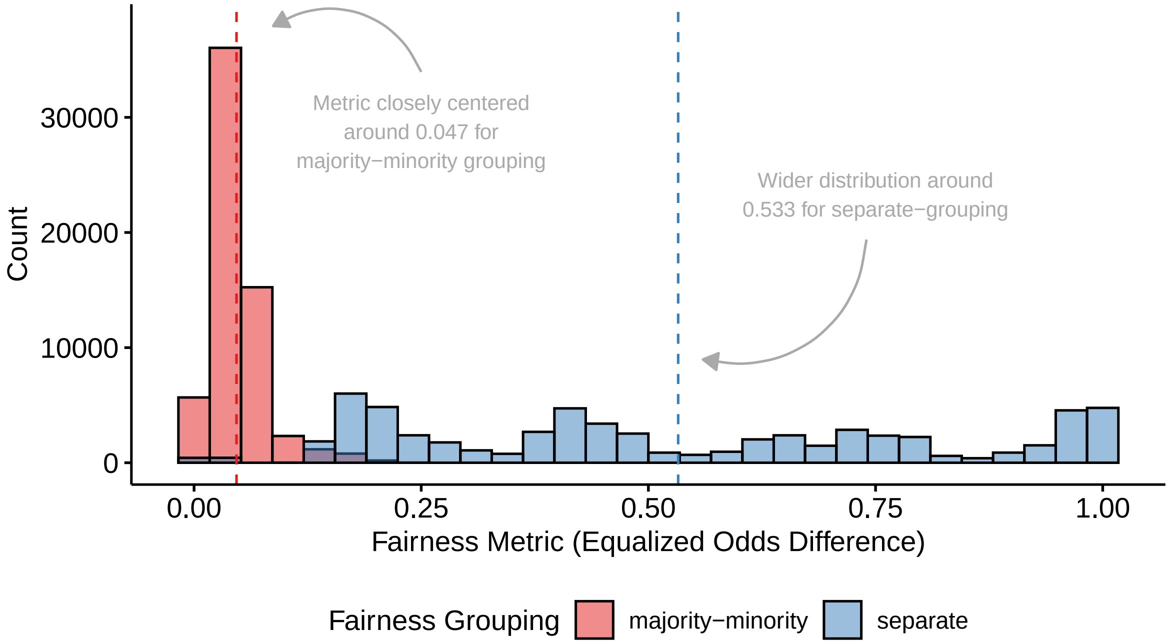

Figure 2: **Using a different grouping of the protected attribute strongly influences the fairness metric.** Distribution of fairness metric (equalized odds difference) split by grouping of the protected attribute with vertical lines corresponding to mean values. Lower values on the fairness metric indicate smaller *TPR* and *FPR* differences across groups.

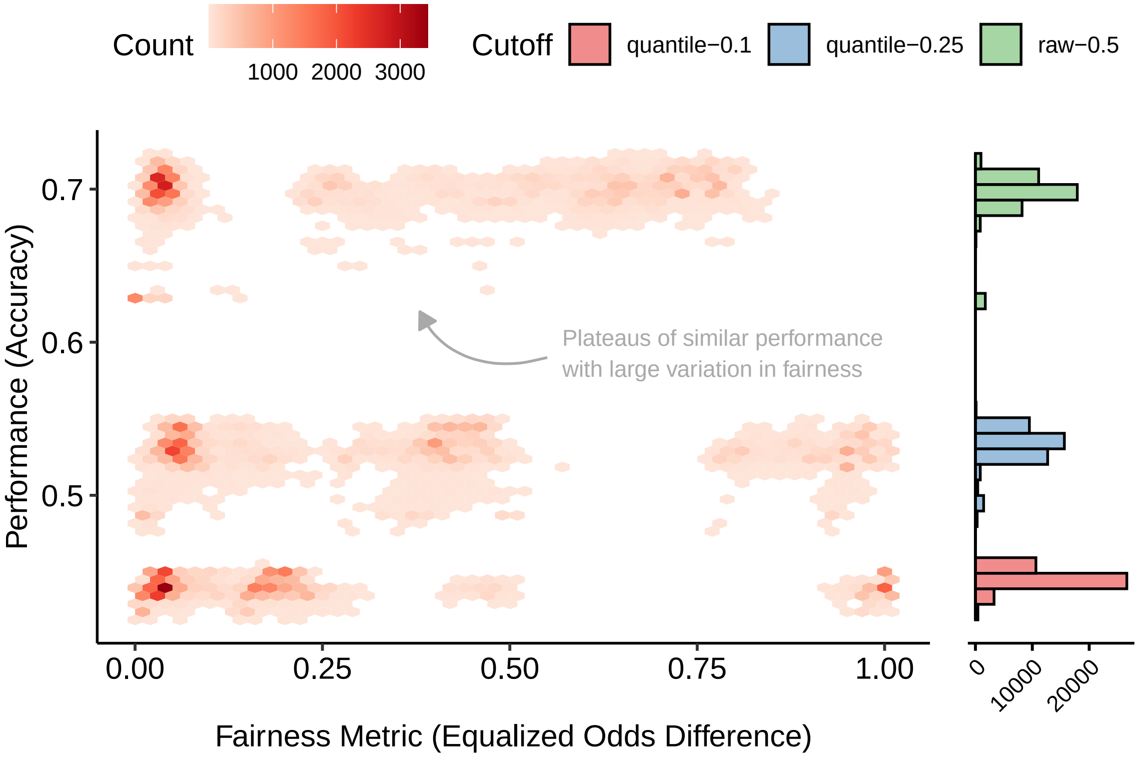

Figure 3: **Performance and fairness are largely unrelated with three plateaus of low variance in performance, but high variance in fairness.** Distribution of overall performance (raw accuracy) and fairness metric (equalized odds difference) across all multiverses. The marginal histogram shows performance for different options of the *Cutoff* decision. A marginal histogram of the fairness metric can be seen in Figure 2.

whereas the largest value from multiple comparisons is used with the *separate* grouping. Because of this, we examined this decision first and in isolation. The implications of the two groupings on the fairness metric are visible in Figure 2: The *majority-minority* grouping leads to a roughly normal distribution of low values on the fairness metric, centered around a mean of $M = 0.047$ ($SD = 0.030$). The *separate* grouping on the other hand leads to much wider distribution of fairness values centered around a mean value of $M = 0.533$ ($SD = 0.291$). Since the decision has such a strong influence on the fairness metric and interacts with many of the other decisions, we chose to conduct further analyses only within these two groupings to differentiate results more clearly and allow for easier interpretation as well as visualization. The following analyses are therefore conducted once for the *majority-minority* grouping and once for the *separate* grouping of the protected attribute; the sample size in each arm of the following analyses is thus $n = 61440$.

## 3.2 Importance of Decisions

We conducted two FANOVAs (Hooker 2007) as described in Hutter, Hoos, and Leyton-Brown (2014) to assess the importance of decisions on the fairness metric. This analysis decomposes the overall variance of the fairness metric into the fractions which are explained by each decision. These variance decompositions are used to assess the relative importance of decisions. Moreover, the FANOVA also allows computing explained variance for interactions of decisions. This is highly useful, as the overall interaction space between decisions is quite large with 511 possible (interaction and main) effects.

Using the resulting importance values from the FANOVA, one can see which decisions are associated with a high variation in fairness scores, whether it be by themselves or in conjunction with others. This allows assessing the most consequential decisions on a one-by-one case. Table 2 contains a ranked list of the most important decisions and decision interactions in our case study alongside their respective importance. When using a *majority-minorit*y grouping the most important decision is whether and which subgroups one chooses to exclude during model training, whereas the most important decisions for the *separate* grouping is how the stratification of the train-test split is performed. The cutoff value used for the final predictions is important in both cases and often has effects in conjunction with other decisions as well. Specifically the interaction of the chosen cutoff value with the stratification strategy is highly important when using the *separate* grouping, accounting for more than 30% of the variance in the fairness metric.

#### 3.2.0.1 Examining Individual Decisions

We analyzed the three most important decisions or decision-interactions per grouping approach to further illustrate the methodology and how one would explore the results of the analysis. The results also highlight why one should investigate the decisions in a detailed manner and not just pick the most-fair and highest-performing universe's model.

Table 2: The 10 most important decisions or decision interactions and their relative importance for both groupings of the protected attribute.

(a) Most important decisions for majority-minority grouping.

| Effect Type | Decision / Interaction of Decisions | Importance | Std. Deviation |
|---|---|---|---|
| main | *ExcludeSubgroups* | 0.254 | 0.001 |
| main | *Cutoff* | 0.211 | 0.001 |
| 2-way int. | *ExcludeFeatures* × *ExcludeSubgroups* | 0.110 | 0.001 |
| 2-way int. | *Cutoff* × *ExcludeSubgroups* | 0.082 | 0.000 |
| 3-way int. | *Cutoff* × *ExcludeFeatures* × *ExcludeSubgroups* | 0.027 | 0.000 |
| 2-way int. | *Model* × *PreprocessIncome* | 0.020 | 0.000 |
| main | *Model* | 0.017 | 0.000 |
| 2-way int. | *ExcludeSubgroups* × *Model* | 0.016 | 0.000 |
| main | *ExcludeFeatures* | 0.014 | 0.000 |
| 3-way int. | *Model* × *PreprocessIncome* × *Scale* | 0.012 | 0.000 |

(b) Most important decisions for separate grouping.

| Effect Type | Decision / Interaction of Decisions | Importance | Std. Deviation |
|---|---|---|---|
| main | *StratifySplit* | 0.375 | 0.001 |
| 2-way int. | *Cutoff* × *StratifySplit* | 0.313 | 0.001 |
| main | *Cutoff* | 0.082 | 0.000 |
| 4-way int. | *Cutoff* × *ExcludeFeatures* × *Model* × *StratifySplit* | 0.007 | 0.000 |
| 3-way int. | *Cutoff* × *Model* × *StratifySplit* | 0.007 | 0.000 |
| 3-way int. | *Cutoff* × *Model* × *PreprocessIncome* | 0.007 | 0.000 |
| 2-way int. | *Model* × *PreprocessIncome* | 0.006 | 0.000 |
| 2-way int. | *ExcludeFeatures* × *Model* | 0.006 | 0.000 |
| 3-way int. | *Model* × *PreprocessIncome* × *Scale* | 0.006 | 0.000 |
| 2-way int. | *Cutoff* × *PreprocessIncome* | 0.005 | 0.000 |

For the *majority-minority* grouping, *Exclude Subgroups* was the most influential decision. Examining the decision leads to surprising results: One might expect that the exclusion of any subgroups in the training data will reduce fairness (as measured by the equalized odds difference), however, results show that dropping the "Other" group from the training data actually lead to slightly fairer models on average (Figure 4 A). Whether one should actually drop the "Other" group from the training data should still be evaluated very carefully, however. The least fair models were produced when retaining only the largest two groups of the protected attribute and dropping all others. The second most important decision was which *Cutoff* value to choose when making the final predictions. Here, the default value of 0.5 typically lead to higher values of the fairness metric (Figure 4 B). It is important, to be aware of the importance of this decision, as one might be forced by practical circumstances to use a different cutoff after deployment in which case model fairness would need to be re-evaluated. Last, examining the interplay between *Exclude Subgroups* and *Exclude Features* in Figure 4 (C), illustrates how the effect of keeping only the two largest groups of race or dropping the "Other" group is greatly amplified when race is included as a predictive feature.

For the *separate* grouping, the effects are slightly more indistinct as computing the maximum fairness metric between all comparisons leads to more volatile results. The decisions *Stratify Split*, *Cutoff* and their interaction account for all three of the most important decisions with this grouping. When examining the decision separately, it can be seen how stratifying by the target variable leads to noticeably less fair models (Figure 4 D, most important) and how the raw cutoff value of 0.5 is suddenly not leading to the most fair models anymore (Figure 4 F, third most important). The effects of both variables become most clear, however, when examining their interaction, which was identified as explaining almost as much variance as the most important decision. While using a cutoff value corresponding to the top 10% quantile leads to the least fair model when stratifying by the target variable it surprisingly leads to the fairest models when using any other stratification strategy (Figure 4 E, second most important).

### 3.3 Scaling Multiverse Analysis for Algorithmic Fairness

To assess the feasibility of running the multiverse analysis on a smaller subset of the grid, we also conducted the FANOVAs on different subsamples of the collected *multiverse* dataset. Specifically, we ran the analysis on random subsets of 1%, 5%, 10% and 20% of the data and calculated the correlation of variance decomposition or importance values with the FANOVA estimated on the full multiverse dataset. The estimates of variance decomposition are highly skewed, with a few highly important decisions and a very larger number of very low-importance decisions. We therefore calculated both, the Pearson correlation which is more sensitive to correlations of the more important decisions and the Spearman rank-correlation which is also sensitive to decisions with low importance estimates. To assess the consistency of this approach we computed the FANOVA on each subsample 50 times and calculated the correlation with the results from the full *multiverse* dataset every time.

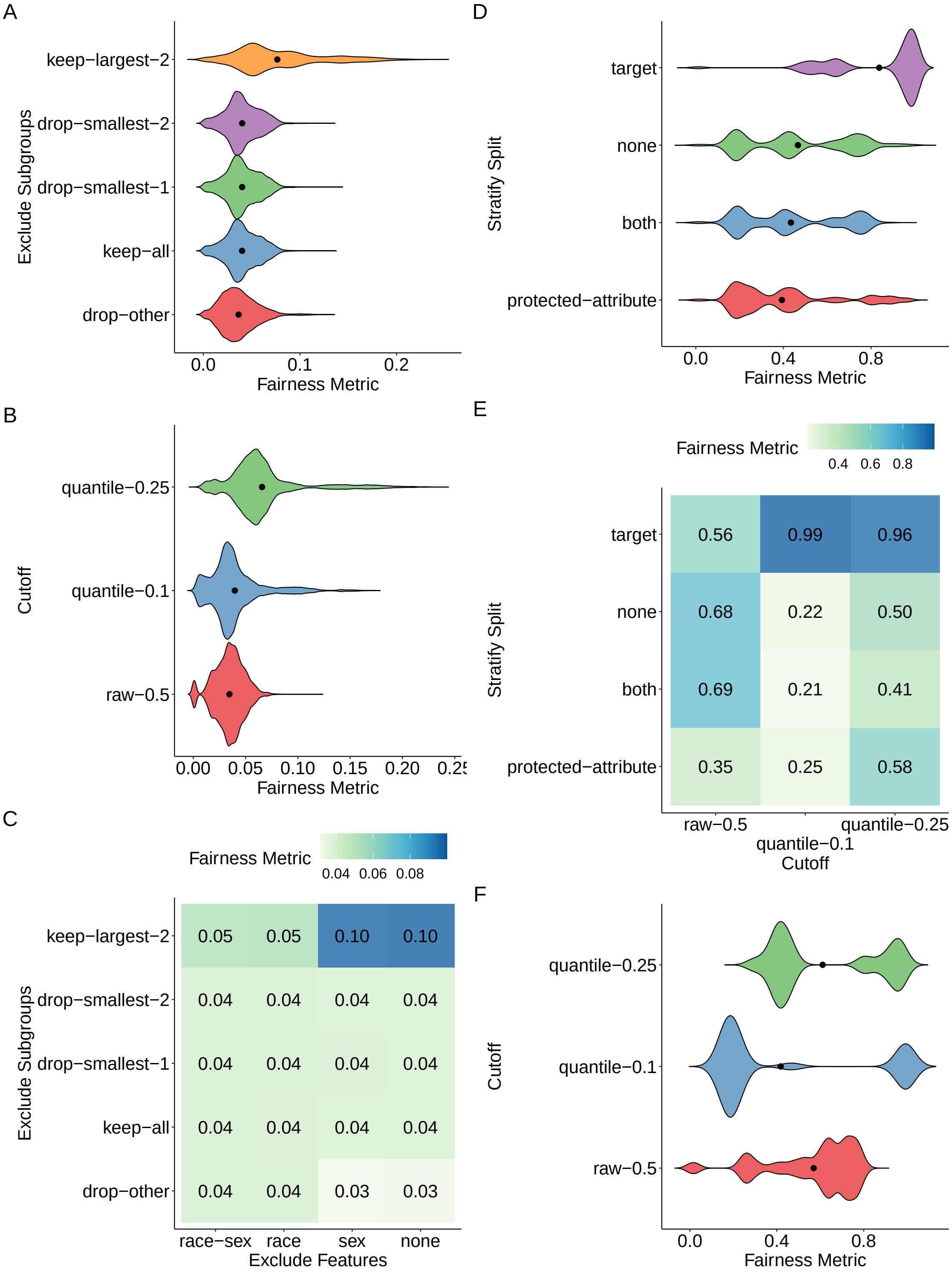

Figure 4: **Detailed analysis of most important decisions.** Visualization of the fairness metric depending on the three most important decision / decision combinations and their respective options for *majority-minority* (A-C) and *separate* (D-F) groupings of the protected attribute.

When calculating the Pearson correlation, the resulting mean correlation coefficient ranged from $\bar{r}_{1\%} = 0.989$ $(SD = 0.004)$ at 1% to $\bar{r}_{20\%} \geq 0.999$ $(SD = 0)$ at 20% for the *majority-minority* grouping and $\bar{r}_{1\%} = 0.997$ $(SD = 0.002)$ at 1% to $\bar{r}_{20\%} \geq 0.999$ $(SD = 0)$ at 20% for the *separate* grouping. Spearman rank-correlations were also high, but lower than the Pearson correlation coefficients and more inconsistent (Figure 5), which indicates that using sparse data to estimate the importance of decisions works well for the most important decisions and less-so to identify nuances between the less-important decisions. The resulting Spearman rank-correlation mean coefficients ranged from $\bar{\rho}_{1\%} = 0.517$ $(SD = 0.032)$ at 1% to $\bar{\rho}_{20\%} = 0.860$ $(SD = 0.012)$ at 20% for the *majority-minority* grouping and $\bar{\rho}_{1\%} = 0.530$ $(SD = 0.03)$ at 1% to $\bar{\rho}_{20\%} = 0.937$ $(SD = 0.007)$ at 20% for the *separate* grouping.

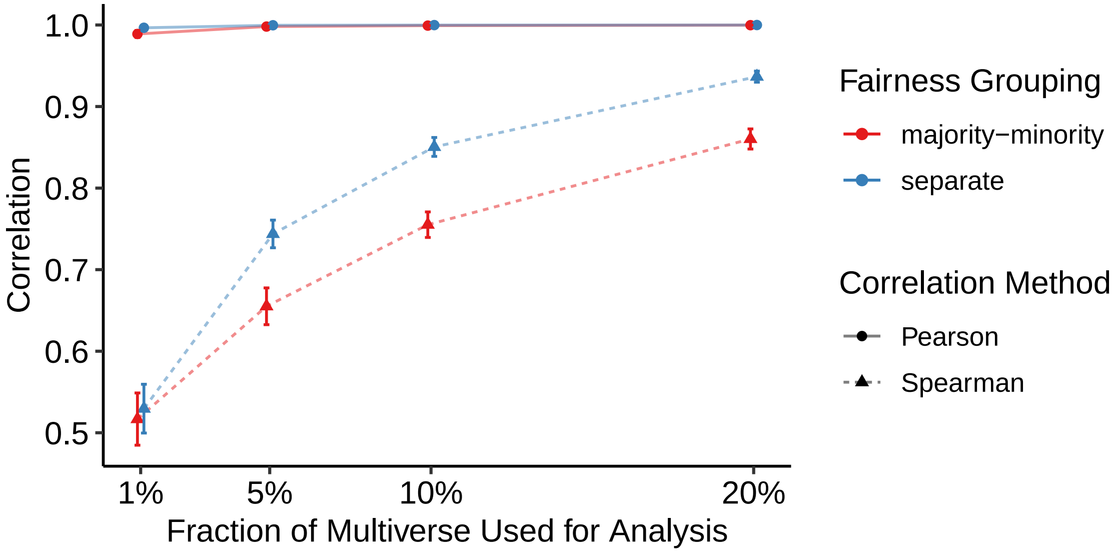

Figure 5: **Conducting the analysis with smaller subsets of the complete multiverse leads to similar results.** Correlations of variance decomposition / importance estimates between full dataset and random subsets of different sizes. Random subsets were drawn 50 times with points corresponding to mean correlations and lines to +/- 1 standard deviation. Pearson correlation coefficients are consistently higher than Spearman correlation coefficients, indicating better estimation of high-importance decisions.

## 4 Discussion

We demonstrate how multiverse analysis for algorithmic fairness provides a useful new method for evaluating the robustness of machine learning and ADM systems with respect to decisions along the modeling pipeline and their implications for algorithmic fairness. Our method provides a promising new methodology which can empower analysts to better understand how their decisions affect a system's fairness and which decisions matter. We highlight the importance of making decisions during model design explicitly rather than implicitly.

By applying this new methodology in a use case of predicting public health care coverage, we demonstrate the feasibility of this approach and show which decisions from our list of decisions affect fairness the most: Unsurprisingly the grouping of the protected attribute used for calculating the fairness metric has considerable effect on algorithmic fairness. More interestingly, this grouping does also affect the influence of almost all other decisions. Besides the grouping, we showed that the cutoff value used for making final decisions has a significant effect on the fairness metric, a decision often implemented post-hoc after model deployment without any consideration of fairness. We also observe that the exclusion of certain subgroups of the protected attribute during training affects fairness downstream, especially when the protected attribute is kept as a feature. Surprisingly, we also saw that the stratification strategy used for the train-test split had strong effects on the fairness metric.

When interpreting the results from a multiverse analysis for algorithmic fairness, one should evaluate results with care and strictly avoid merely selecting the combination of decisions with the best fairness metric. Results should be seen as an indication of how susceptible the fairness of the model is to design decisions and which decisions warrant closer examination. Results from the analysis can also be used to guide the search of new options for the most important decisions. Final choices regarding the design of the system should be made using a combination of empirical results from the multiverse analysis and practical as well as ethical considerations within the context of the use case. This can often mean that decisions may be made that do not correspond to the optimal value in the fairness metric for a decision. The main goal of a multiverse analysis for algorithmic fairness is to facilitate making educated and explicit decisions. We recommend including complete results from the analysis alongside the final system.

As we explored only a single use-case, we do not make any generalizable claims regarding the importance of any particular decisions, beyond the fact that these decisions *can* matter and are worth investigating. Another limitation of this case-study is that we only examined 10 distinct decisions, with many plausible alternative decisions which could've been examined in their place or additionally. As there is an infinite space of decisions one may consider, we decided to draw the line at these 10 decisions for illustrative purposes. A successful adoption of multiverse analysis for algorithmic fairness in different use cases and reporting of results could help identify a more exhaustive list of the most important decisions across contexts. Potential concerns regarding the computational cost of conducting a multiverse analysis for algorithmic fairness are valid, but can be addressed as we demonstrate that estimates of importance are robustly detected for important decisions even when exploring only 1% of the full *multiverse*.

We encourage the use of the method during the design of future machine learning or ADM systems and provide an overview of the most important areas of decisions to guide analysts when adapting multiverse analysis for algorithmic fairness in their own context. We further provide a non-exhaustive list of exemplary decisions to serve as inspiration to identify potentially relevant decisions and a modular implementation that makes adoption to different use-cases easy. We posit that results from a multiverse analysis for algorithmic fairness can critically inform discussions between developers and stakeholders and advise joint reflections on the ultimate design of ADM systems.

By successfully adapting multiverse analysis across disciplines we also highlight the feasibility of adapting methods and techniques between disciplines and hope that this will inspire further cross-discipline pollination of ideas and methodologies between research fields.

## 5 References

Agarwal, Alekh, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. "A Reductions Approach to Fair Classification."

Agrawal, Ashrya, Florian Pfisterer, Bernd Bischl, Francois Buet-Golfouse, Srijan Sood, Jiahao Chen, Sameena Shah, and Sebastian Vollmer. 2021. "Debiasing Classifiers: Is Reality at Variance with Expectation?" https://doi.org/10.48550/arXiv.2011.02407.

Allaire, J. J., Charles Teague, Carlos Scheidegger, Yihui Xie, and Christophe Dervieux. 2022. *Quarto*. https://doi.org/10.5281/zenodo.5960048.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias." *ProPublica*, May, 254264. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Bao, Michelle, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. 2022. "It's COMPASlicated: The Messy Relationship Between RAI Datasets and Algorithmic Fairness Benchmarks." https://doi.org/10.48550/arXiv.2106.05498.

Bell, Samuel J., Onno P. Kampman, Jesse Dodge, and Neil D. Lawrence. 2022. "Modeling the Machine Learning Multiverse." https://doi.org/https://doi.org/10.48550/arXiv.2206.05985.

Bird, Sarah, Miroslav Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, Kathleen Walker, and Allovus Design. 2020. "Fairlearn: A Toolkit for Assessing and Improving Fairness in AI."

Bischl, Bernd, Martin Binder, Michel Lang, Tobias Pielok, Jakob Richter, Stefan Coors, Janek Thomas, et al. 2023. "Hyperparameter Optimization: Foundations, Algorithms, Best Practices, and Open Challenges." *WIREs Data Mining and Knowledge Discovery* 13 (2). https://doi.org/10.1002/widm.1484.

Black, Emily, Manish Raghavan, and Solon Barocas. 2022. "Model Multiplicity: Opportunities, Concerns, and Solutions."

Breznau, Nate, Eike Mark Rinke, Alexander Wuttke, Hung H. V. Nguyen, Muna Adem, Jule Adriaans, Amalia Alvarez-Benjumea, et al. 2022. "Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty." *Proceedings of the National Academy of Sciences* 119 (44): e2203150119. https://doi.org/10.1073/pnas.2203150119.

Bureau, US Census. 2021. “ACS Health Insurance Coverage Recoding Programming Code.” https://www.census.gov/topics/health/health-insurance/guidance/programming-code/acs-recoding.html.

Caton, Simon, Saiteja Malisetty, and Christian Haas. 2022. “Impact of Imputation Strategies on Fairness in Machine Learning.” *Journal of Artificial Intelligence Research* 74 (September). https://doi.org/10.1613/jair.1.13197.

contributors, nteract. 2017. *Papermill: Parametrize and Run Jupyter and Nteract Notebooks*. https://github.com/nteract/papermill.

Cox, David R. 1958. “The Regression Analysis of Binary Sequences.” *Journal of the Royal Statistical Society: Series B (Methodological)* 20 (2): 215232.

Ding, Frances, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. “Retiring Adult: New Datasets for Fair Machine Learning,” 13.

Fabris, Alessandro, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. “Algorithmic Fairness Datasets: The Story so Far.” *Data Mining and Knowledge Discovery*, September. https://doi.org/10.1007/s10618-022-00854-z.

Faliagka, Evanthia, Kostas Ramantas, and Giannis Tzimas. 2012. “Application of Machine Learning Algorithms to an Online Recruitment System.”

Feurer, Matthias, and Frank Hutter. 2019. “Hyperparameter Optimization.” In, edited by Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren, 3–33. The Springer Series on Challenges in Machine Learning. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-05318-5_1.

Friedman, Jerome H. 2001. “Greedy Function Approximation: A Gradient Boosting Machine.” *The Annals of Statistics* 29 (5): 1189–1232. https://doi.org/10.1214/aos/1013203451.

Gelman, Andrew, and Eric Loken. 2014. “The Statistical Crisis in Science.” *American Scientist* 102 (6): 460.

Hardt, Moritz, Eric Price, Eric Price, and Nati Srebro. 2016. “Equality of Opportunity in Supervised Learning.”

Henriques-Gomes, Luke. 2023. “Robodebt: Five Years of Lies, Mistakes and Failures That Caused a $1.8bn Scandal.” *The Guardian*, March. https://www.theguardian.com/australia-news/2023/mar/11/robodebt-five-years-of-lies-mistakes-and-failures-that-caused-a-18bn-scandal.

Ho, Tin Kam. 1995. “Random Decision Forests.” In, 1:278282. IEEE.

Hooker, Giles. 2007. “Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables.” *Journal of Computational and Graphical Statistics* 16 (3): 709–32. https://www.jstor.org/stable/27594267.

Hutter, Frank, Holger Hoos, and Kevin Leyton-Brown. 2014. “International Conference on Machine Learning.” In, 754–62. PMLR. https://proceedings.mlr.press/v32/hutter14.html.

International, Amnesty. 2021. “Xenophobic Machines.” https://www.amnesty.org/en/wp-content/uploads/2021/10/EUR3546862021ENGLISH.pdf.

Kassambara, Alboukadel. 2023. *Ggpubr: 'Ggplot2' Based Publication Ready Plots*. https://CRAN.R-project.org/package=ggpubr.

Keisler-Starkey, Katherine, and Lisa N Bunch. 2022. “Health Insurance Coverage in the United

States: 2021 - Appendix Table C3." https://www.census.gov/content/dam/Census/library/publications/2022/demo/p60-278.pdf.

Kern, Christoph, Ruben L. Bach, Hannah Mautner, and Frauke Kreuter. 2021. "Fairness in Algorithmic Profiling: A German Case Study." https://doi.org/10.48550/arXiv.2108.04134.

Kohavi, Ronny, and Barry Becker. 1996. "Adult Data Set." *UCI Machine Learning Repository* 5: 2093.

Kuhn, Max, and Kjell Johnson. 2020. *Feature engineering and selection: a practical approach for predictive models*. Chapman & Hall/CRC data science series. Boca Raton London New York: CRC Press, Taylor & Francis Group. www.feat.engineering.

Le Quy, Tai, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. 2022. "A Survey on Datasets for Fairness-Aware Machine Learning." *WIREs Data Mining and Knowledge Discovery* 12 (3): e1452. https://doi.org/10.1002/widm.1452.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. "A Survey on Bias and Fairness in Machine Learning." *ACM Computing Surveys* 54 (6): 115:1115:35. https://doi.org/10.1145/3457607.

Mukerjee, Amitabha, Rita Biswas, Kalyanmoy Deb, and Amrit P. Mathur. 2002. "Multi–objective Evolutionary Algorithms for the Risk–return Trade–off in Bank Loan Management." *International Transactions in Operational Research* 9 (5): 583–97. https://doi.org/10.1111/1475-3995.00375.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366 (6464): 447–53. https://doi.org/10.1126/science.aax2342.

OPEN SCIENCE COLLABORATION. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251): aac4716. https://doi.org/10.1126/science.aac4716.

Ortiz-Ospina, Esteban, and Max Roser. 2017. "Healthcare Spending." *Our World in Data*, June. https://ourworldindata.org/financing-healthcare.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 28252830.

Perrone, Valerio, Michele Donini, Muhammad Bilal Zafar, Robin Schmucker, Krishnaram Kenthapadi, and Cédric Archambeau. 2021. "AIES '21: AAAI/ACM Conference on AI, Ethics, and Society." In, 854–63. Virtual Event USA: ACM. https://doi.org/10.1145/3461702.3462629.

Rodolfa, Kit T., Pedro Saleiro, and Rayid Ghani. 2020. "Bias and Fairness." In, 2nd ed. Chapman; Hall/CRC.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66. https://doi.org/10.1177/0956797611417632.

Simonsohn, Uri. 2018. "Two Lines: A Valid Alternative to the Invalid Testing of U-Shaped Relationships With Quadratic Regressions." *Advances in Methods and Practices in Psychological Science* 1 (4): 538–55. https://doi.org/10.1177/2515245918805755.

Sommers, Benjamin D., Atul A. Gawande, and Katherine Baicker. 2017. "Health Insurance Coverage and Health — What the Recent Evidence Tells Us." *New England Journal of Medicine* 377 (6): 586–93. https://doi.org/10.1056/NEJMsb1706645.

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. "Increasing Transparency Through a Multiverse Analysis." *Perspectives on Psychological Science* 11 (5): 702–12. https://doi.org/10.1177/1745691616658637.

Sukthanker, Rhea, Samuel Dooley, John P. Dickerson, Colin White, Frank Hutter, and Micah Goldblum. 2022. "On the Importance of Architectures and Hyperparameters for Fairness in Face Recognition." https://doi.org/10.48550/arXiv.2210.09943.

Team, Pipenv Maintainer. 2017. *Pipenv: Python Development Workflow for Humans.* https://github.com/pypa/pipenv.

Team, R Core. 2022. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

team, The pandas development. 2020. *Pandas-Dev/Pandas: Pandas.* Zenodo. https://doi.org/10.5281/zenodo.3509134.

Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual.* Scotts Valley, CA: CreateSpace.

Weerts, Hilde J. P. 2021. "An Introduction to Algorithmic Fairness." https://doi.org/10.48550/arXiv.2105.05595.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the Tidyverse." https://joss.theoj.org.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67 (2): 301–20. https://www.jstor.org/stable/3647580.