One Model Many Scores: Using Multiverse Analysis to Prevent Fairness Hacking and Evaluate the Influence of Model Design Decisions

Jan Simson
LMU Munich
Munich, Germany
Munich Center for Machine Learning
(MCML)
Munich, Germany
jan.simson@lmu.de

Florian Pfisterer
LMU Munich
Munich, Germany

Christoph Kern
LMU Munich
Munich, Germany
Munich Center for Machine Learning
(MCML)
Munich, Germany
University of Maryland
College Park, USA
christoph.kern@lmu.de

ABSTRACT

A vast number of systems across the world use algorithmic decision making (ADM) to (partially) automate decisions that have previously been made by humans. The downstream effects of ADM systems critically depend on the decisions made during a systems' design, implementation, and evaluation, as biases in data can be mitigated or reinforced along the modeling pipeline. Many of these decisions are made implicitly, without knowing exactly how they will influence the final 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 decisions during design and evaluation 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 investigate the variability and robustness of fairness scores and see how and which decisions impact fairness. We demonstrate how multiverse analyses can be used to better understand fairness implications of design and evaluation decisions using an exemplary case study of predicting public health care coverage for vulnerable populations. Our results highlight how decisions regarding the evaluation of a system can lead to vastly different fairness metrics for the same model. This is problematic, as a nefarious actor could optimise or "hack" a fairness metric to portray a discriminating model as fair merely by changing how it is evaluated. We illustrate how a multiverse analysis can help to address this issue.

CCS CONCEPTS

• **Social and professional topics** → **User characteristics**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

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

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 [43], the labor market for hiring decisions or filtering resumes [22], and the criminal justice system to assess risks of recidivism [5]. While these systems are promising when designed well, raising hopes of more accurate and objective decisions, their impact can be quite the opposite when designed incorrectly. There are many examples of ADM systems discriminating against people [41]. 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 [28]. Other examples include the Dutch childcare benefits system using an ADM system that was much more likely to accuse immigrants of having committed fraud [32].

These fairness problems often occur because algorithms replicate biases in the underlying training data. However, biases can also be amplified throughout the machine learning pipeline depending on how exactly data is processed and turned into outputs [36, 49]. Unfortunately, no silver bullet exists to prevent biases in the machine learning pipeline [2] and legislation usually provides little guidance. 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 [10]. Alarmingly, we demonstrate how the evaluation of the same model can be modified to achieve large variability in a fairness metric, potentially allowing the *hacking* of fairness metrics. Related issues regarding the hacking or washing of fairness metrics have recently been raised in fair ML research [3, 40]. As a result, preventing algorithms from introducing, reinforcing or hiding biases requires careful study and evaluation of the – often implicit – decisions made while designing and evaluating a machine learning system. To address 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 [56]. We adapt this methodology across domains to work in the context of machine learning with a focus on evaluating metrics of algorithmic fairness. We present two variations of this method demonstrating its usefulness: (1) as a guidance during the design of the model and preprocessing pipeline and (2) as an estimator of robustness of a fairness metric and to protect against fairness hacking.

In the following, we present a generalizable approach of using multiverse analysis to estimate the effect of decisions during the design and evaluation of a machine learning or ADM system on fairness outcomes. Using a case study of predicting public health coverage in US census data we demonstrate how design decisions can be better understood and fairness hacking can be addressed. 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. [56] in response to the reproducibility crisis affecting the field [15]. 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 [51], 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* [26], 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 [11].

Multiverse analysis focuses on the preprocessing 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 preprocessing 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 extend this methodology to also examine the influence of variation in evaluation and adapt it for the machine learning context with a special focus on using it to generate insights on metrics of algorithmic fairness.

In addition to multiverse analysis, a related type of analysis, called specification curve analysis [53] emerged in the social sciences literature. Its goal is to assess the strength of an effect of interest under the different modelling decisions contained in the complete grid of possible decision combinations. Results are aggregated in a specification curve, a graph displaying the distribution of the effect size or coefficient of interest, yielding a single curve that allows assessing the robustness of a measured association across modelling decisions. In contrast, our approach is not only interested in the robustness, but we aim to also identify decisions that impact the resulting fairness metrics for further investigation.

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 and evaluation 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 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 preprocessing of data, in line with the original approach of multiverse analyses [56]. 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 decision combinations, the multiverse. For each of these universes, we compute the resulting fairness and performance metrics of the machine learning system and collect them 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).

Another novelty in our approach is our introduction of two distinct perspectives on multiverse analyses: One with a focus on preprocessing, fostering the understanding of how decisions affect models in a fairness context and a second, focusing on robust
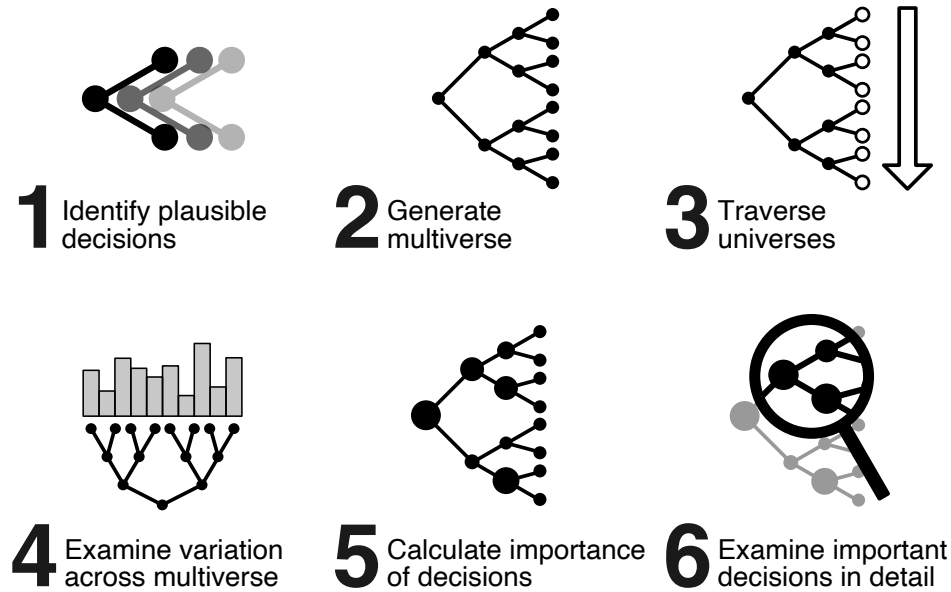
Figure 1: Steps to conduct a multiverse analysis for algorithmic fairness. Steps 1 - 4 apply to multiverse analyses in general, whereas steps 5 - 6 are unique to larger multiverse analyses for algorithmic fairness.

fairness evaluation of ML systems and protecting against cherry picking of evaluation criteria.

1.3 Related Research

Existing work has described the effects of specific preprocessing or modeling decisions in isolation, such as the influence of different imputation methods [14], of the model architecture, and of hyperparameters [20] on fairness in different contexts. Multiverse analyses have also been used to model the performance distribution in hyperparameter-space [8], but not yet to analyze algorithmic fairness. Research into model multiplicity has discovered multiple sources of arbitrariness that can influence model predictions and fairness: Random samples of a dataset can lead to different predictions on the individual level [17, 24], the selection of different target variables can strongly affect model fairness [61] and even the original sampling during the creation of a dataset can be considered arbitrary [42].

In terms of manipulating fairness, prior work has demonstrated the possibility of generating surrogate models that show little dependence on protected features for unfair models, a process termed "fairwashing" [3]. Under an assumption of "fairness through unawareness", these surrogate models could then be presented as fair models. This assumption is unrealistic in practice, however, as there are commonly proxy variables available for protected attributes [7]. Recent parallel work has demonstrated a process of using completely different fairness metrics to then report only the one with the most optimal score in a process also termed "fairness hacking" [40]. In this work, we demonstrate how there is no need to vary the chosen fairness metric itself, if one is willing to shift evaluation criteria in order to manipulate its scores. We believe both of these

approaches are troublesome and fall under the term "fairness hacking". They closely mirror practices of varying evaluation criteria to achieve significant p-values, a practice commonly referred to as "p-hacking", which gave rise to the introduction of multiverse analysis in psychology in the first place [52].

The field of hyperparameter-optimization (HPO) [9, 23] tries to optimize the process of tuning machine learning model hyperparameters. This field typically focuses on optimizing algorithm performance by employing efficient search strategies that allow optimizing performance without requiring the exploration of the complete hyperparameter space. However, adaptive search patterns such as, e.g. Bayesian Optimization [54], usually focus on efficiently finding the optimal configuration and yield non-i.i.d. optimization traces. This makes them unsuitable for assessing the influence and robustness of any particular decision as post-hoc analysis relies on representative, i.i.d. data. While algorithmic fairness is also explored in the context of HPO [47, 48], the focus is only on finding models with favourable performance-fairness trade-offs instead of understanding the effects of individual decisions or assessing overall robustness. Here, we draw on insights and methodology from the field of HPO, in particular the functional analysis of variance (FANOVA) [30, 31] 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.

1.4 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 in 2020 [45]. Whether or not someone is covered by health insurance can have large effects on their health and financial situation: According to Sommers et al. [55], people with insurance have better self-reported health, have more preventative doctor's appointments, improved depression outcomes, and fewer personal bankruptcies.

We implement our case study using the ACSPublicCoverage dataset [19], with data from the American Community Survey (ACS) Public Use Microdata Sample (PUMS) [13]. 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 [6, 19, 21]: UCI Adult [37], the most popular dataset in the fairness literature [21], uses an arbitrary threshold of $50,000 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 [19]. 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 [12]. In line with the original task setup by Ding et al. [19], 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 [35].

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* [1, 27] 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 [62] 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 [44] and matching the original task [19].

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 [38, 39] we identified five distinct categories to structure and guide the identification of decisions: Data Selection, Preprocessing, Modeling, Post-Hoc and Evaluation decisions (Table 1). 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.

We chose to examine evaluation decisions separately from preprocessing decisions to demonstrate the two main uses of a multiverse analysis for algorithmic fairness: Understanding fairness implications of design decisions during model development and studying robustness of fairness scores in model evaluations. We therefore split the list of decisions as well as the following analyses into *Study 1* examining the impact of design decisions on models and *Study 2* examining the variation that can arise from differences in evaluation decisions. An overview of all decisions and their respective options can be seen in Table 1, and a detailed description of each is provided below.

2.2.1 Study 1: Model Design Decisions. We consider 9 distinct and orthogonal design decisions. Each of these decisions has two to five unique choice options, leading to a total of $N = 61440$ combinations of decisions or universes. We consider decisions roughly in the order they would be made during a typical analysis.

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 [63]. 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 from modelling.

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, due to unreliability in the data

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.

Category	Decision	Options
		Decisions and Options Examined in Case Study
Decisions examined in Study 1		
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
	Stratify Split	(1) none; (2) target; (3) protected-attribute; (4) both
Post-Hoc	Cutoff	(1) raw-0.5; (2) quantile-0.1; (3) quantile-0.25
Decisions examined in Study 2		
Evaluation	Eval Fairness Grouping	(1) majority-minority; (2) separate
	Eval Exclude Subgroups	(1) exclude-in-eval; (2) keep-in-eval
	Eval On Subset	(1) full; (2) locality-largest-only; (3) locality-most-privileged; (4) locality-city-la; (5) locality-city-sf; (6) exclude-military; (7) exclude-non-citizens

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. Groups were *not* dropped from the test data used for evaluation as part of this decision. 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%).

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.

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.

Encoding of Categorical Variables (Encode Categorical). Another common preprocessing 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.

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 other choices, as HPO has already been studied elsewhere [47]. We therefore support the following model types as options for this decision: (1) logistic regression [18], (2) random forest [29], (3) gradient boosting machine [25], and (4) elastic net [65] trained with their default hyperparameters.

Stratification of Train-Test Split (Stratify Split). Training and test sets are often created by simple random splitting of the

---

[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.

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.

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*.

2.2.2 Study 2: Evaluation. We consider 3 distinct and orthogonal decisions, all focusing on evaluation only. Each decision has between 2 and 7 options each. Together these produce a total of $N = 28$ unique evaluation strategies for any given model, without modifying the model or its predictions.

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 default behaviour in Weerts et al. [62]). Naturally, this has a strong influence on the fairness metric. 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*).

Exclusion of Subgroups during Evaluation (Eval Exclude Subgroups). Similarly to how subgroups of the protected attribute may be excluded from the training data, they may also be excluded from the test data used for evaluation, with potentially even greater

adverse impact. We examine the exclusion of the same subgroups as in the decision *Exclude Subgroups* in Study 1 (Section 2.2.1) and vary whether or not subgroups are also excluded from the test dataset. The same warnings raised for that decision are even more relevant for this decision and we *strongly* discourage the exclusion of subgroups in any system.

Evaluation using a Subset of the Data (Eval on Subset). When assessing the fairness of a system, the evaluation may happen on only a subset of the eventual target population, for example because some populations may be easier to reach or because the model deployment context changes over time. While this practice is obviously not desirable, it may be necessary in certain situations due to real-world limitations in resources. An example of this is the popular COMPAS dataset [5] which was constructed using only data from a single county (Broward County, Florida), as a larger-scale construction of such a dataset would not have been feasible. We examine the following options for this decision, to represent possible population subsets one may use for evaluation: (1) examining only the largest geographical region (in terms of sample size), (2) examining the geographical region with the largest fraction of the privileged group; examining only data from the counties of (3) Los Angeles or (4) San Francisco, (5) examining a subset of only non-military people (as former military status may affect healthcare status), (6) examining only U.S. citizens and (7) not examining any subset, but rather using the full test data for evaluation.

2.3 Software

Analyses were conducted using Python Version 3.8 [60] and pipenv [57] for reproducibility. The Python package scikit-learn [46] was used for preprocessing and fitting of models, pandas [59] for loading and modification of data, folktables [19] for retrieval of data, fairlearn [62] for computation of fairness metrics, fANOVA [31] for calculation of variable importance and papermill [16] for parameterized computation of decision universes. This reproducible document was generated using quarto [4], R [58] Version 4.2, the R packages from the tidyverse [64] and ggpubr [34] for generation of figures. The source code of the analyses and this publication is available at https://github.com/reliable-ai/fairml-multiverse. We purposefully created source code in a modular fashion to allow for easy adoption of the multiverse method in other fair ML contexts. An interactive analysis of a subset of the results is available at https://reliable-ai.github.io/fairml-multiverse/.

3 RESULTS

3.1 Study 1: Model Design

The multiverse analysis examining the influence of model design decisions produced a total of $N = 61440$ values of the fairness metric in Study 1[3]. 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 F_1 scores between 0 and

---

[2]**Majority group**: 'White alone'; **Minority group(s)**: '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]In Study 1, we evaluated all models using the same strategy, namely not aggregating groups of the protected attribute, not excluding any subgroups during evaluation, and evaluating on the complete test set.
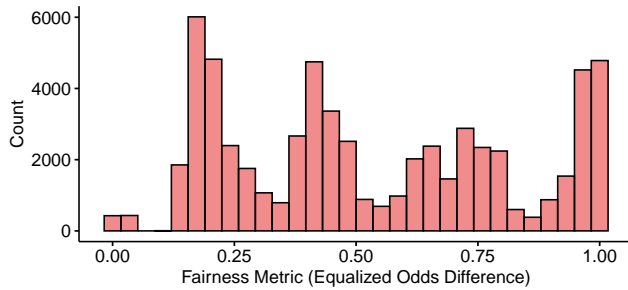
Figure 2: Variation in the multiverse spans the entirety of possible values of the fairness metric. Distribution of fairness metric (equalized odds difference) across universes. Lower values on the fairness metric indicate smaller *TPR* and *FPR* differences across groups.
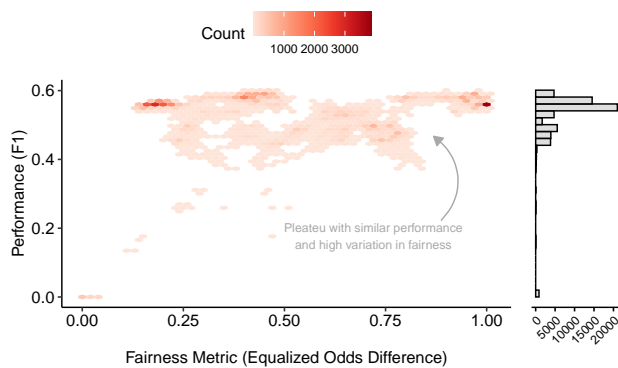

Figure 3: Performance and fairness are largely unrelated with plateaus of low variance in performance, but high variance in fairness. Distribution of overall performance as F_1 score and fairness metric (equalized odds difference) across all multiverses. Marginal histogram shows distribution of performance. A marginal histogram of the fairness metric can be seen in Figure 2, similar figures for raw and balanced accuracy can be seen in Figure A6. An interactive version of this figure is available.

0.598 and raw accuracies between 0.419 and 0.722. Performance and the fairness metric were only weakly correlated with a Pearson correlation of $r = 0.149$ for F_1 scores and $r = 0.192$ for raw accuracy. For the F_1 score, the majority of universes fell into a similar range of performance, but exhibited large variation on the fairness metric (Figure 3), highlighting the opportunity to optimize algorithmic fairness without sacrificing performance in line with Islam et al. [33]. Raw accuracy exhibited similar opportunities, varying largely based on the decision *Cutoff*, with three large clusters of similar performance (Figure A6 A). For balanced accuracy the distribution of fairness and performance values was slightly more complex, exhibiting a slight fairness-performance trade-off (Figure A6 B).

3.1.1 Importance of Decisions. We conducted a FANOVA [30] as described in Hutter et al. [31] 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.

As can be seen in Table 2, the most important decision is how the stratification of the train-test split is performed. Moreover, the interaction of the chosen cutoff value with the stratification strategy is highly important, accounting for more than 30% of the variance in the fairness metric. It also becomes apparent that especially the *interactions* of decisions are relevant here, with all decisions among the top 10 except the stratification and cutoff being interactions rather than sole decisions.

We analyzed the three most important decisions or decision-interactions 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. The decisions *Stratify Split*, *Cutoff* and their interaction account for all three of the most important decisions. When examining the decision separately, it can be seen how stratifying by the target variable leads to noticeably lower fairness scores (Figure 4 A, most important) and how the raw cutoff value of 0.5 is suddenly not leading to the best fairness scores anymore (Figure 4 B, 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 models with the best average fairness metric when using any other stratification strategy (Figure 4 C, second most important).

As variation in random train-test splits can affect fairness and performance of machine learning models [17, 24], we repeated the complete multiverse analysis five times with different random seeds, achieving highly similar results regarding both the overall variation of the fairness metric (Figure A7) and the relative importance of decisions (Figure A8).

3.1.2 Scaling the Analysis. Conducting a multiverse analysis can be computationally expensive. Especially if the multiverse is particularly large or computational resources are limited, it may not be possible to explore the complete grid of universes. 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

Table 2: The 10 most important decisions or decision interactions and their relative importance.

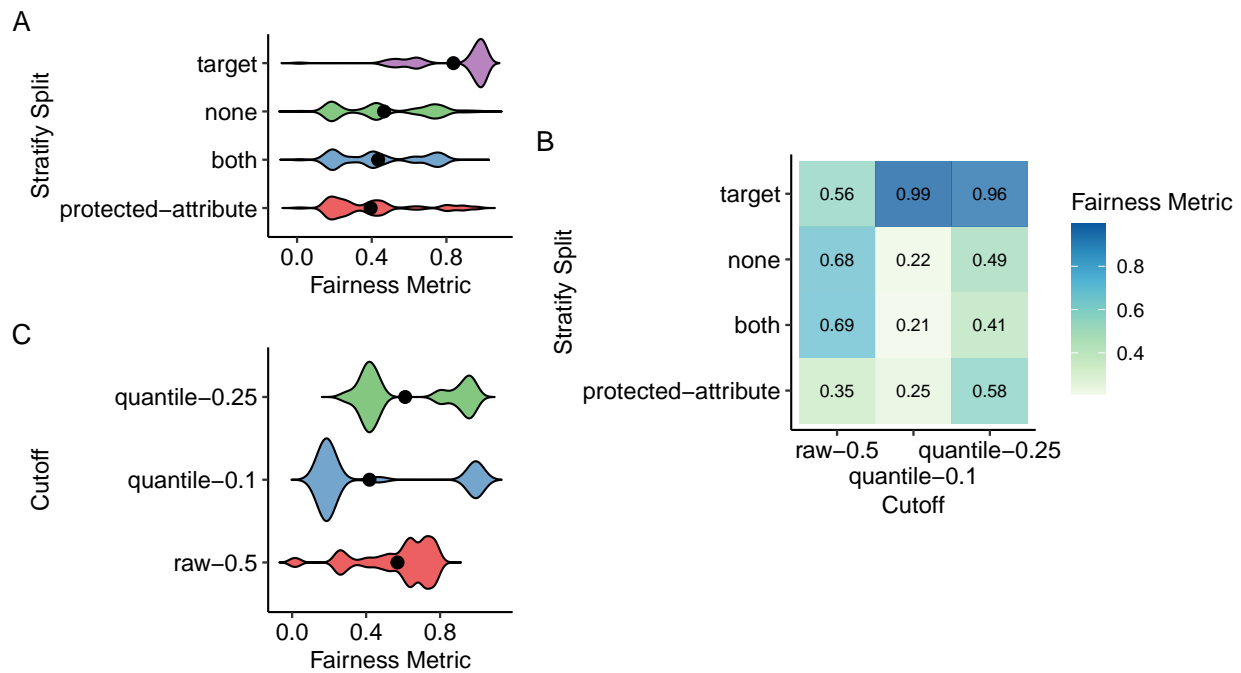| Effect Type	Decision / Interaction of Decisions	Importance	Std. Deviation
main	*StratifySplit*	0.375	0.001
2-way int.	*Cutoff × StratifySplit*	0.313	0.000
main	*Cutoff*	0.081	0.000
4-way int.	*Cutoff × ExcludeFeatures × Model × StratifySplit*	0.008	0.000
3-way int.	*Cutoff × Model × StratifySplit*	0.007	0.000
3-way int.	*Cutoff × Model × PreprocessIncome*	0.007	0.000
2-way int.	*Model × PreprocessIncome*	0.007	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

Figure 4: The influence of decisions on the fairness metric can only be understood when examining interactions on top of individual decisions. Visualization of the fairness metric depending on the three most important decision / decision combinations (from A - C by importance) and their respective options.

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.

When calculating the Pearson correlation, the resulting mean correlation coefficient ranged from $\bar{r}_{1\%} = 0.996 \, (SD = 0.003)$ at 1% to $\bar{r}_{20\%} \geq 0.999 \, (SD = 0)$ at 20%. Spearman rank-correlations were also high, but lower than the Pearson correlation coefficients and more inconsistent (Figure A9), which indicates that using sparse data to estimate the importance of decisions works well for important decisions and less-so to identify nuances between less-important decisions. The resulting Spearman rank-correlation mean coefficients ranged from $\bar{\rho}_{1\%} = 0.529 \, (SD = 0.031)$ at 1% to $\bar{\rho}_{20\%} = 0.937 \, (SD = 0.007)$ at 20%.

3.2 Study 2: Evaluation

By combining the different evaluation decisions we end up with $N = 28$ possible evaluation strategies for any given model. We computed each of these for each of the universes from Study 1. This lead to a total of $N = 1,720,320$ values of the fairness metric with a mean value of $M = 0.339$. Similar to Study 1, these fairness values exhibited a high degree of variation. However, variation stayed high, even when examining values for the *exact same model*. We observe a full spread of the fairness metric from 0 to 1 ($\Delta = 1$) for 5.80% of the models, only by varying their evaluation. Alarmingly, we observe a spread of at least $\Delta \geq 0.9$ on the fairness metric for 94.51% of models. In the following we examine variation due to evaluation decisions for a single model in more detail.

We examined the variation of two individual models in more detail to illustrate the impact of evaluation decisions on algorithmic fairness for a single model. We chose to illustrate our point with one model exhibiting a median degree of variance based on evaluation decisions and one exhibiting a high degree. Neither model resulted from a particularly extreme combination of options.[4]

The overall distribution of the fairness metric alongside a detailed breakdown by decisions can be seen in Figure 5 for the model with median variation and Figure A10 for the model with high variation. Under the evaluation strategy used in Study 1, the chosen model with high variance would be considered highly unfair with a metric of $m_{EqOdds} = 1.000$ and the model with median variance slightly fairer with $m_{EqOdds} = 0.638$. However, as can be seen in Figure 5, there exist ample opportunities to tweak the evaluation strategy to achieve significantly better scores on the fairness metric. Indeed, both models can achieve a perfect score of 0 on the fairness metric, only by varying how they are evaluated. Given that the models stay exactly the same, we consider this practice "fairness hacking".

An overview of how evaluation decisions affect the fairness metric across the complete multiverse can be seen in Figure A11, illustrating how e.g. the fairness grouping can consistently mask disparate treatment of minority groups.

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. We highlight the importance of making decisions during model design and evaluation 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 as well as how fairness metrics can be manipulated through evaluation strategies. We further show which decisions during model design affect fairness the most: Surprisingly, we see that the stratification strategy used for the train-test split has strong effects on the fairness metric. We also observe that the cutoff value

used for making final decisions is important, a decision often implemented post-hoc after model deployment without consideration of fairness.

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 a model is to design decisions and which decisions warrant closer examination. Relative scores of decision importance should always be interpreted in light of the overall degree of observed variation. 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. 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 nine design and three evaluation decisions, with many plausible alternative decisions which could have 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 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 important decisions are robustly detected even when exploring only 1% of the full *multiverse*.

There are varying degrees of conducting a multiverse analysis of algorithmic fairness, each providing unique value and requiring different amounts of computation: We believe there is already significant value in (1) merely thinking about (implicit) decisions taken during system design and the consideration of potential alternatives, (2) performing a multiverse analysis of a fixed model with different evaluation strategies as a computationally inexpensive option to provide more robust evaluations and combat fairness hacking, (3) conducting a partial multiverse analysis of a subset of the full multiverse (e.g. 1%) and (4) an analysis of the full multiverse as the most thorough option.

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 source code 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. We further advocate for the use of multiverse

---

[4]The options for the model with median variance are: Cutoff = raw-0.5, Encode Categorical = ordinal, Exclude Features = race, Exclude Subgroups = drop-smallest-2, Model = rf, Preprocess Age = quantiles-4, Preprocess Income = bins-10000, Scale = scale, Stratify Split = none. The options for the model with high variance are: Cutoff = quantile-0.1, Encode Categorical = one-hot, Exclude Features = race, Exclude Subgroups = drop-other, Model = rf, Preprocess Age = quantiles-4, Preprocess Income = none, Scale = scale, Stratify Split = none.
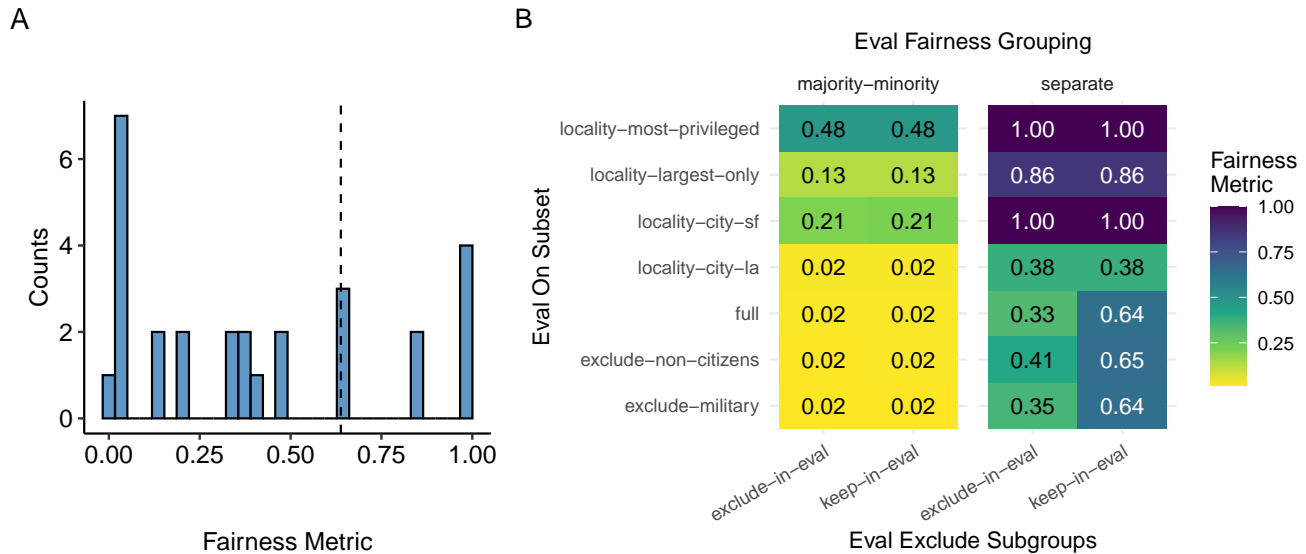
Figure 5: The fairness metric of the exact same model can be significantly altered by varying its evaluation strategy alone (A) and especially the interaction of different evaluation decisions leads to changes in the fairness metric (B). Overall distribution (A) and raw values (B) of fairness metric (equalized odds difference) for a single model over different decisions regarding its evaluation. The dashed line in A corresponds to the evaluation strategy used in Study 1[3]. Both plots display scores for a model showing median variation, to see the same figure for the model with high variation see Figure A10 in the Appendix. An interactive version of A is available, allowing examination of the distribution for any model in the multiverse analysis.

analysis in fairness evaluations to understand the distribution of fairness scores that can be evoked by the same model under different evaluation scenarios and to reduce the risk of potential fairness hacking by transparently reporting the entirety of results.

RESEARCH ETHICS AND SOCIAL IMPACT

Ethics Statement

Our selection of preprocessing and evaluation decisions builds on common practices observed in machine learning publications. While some of these practices such as excluding minority groups in preprocessing and evaluation are highly questionable and should not be normalized, we decided to include them in our case study to highlight their fairness implications and stimulate critical reflection. We further decided that criticism of individual manuscripts which implement such practices would not add much utility to our work, while potentially leading to (limited) negative consequences for their authors. Therefore, we present the implications of such data practices without singling out individual manuscripts.

Positionality Statement

All authors are affiliated with organizations from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) countries, in line with a common pattern in the fair ML research community [50]. This background inherently influenced the research practice of this study, including the case study and data that was chosen, which ultimately predetermined the design and evaluation decisions we focused on. We posit, however, that the proposed methodology

can be applied in a wide range of contexts, tasks, and with various different data modalities and protected attributes.

Adverse Impact Statement

We condemn potential misuses of our proposed method that contrast its objective of promoting transparency and reliability in machine learning practice and identified the following potential adverse impacts and misconceptions.

- We do not interpret fairness as an optimization problem. A multiverse analysis allows to understand the *variation of fairness scores* as a result of design decisions that researchers and developers might not have related to fairness in standard modeling practice and although fairness scores can imply real fairness they are only an indicator and not proof of fairness. While its results can inform discussions on sensible design decisions, the social impacts of an ADM system can only be understood by considering its specific implementation context and the interactions with the social environment in which it is placed.

- A multiverse analysis critically depends on the careful identification of *relevant design decisions*. While the decisions we examined in our case study may serve as a starting point, they do not present an exhaustive list by any means. Specifying a multiverse analysis requires researchers to carefully reflect on the data practices, processing and modeling decisions, embedded in their respective application context.

- A multiverse analysis should not be used to search for the evaluation strategy which displays the best fairness score. On the contrary, it presents a tool whose usage can be requested by stakeholders to instead *prevent selective reporting* and promote transparency by presenting the distribution of fairness scores across multiple evaluation schemes. It re-centers the discussion on how and for whom fairness metrics are computed, and acknowledges the susceptibility and instability of metrics to (small) changes in the evaluation protocol.

ACKNOWLEDGMENTS

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. (2018).
[2] Ashrya Agrawal, 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? (2021). https://doi.org/10.48550/arXiv.2011.02407
[3] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. 2019. Fairwashing: the risk of rationalization. In *International Conference on Machine Learning*. PMLR, 161–170.
[4] J.J. Allaire, Charles Teague, Carlos Scheidegger, Yihui Xie, and Christophe Dervieux. 2022. *Quarto*. https://doi.org/10.5281/zenodo.5960048 DOI: 10.5281/zenodo.5960048.
[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica* (05 2016), 254–264. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[6] Michelle Bao, 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. (2022). https://doi.org/10.48550/arXiv.2106.05498
[7] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. Classification - No Fairness through Unawareness. In *Fairness and Machine Learning: Limitations and Opportunities*. The MIT Press, Cambridge, Massachusetts.
[8] Samuel J. Bell, Onno P. Kampman, Jesse Dodge, and Neil D. Lawrence. 2022. Modeling the Machine Learning Multiverse. (2022). https://doi.org/10.48550/arXiv.2206.05985
[9] Bernd Bischl, Martin Binder, Michel Lang, Tobias Pielok, Jakob Richter, Stefan Coors, Janek Thomas, Theresa Ullmann, Marc Becker, Anne-Laure Boulesteix, Difan Deng, and Marius Lindauer. 2023. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Mining and Knowledge Discovery* 13, 2 (03 2023). https://doi.org/10.1002/widm.1484
[10] Emily Black, Manish Raghavan, and Solon Barocas. 2022. Model Multiplicity: Opportunities, Concerns, and Solutions. (2022).
[11] Nate Breznau, Eike Mark Rinke, Alexander Wuttke, Hung H. V. Nguyen, Muna Adem, Jule Adriaans, Amalia Alvarez-Benjumea, Henrik K. Andersen, Daniel Auer, Flavio Azevedo, Oke Bahnsen, Dave Balzer, Gerrit Bauer, Paul C. Bauer, Markus Baumann, Sharon Baute, Verena Benoit, Julian Bernauer, Carl Berning, Anna Berthold, Felix S. Bethke, Thomas Biegert, Katharina Blinzler, Johannes N. Blumenberg, Licia Bobzien, Andrea Bohman, Thijs Bol, Amie Bostic, Zuzanna Brzozowska, Katharina Burgdorf, Kaspar Burger, Kathrin B. Busch, Juan Carlos-Castillo, Nathan Chan, Pablo Christmann, Roxanne Connelly, Christian S. Czymara, Elena Damian, Alexander Ecker, Achim Edelmann, Maureen A. Eger, Simon Ellerbrock, Anna Forke, Andreas Forster, Chris Gaasendam, Konstantin Gavras, Vernon Gayle, Theresa Gessler, Timo Gnambs, Amélie Godefroidt, Max Grömping, Martin Groß, Stefan Gruber, Tobias Gummer, Andreas Hadjar, Jan Paul Heisig, Sebastian Hellmeier, Stefanie Heyne, Magdalena Hirsch, Mikael Hjerm, Oshrat Hochman, Andreas Hövermann, Sophia Hunger, Christian Hunkler, Nora Huth, Zsófia S. Ignácz, Laura Jacobs, Jannes Jacobsen, Bastian Jaeger, Sebastian Jungkunz, Nils Jungmann, Mathias Kauff, Manuel Kleinert, Julia Klinger, Jan-Philipp Kolb, Marta Kołczyńska, John Kuk, Katharina Kunißen, Dafina Kurti Sinatra, Alexander Langenkamp, Philipp M. Lersch, Lea-Maria Löbel, Philipp Lutscher, Matthias Mader, Joan E. Madia, Natalia Malancu, Luis Maldonado, Helge Marahrens, Nicole Martin, Paul Martinez, Jochen Mayerl, Oscar J. Mayorga, Patricia McManus, Kyle McWagner, Cecil Meeusen, Daniel Meierrieks, Jonathan Mellon, Friedolin Merhout, Samuel Merk, Daniel Meyer,

Leticia Micheli, Jonathan Mijs, Cristóbal Moya, Marcel Neunhoeffer, Daniel Nüst, Olav Nygård, Fabian Ochsenfeld, Gunnar Otte, Anna O. Pechenkina, Christopher Prosser, Louis Raes, Kevin Ralston, Miguel R. Ramos, Arne Roets, Jonathan Rogers, Guido Ropers, Robin Samuel, Gregor Sand, Ariela Schachter, Merlin Schaeffer, David Schieferdecker, Elmar Schlueter, Regine Schmidt, Katja M. Schmidt, Alexander Schmidt-Catran, Claudia Schmiedeberg, Jürgen Schneider, Martijn Schoonvelde, Julia Schulte-Cloos, Sandy Schumann, Reinhard Schunck, Jürgen Schupp, Julian Seuring, Henning Silber, Willem Sleegers, Nico Sonntag, Alexander Staudt, Nadia Steiber, Nils Steiner, Sebastian Sternberg, Dieter Stiers, Dragana Stojmenovska, Nora Storz, Erich Striessnig, Anne-Kathrin Stroppe, Janna Teltemann, Andrey Tibajev, Brian Tung, Giacomo Vagni, Jasper Van Assche, Meta van der Linden, Jolanda van der Noll, Arno Van Hootegem, Stefan Vogtenhuber, Bogdan Voicu, Fieke Wagemans, Nadja Wehl, Hannah Werner, Brenton M. Wiernik, Fabian Winter, Christof Wolf, Yuki Yamada, Nan Zhang, Conrad Ziller, Stefan Zins, and Tomasz Żółtak. 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 (11 2022), e2203150119. https://doi.org/10.1073/pnas.2203150119 Publisher: Proceedings of the National Academy of Sciences.
[12] US Census Bureau. 2021. ACS Health Insurance Coverage Recoding Programming Code. https://www.census.gov/topics/health/health-insurance/guidance/programming-code/acs-recoding.html Section: Government.
[13] US Census Bureau. 2021. Understanding and using the American Community Survey public use microdata sample files: What data users need to know.
[14] Simon Caton, Saiteja Malisetty, and Christian Haas. 2022. Impact of Imputation Strategies on Fairness in Machine Learning. *Journal of Artificial Intelligence Research* 74 (09 2022). https://doi.org/10.1613/jair.1.13197
[15] OPEN SCIENCE COLLABORATION. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (08 2015), aac4716. https://doi.org/10.1126/science.aac4716 Publisher: American Association for the Advancement of Science.
[16] nteract contributors. 2017. *papermill: Parametrize and run Jupyter and nteract Notebooks*. https://github.com/nteract/papermill
[17] A. Feder Cooper, Katherine Lee, Madiha Zahrah Choksi, Solon Barocas, Christopher De Sa, James Grimmelmann, Jon Kleinberg, Siddhartha Sen, and Baobao Zhang. 2024. Arbitrariness and Social Prediction: The Confounding Role of Variance in Fair Classification. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 20 (March 2024), 22004–22012. https://doi.org/10.1609/aaai.v38i20.30203
[18] David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)* 20, 2 (1958), 215–232. Publisher: Wiley Online Library.
[19] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. (2021), 13.
[20] Samuel Dooley, Rhea Sukthanker, John Dickerson, Colin White, Frank Hutter, and Micah Goldblum. 2024. Rethinking bias mitigation: Fairer architectures make for fairer face recognition. *Advances in Neural Information Processing Systems* 36 (2024).
[21] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery* (09 2022). https://doi.org/10.1007/s10618-022-00854-z
[22] Evanthia Faliagka, Kostas Ramantas, and Giannis Tzimas. 2012. Application of Machine Learning Algorithms to an online Recruitment System. (2012).
[23] Matthias Feurer and Frank Hutter. 2019. *Hyperparameter Optimization*. Springer International Publishing, Cham, 3–33. https://doi.org/10.1007/978-3-030-05318-5_1 DOI: 10.1007/978-3-030-05318-5_1.
[24] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*. 329–338.
[25] Jerome H. Friedman. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 29, 5 (10 2001), 1189–1232. https://doi.org/10.1214/aos/1013203451 Publisher: Institute of Mathematical Statistics.
[26] Andrew Gelman and Eric Loken. 2014. The Statistical Crisis in Science. *American Scientist* 102, 6 (2014), 460. Publisher: Sigma XI-The Scientific Research Society.
[27] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. (2016).
[28] Luke Henriques-Gomes. 2023. Robodebt: five years of lies, mistakes and failures that caused a $1.8bn scandal. *The Guardian* (03 2023). https://www.theguardian.com/australia-news/2023/mar/11/robodebt-five-years-of-lies-mistakes-and-failures-that-caused-a-18bn-scandal
[29] Tin Kam Ho. 1995. Random decision forests, Vol. 1. IEEE, 278–282.
[30] Giles Hooker. 2007. Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables. *Journal of Computational and Graphical Statistics* 16, 3 (2007), 709–732. https://www.jstor.org/stable/27594267
[31] Frank Hutter, Holger Hoos, and Kevin Leyton-Brown. 2014. International Conference on Machine Learning. PMLR, 754–762. https://proceedings.mlr.press/

v32/hutter14.html ISSN: 1938-7228.

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

[33] Rashidul Islam, Shimei Pan, and James R. Foulds. 2021. AIES '21: AAAI/ACM Conference on AI, Ethics, and Society. ACM, Virtual Event USA, 586–596. https://doi.org/10.1145/3461702.3462614

[34] Alboukadel Kassambara. 2023. *ggpubr: 'ggplot2' Based Publication Ready Plots*. https://CRAN.R-project.org/package=ggpubr

[35] Katherine Keisler-Starkey and Lisa N Bunch. 2022. *Health Insurance Coverage in the United States: 2021 - Appendix Table C3*. Technical Report. https://www.census.gov/content/dam/Census/library/publications/2022/demo/p60-278.pdf

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

[37] Ronny Kohavi and Barry Becker. 1996. Adult data set. *UCI machine learning repository* 5 (1996), 2093.

[38] Max Kuhn and Kjell Johnson. 2020. *Feature engineering and selection: a practical approach for predictive models*. CRC Press, Taylor & Francis Group, Boca Raton London New York. www.feat.engineering

[39] Tai Le Quy, 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 (2022), e1452. https://doi.org/10.1002/widm.1452 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1452.

[40] Kristof Meding and Thilo Hagendorff. 2024. Fairness Hacking: The Malicious Practice of Shrouding Unfairness in Algorithms. *Philosophy & Technology* 37, 1 (Jan. 2024), 4. https://doi.org/10.1007/s13347-023-00679-8

[41] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6 (07 2021), 115:1–115:35. https://doi.org/10.1145/3457607

[42] Anna P. Meyer, Aws Albarghouthi, and Loris D'Antoni. 2023. The Dataset Multiplicity Problem: How Unreliable Data Impacts Predictions. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 193–204. https://doi.org/10.1145/3593013.3593988

[43] Amitabha Mukerjee, 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 (2002), 583–597. https://doi.org/10.1111/1475-3995.00375 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1475-3995.00375.

[44] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (10 2019), 447–453. https://doi.org/10.1126/science.aax2342 Publisher: American Association for the Advancement of Science.

[45] Esteban Ortiz-Ospina and Max Roser. 2017. Healthcare Spending. *Our World in Data* (06 2017). https://ourworldindata.org/financing-healthcare

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[47] Valerio Perrone, Michele Donini, Muhammad Bilal Zafar, Robin Schmucker, Krishnaram Kenthapadi, and Cédric Archambeau. 2021. AIES '21: AAAI/ACM Conference on AI, Ethics, and Society. ACM, Virtual Event USA, 854–863. https://doi.org/10.1145/3461702.3462629

[48] F. Pfisterer, S. Coors, J. Thomas, and B. Bischl. 2019. Multi-Objective Automatic Machine Learning with AutoxgboostMC. *arXiv* 1908.10796 [stat.ML] (2019).

[49] Kit T. Rodolfa, Pedro Saleiro, and Rayid Ghani. 2020. *Bias and Fairness* (2 ed.). Chapman and Hall/CRC. Num Pages: 32.

[50] Ali Akbar Septiandri, Marios Constantinides, Mohammad Tahaei, and Daniele Quercia. 2023. WEIRD FAccTs: How Western, Educated, Industrialized, Rich, and Democratic is FAccT?. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023*. ACM, 160–171. https://doi.org/10.1145/3593013.3593985

[51] Joseph P. Simmons, 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 (11 2011), 1359–1366. https://doi.org/10.1177/0956797611417632

[52] Uri Simonsohn, Leif D. Nelson, and Joseph P. Simmons. 2014. P-Curve: A Key to the File-Drawer. *Journal of Experimental Psychology: General* 143, 2 (2014), 534–547. https://doi.org/10.1037/a0033242

[53] Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. 2020. Specification Curve Analysis. *Nature Human Behaviour* 4, 11 (Nov. 2020), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

[54] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems* 25 (2012).

[55] Benjamin D. Sommers, 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 (08 2017), 586–593. https://doi.org/10.1056/NEJMsb1706645

[56] Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science* 11, 5 (09 2016), 702–712. https://doi.org/10.1177/1745691616658637 Publisher: SAGE Publications Inc.

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

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

[59] The pandas development team. 2020. *pandas-dev/pandas: Pandas*. Zenodo. https://doi.org/10.5281/zenodo.3509134 DOI: 10.5281/zenodo.3509134.

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

[61] Jamelle Watson-Daniels, Solon Barocas, Jake M. Hofman, and Alexandra Chouldechova. 2023. Multi-Target Multiplicity: Flexibility and Fairness in Target Specification under Resource Constraints. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 297–311. https://doi.org/10.1145/3593013.3593998

[62] Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. 2023. Fairlearn: Assessing and Improving Fairness of AI Systems. *Journal of Machine Learning Research* 24, 257 (2023), 1–8. http://jmlr.org/papers/v24/23-0389.html

[63] Hilde J. P. Weerts. 2021. An Introduction to Algorithmic Fairness. (2021). https://doi.org/10.48550/arXiv.2105.05595

[64] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Pedersen, Evan Miller, Stephan Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. Welcome to the Tidyverse. https://joss.theoj.org DOI: 10.21105/joss.01686.

[65] Hui Zou and Trevor Hastie. 2005. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67, 2 (2005), 301–320. https://www.jstor.org/stable/3647580 Publisher: [Royal Statistical Society, Wiley].
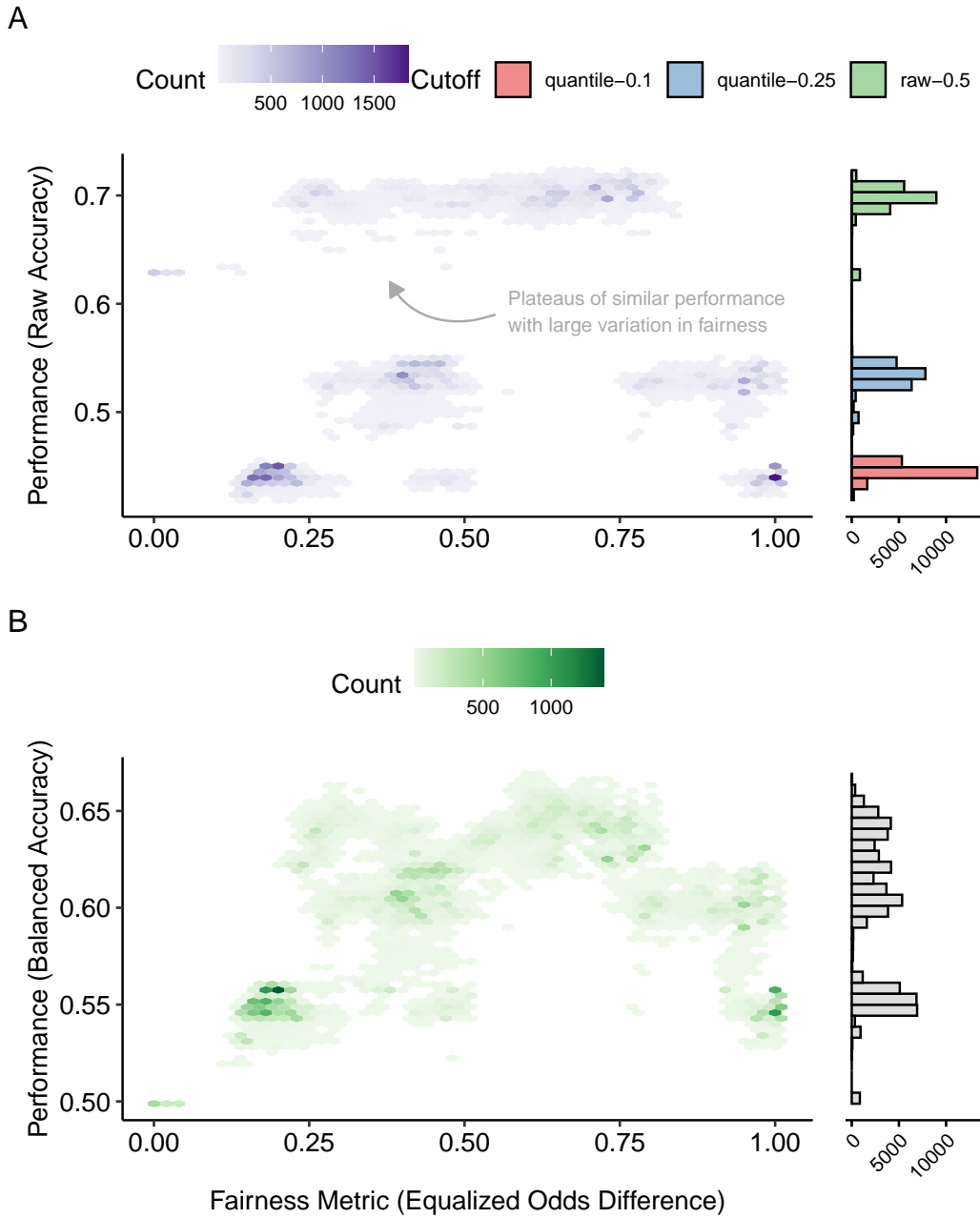
A SUPPLEMENTARY FIGURES

Figure A6: Performance and fairness are largely unrelated with clusters of low variance in performance, but high variance in fairness. Distribution of overall performance as raw (A) or balanced (B) accuracy and fairness metric (equalized odds difference) across all multiverses. Marginal histograms show distribution of performance for different options of the *Cutoff* decision in A and overall in B. A marginal histogram of the fairness metric can be seen in Figure 2. This figure is analogous to Figure 3 in the main text.
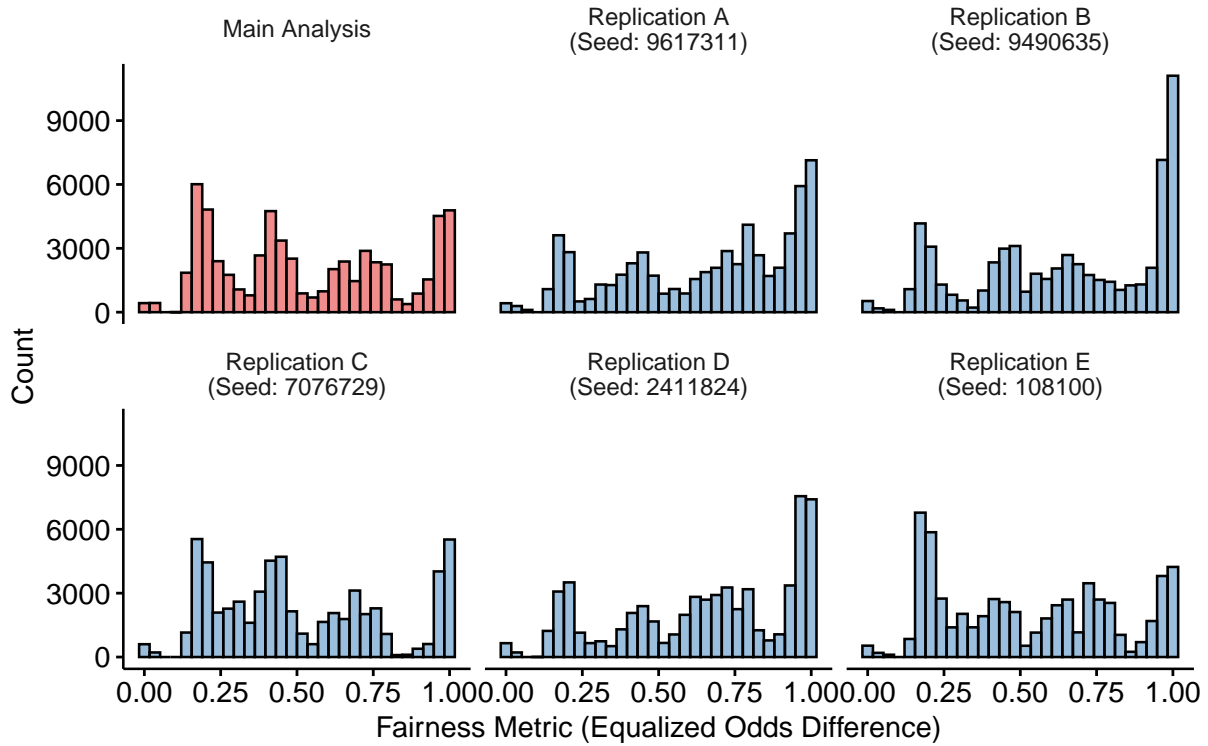
Figure A7: Overall variation in the multiverse is highly similar across different replications. Distribution of fairness metric (equalized odds difference) across universes in five different replications alongside the results reported in the main body of the paper. Lower values on the fairness metric indicate smaller *TPR* and *FPR* differences across groups.
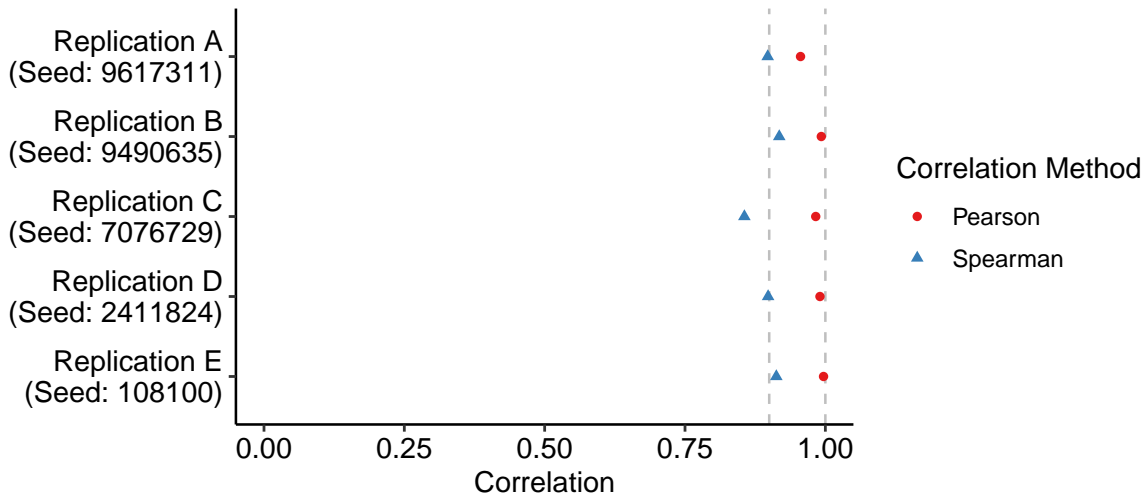

Figure A8: Estimates of decision importance are similar across replications of the analysis. Correlations of variance decomposition / importance estimates between the analysis reported in the main body of the paper and five replications. Pearson correlation coefficients are consistently higher than Spearman correlation coefficients, indicating better estimation of high-importance decisions. Dashed lines were inserted at 0.9 and 1.0 to indicate high correlation values.
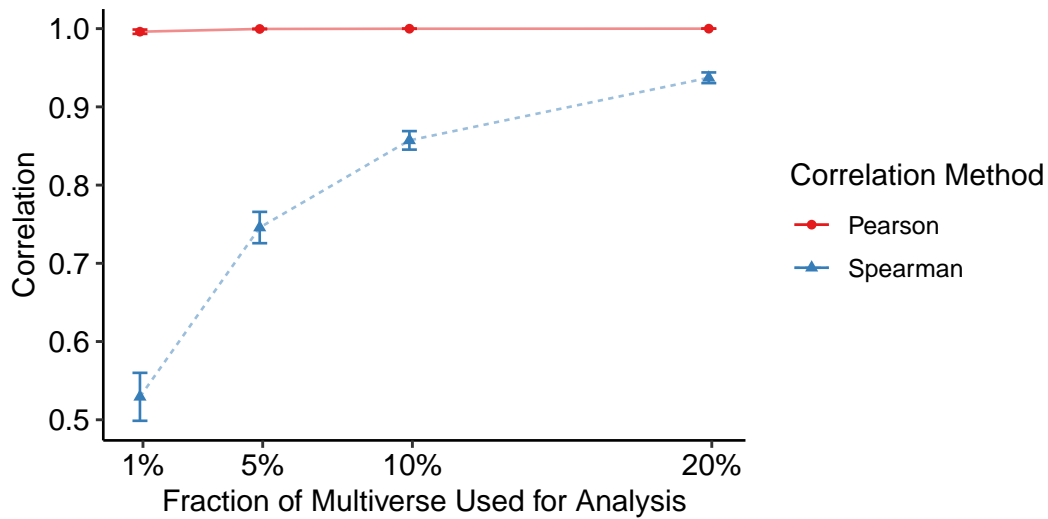
Figure A9: Conducting the analysis with smaller subsets of the complete multiverse leads to similar results. Correlations of variance decomposition / importance estimates between full dataset and random subsets of different sizes. Random subsets were drawn 50 times with points corresponding to mean correlations and lines to +/- 1 standard deviation. Pearson correlation coefficients are consistently higher than Spearman correlation coefficients.
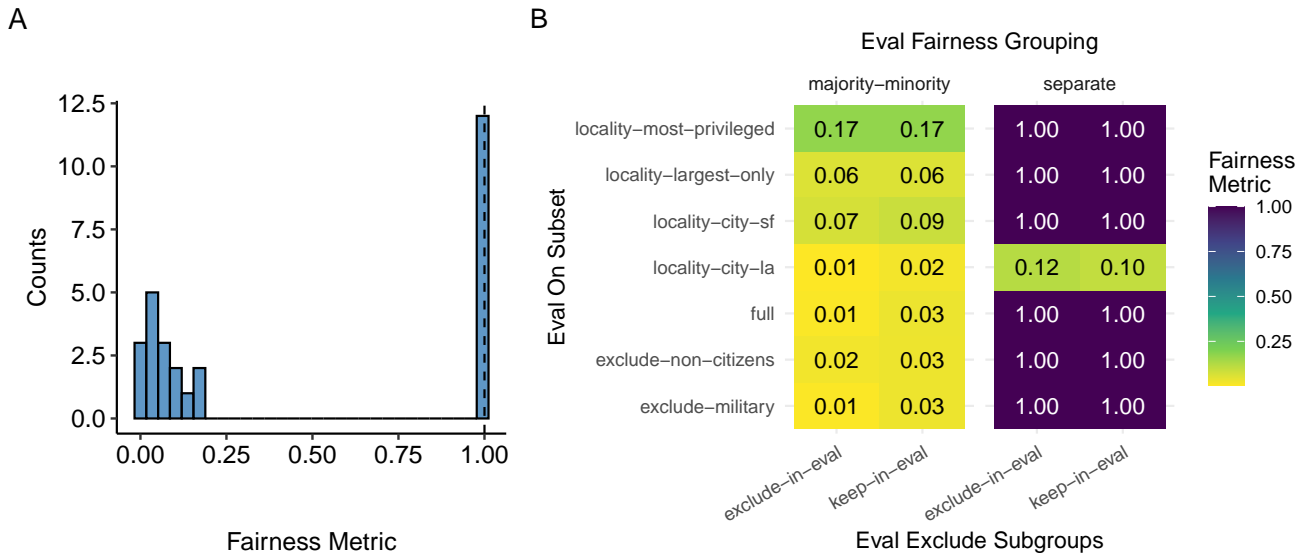

Figure A10: Evaluation decisions can strongly interact in their effect on the fairness metric. Overall distribution (A) and raw values (B) of the fairness metric for a single model exhibiting high variation over different decisions regarding its evaluation. This figure is analogous to Figure 5 in the main text.
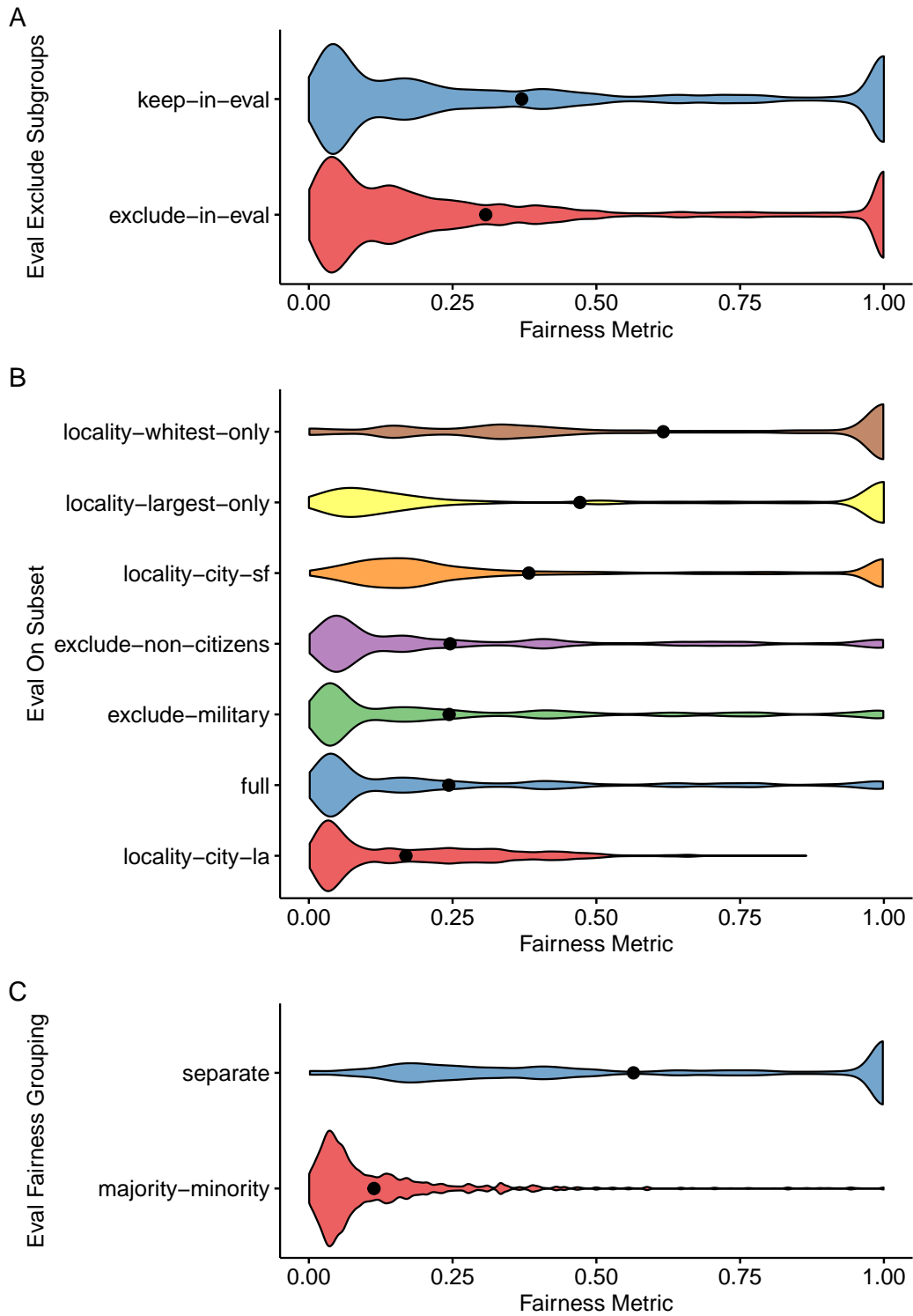
Figure A11: Despite strong interactions for the same model, evaluation decisions exhibit general tendencies in how they affect algorithmic fairness. Distribution of the fairness metric for different evaluation decisions across the complete multiverse of design decisions from studies 1 and 2.