Jan Simson

Re-Evaluating the Machine Learning Pipeline to Improve Fairness and Reliability

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht am 3.11.2025

Jan Simson

Re-Evaluating the Machine Learning Pipeline to Improve Fairness and Reliability

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht am 3.11.2025

Erster Berichterstatter: Prof. Dr. Christoph Kern
Zweiter Berichterstatter: Prof. Dr. Frauke Kreuter
Dritter Berichterstatter: Prof. Dr. Roger Peng

Tag der Disputation: 20.01.2026

Acknowledgments

A *huge thank you* to all the amazing people who have made the last three years not only possible, but actually incredibly fun. It was a real privilege getting to know you all, and every single one of you made a positive difference along this journey.

- First and foremost, I want to thank my advisors, *Christoph* and *Frauke*, for their continuous support, trust, and guidance. I especially appreciate the flexibility and freedom, alongside gentle nudges, that you provided, allowing my research to be shaped by my interests while staying on track. Thank you for the opportunity to do a PhD under your supervision!
- A big thank you to *Roger Peng* for acting as the third reviewer on this thesis, as well as *Joseph Sakshaug* and *Helmut Küchenhoff* for forming the examination committee for the disputation together.
- Thank you to the wonderful collaborators I got to work with over the past three years: *Alessandro*, *Sam*, *Fiona*, *Florian*, *Cosima*, *Malte*, and *Olga*. I appreciate the constructive discussions, valuable contributions, and great company you have all provided for these works. A special thank you to *Sam* for his advice and guidance in navigating the academic landscape. I hope and look forward to future collaborations with all of you!
- Thank you to *Patrick*, *Yegi*, and *Cynthia* for reading earlier drafts of this thesis and the valuable feedback they provided to improve it.
- Thank you to the past and present members of SODA. Thank you to *Ailin*, *Andreas*, *Anna*, *Bolei*, *Caro*, *Clara*, *Felicitas*, *Felix*, *Fiona*, *Julia*, *Leo*, *Lisa*, *Luis*, *Malte*, *Marcus*, *Markus*, *Olga*, *Patrick*, *Ruben*, *Sarah*, *Sofia*, *Unai*, and *Wiebke*. I appreciate the mensa visits, coffee chats, Pfälzer celebrations, board game evenings, and especially the trips and visits together with many of you. Special thanks to *Jacob* for the warm welcome and rising to YouTube fame together and to *Leah*, without whom I would've never known about this opportunity. I am grateful to get to call so many of you friends rather than colleagues.
- Thank you to *Pia*, *Yvonne*, *Christina*, *Brigitte*, and *Denise* for their continued help with the hardest part of research. You make it almost seem easy to navigate German university bureaucracy, and I enjoy every opportunity for a small chat on the 1st floor.
- I would also like to thank all the other lovely people at the Institute of Statistics here at LMU. Everyone has been incredibly welcoming, and it has been a joy to get to know and befriend many of you. It's truly an extraordinary institute. A special thank you goes out to *Yegi*, who has been the best of friends from the very beginning and to whom I owe more things than can be listed here – your place in Munich is empty. Special thanks also to *Alex* and *Martje* for more than a card deck's worth of fond memories and *Yin* for being extra welcoming from day one.
- Thank you also to the newfound friends outside of the institute, who have played a significant role in making Munich feel like home. Thank you for the movie nights, Padel sessions, concerts, and, most importantly, of course, all the hikes.
- Last, I'd like to thank my parents and brother for being the solid foundation that has always made it easy to explore.

Abstract

Across the world, an increasing number of decisions are, at least partially, automated using machine learning (ML) or artificial intelligence (AI) models, in what is typically referred to as algorithmic decision-making (ADM). The introduction of ADM systems has brought about significant transformations across a wide range of domains. However, these developments have not come without challenges: serious concerns have been raised about the ethical implications of automating (high-stakes) decisions, and numerous real-world examples illustrate how such systems can facilitate harm. Issues such as bias, discrimination, and a lack of transparency or accountability can often lead to outcomes that are unfair.

Building an ML, AI, or ADM system is typically a complex process involving multiple steps or stages that collectively form a pipeline. The stages along this pipeline include, but are not limited to, creation or choice of data, (pre-)processing of data, selection of a model type and architecture, model training, evaluation, and ultimately deployment and potentially decision-making. Each of the steps along this pipeline comes with a multitude of decisions. However, these decisions are often made ad hoc and might not even be recognized as such when they are made or as having alternative options. As decisions are often independent of each other, they together form a garden of forking paths, with the number of combinations of choices between decisions growing exponentially. Each pathway through this garden of forking paths can be seen as a plausible universe of choices. The resulting set of universes is often referred to as a multiverse and systematically explored in a "multiverse analysis". Navigating this multiverse is important as decisions affect not only real-world systems but also the development and evaluation of new methodologies.

This thesis offers a critical re-evaluation of the traditional machine learning pipeline. In this work, I systematically examine the different stages across the machine learning pipeline and the decisions made within each, questioning common defaults in search of better alternatives.

The thesis begins by describing problematic data practices we observed when reviewing research on fairness in machine learning. It continues by proposing multiverse analyses as a methodology for machine learning and, in particular, algorithmic fairness to evaluate robustness and mitigate potential issues of fairness hacking. Next, it shows how participatory input can be used to address problematic data practices, followed by the introduction of a new corpus and analysis of datasets in fair classification research. The thesis concludes with introducing a software implementation that explores the integration of machine learning-based suggestions within an interactive survey context.

Zusammenfassung

Weltweit werden immer mehr Entscheidungen mithilfe von Machine Learning (ML) oder künstlicher Intelligenz (KI; engl.: artificial intelligence, AI) teilweise oder vollständig automatisiert. Dies wird üblicherweise als algorithmische Entscheidungsfindung (engl.: algorithmic decision-making, ADM) bezeichnet. Die Einführung von ADM-Systemen hat in einer Vielzahl von Bereichen bereits zu starken Veränderungen geführt, allerdings nicht ohne Herausforderungen: Die Automatisierung (potenziell folgenschwerer) Entscheidungen wirft ernsthafte Bedenken hinsichtlich ethischer Implikationen auf. Zusätzlich veranschaulichen zahlreiche Beispiele aus der Praxis, wie Systeme dieser Art Schaden anrichten können: Ungleiche Entscheidungstendenzen, Verantwortungsdiffusion, Diskriminierung sowie ein Mangel an Transparenz können oft zu unfairen Ergebnissen führen.

Ein ML-, KI- oder ADM-Systems zu designen ist in der Regel ein komplexer Prozess und umfasst mehrere Schritte, die zusammen eine Pipeline bilden. Zu den Phasen dieser Pipeline gehören unter anderem das Erstellen oder die Auswahl von Datensätzen, die (Vor-)Verarbeitung von Daten, die Auswahl eines Modelltyps / einer Architektur, das Trainieren eines Modells, die Evaluation des Modells und schließlich die Inbetriebnahme und automatisierte Entscheidungsfindung. Jeder der Schritte entlang dieser Pipeline ist mit einer Vielzahl von Entscheidungen verbunden. Diese Entscheidungen werden jedoch häufig ad hoc getroffen und möglicherweise nicht immer als solche erkannt. Selbst wenn eine Entscheidung bewusst getroffen wird, besteht nicht immer Bewusstsein über mögliche Alternativen. Da die Entscheidungen oft unabhängig voneinander sind, bilden sie zusammen einen Garten verzweigter Pfade (engl.: garden of forking paths), wobei die Anzahl der möglichen Kombinationen zwischen Entscheidungen mit jeder zusätzlichen Entscheidung exponentiell wächst. Jeder Pfad durch diesen Garten kann als plausibles Universum an Entscheidungen betrachtet werden. Die daraus resultierenden Universen werden oft als Multiverse bezeichnet und in einer Multiverse-Analyse (engl.: multiverse analysis) systematisch untersucht. Die Navigation durch dieses Multiverse ist wichtig, da Entscheidungen nicht nur aktiv genutzte Systeme beeinflussen können, sondern auch die Entwicklung neuer Methoden und Algorithmen.

In dieser Dissertation stelle ich eine kritische Re-Evaluation der traditionellen ML-Pipeline vor. Ich untersuche systematisch die verschiedenen Schritte entlang der ML-Pipeline und hinterfrage gängige Entscheidungen in jedem Schritt auf der Suche nach potenziell besseren Alternativen.

Die Dissertation beginnt mit einer Beschreibung problematischer Praktiken beim Umgang mit Daten in der Forschung zu Fairness in ML und AI. Anschließend werden Multiverse-Analysen als Methodik für ML adaptiert. Als Nächstes wird demonstriert, wie Multiverse-Analysen für Untersuchungen zu Fairness in ML genutzt werden können, um u.a. potenziellen “Fairness-Hacking”-Problemen vorzubeugen. Danach wird zunächst gezeigt, wie partizipatives Design (engl.: participatory design) dabei helfen kann, einen Teil der zuvor genannten, problematischen Praktiken im Umgang mit Daten zu verhindern. Daraufhin wird ein neuer Korpus aus Datensätzen aus der Forschung zu fairen Klassifizierungsalgorithmen vorgestellt, inklusive einer Analyse dieser Datensätze. Die Dissertation schließt mit mehreren Softwareprojekten, unter anderem einer Applikation, welche die interaktive Nutzung von ML-generierten Vorschlägen in Umfragen untersucht.

Contents

Part I.

Introduction

1. Introduction

Artificial intelligence (AI) is advancing and being adopted at an unprecedented rate. While a lot of attention is focused on large language models, machine learning (ML) systems are being deployed across various domains, introducing more and more touchpoints within our lives every day. In this light, we observe an increasing number of decision-making processes that are either informed by or entirely delegated to algorithms and AI/ML systems – typically referred to as *algorithmic* or *automated decision-making* (ADM). Examples of ADM are abundant and span a multitude of domains, including the filtering of resumes and hiring decisions in the labor market (Faliagka et al., 2012; Fabris et al., 2025; Kappen and Naber, 2021), approval or denial of medical interventions by insurers (Mello and Rose, 2024; Atherton, 2023; United States District Court, District of Minnesota, 2023), identification of patients in need of additional care (Obermeyer et al., 2019), loan approval in finance (Mukerjee et al., 2002), and recidivism risk assessment in the criminal justice system (Angwin et al., 2016; Bao et al., 2022). ADM systems offer compelling promises of making decisions that are more efficient, fair, and just (Kappen and Naber, 2021; Christian Sandvig, 2014). When not designed well, however, ADM systems can amplify existing biases and lend a false sense of objectivity to opaque decision-making processes (Burrell, 2016; Obermeyer et al., 2019; Angwin et al., 2016). Although ADM systems are already deployed in many domains, we are still refining our understanding of them and how to build these systems well.

Unfortunately, there are many examples of the harm that occurs when systems are *not* designed well. Prominent examples include accusations against United Healthcare, a U.S. health insurance company, of using a faulty algorithm with an error rate as high as 90% to deny health care coverage against medical advice (Mello and Rose, 2024; Atherton, 2023; United States District Court, District of Minnesota, 2023), and the Dutch childcare benefits scandal (Persoonsgegevens and Belastingdienst, 2021; Amnesty International, 2021) where a discriminatory algorithm wrongly accused thousands of families of committing fraud all the while they were being denied the right to appeal (Henley, 2021). A database tracking harmful incidents related to AI has logged 1,230 incidents to date (McGregor, 2025).

When it comes to preventing such issues in the future, the majority of the work focuses on ML models *in isolation* and how to improve them through various means of *processing* (Barocas et al., 2023; Pessach and Shmueli, 2022). However, in this thesis, I argue – and demonstrate – that we need a broadening of perspective and consider not just models, but the ML pipeline in its entirety. How and which data are used and processed plays a crucial role in how it affects ML models (Simson et al., 2024a, 2025b; Caton et al., 2022), as do decision thresholds and evaluation protocols (Simson et al., 2024b; Meding and Hagendorff, 2024). An ML pipeline can both amplify and dampen existing biases in data, and will require careful study and evaluation to consciously choose between the two.

To substantiate this argument, this thesis is structured as a collection of studies, each investigating relevant stages of the ML pipeline. I critically examine each step in the machine learning pipeline,

re-evaluating default approaches in search of better alternatives, and assessing the importance of decisions along this pipeline. The thesis begins with Part I, including this introduction (Chapter 1) and a background chapter (Chapter 2) that provides an in-depth overview of the relevant research areas. Then, I provide an overview of the different publications that comprise this thesis, briefly summarizing each and positioning them in relation to each other (Chapter 3), before including the actual contributions themselves in Part II. The different publications are each making unique contributions along different sections of the ML pipeline. I don't just raise issues along the ML pipeline in these contributions, but also try to provide solutions for them.

The contributions of this thesis begin at the start of the pipeline, with *data*. The paper, "Lazy Data Practices Harm Fairness Research" (Chapter 4), provides a detailed and critical examination of datasets and common practices of how they are handled in fairness research. It identifies three particularly concerning data practices observed in the literature: *neglected identities* who are not represented in datasets, *omitted populations* who are removed from datasets during preprocessing, and *opaque preprocessing* where there is a lack of documentation regarding the (pre-)processing of datasets in the literature. This lack of documentation and transparency is problematic, as we also find that publications which do provide preprocessing details show considerable variation in their approaches. Using an illustrative case study, we demonstrate that these differences in data handling have a significant impact on fairness-related metrics. While the paper focuses on these three practices in particular, it also acknowledges other issues, such as the continuing widespread usage of datasets with known flaws.

Moving further along the ML pipeline, the second paper, "One Model Many Scores: Using Multiverse Analysis to Prevent Fairness Hacking and Evaluate the Influence of Model Design Decisions" (Chapter 5), focuses on design and evaluation of ML models. Using a case study, it examines the influence of design and evaluation decisions on algorithmic fairness. The paper examines decisions related to pipeline design, including the omission of populations described in the prior paper, and four key evaluation decisions. The paper highlights how seemingly unimportant decisions, which are rarely examined systematically, as well as their interactions, can have strong downstream effects on fairness. It also demonstrates how a multiverse analysis across evaluation decisions can be useful in increasing the transparency of reporting algorithmic fairness scores, which can otherwise be easily manipulated in a practice referred to as *fairness hacking*.

Addressing the complex decision-space identified in the previous paper, the third paper, "Preventing Harmful Data Practices by using Participatory Input to Navigate the Machine Learning Multiverse" (Chapter 6), explores the opportunity of navigating and restricting the decision space in the machine learning multiverse through participatory input. We demonstrate that laypeople can provide meaningful input on the design of a machine learning pipeline, with their input generally leading to fair and high-performing ML models. Participatory input was particularly effective at preventing the practice of omitting populations. The paper presents a promising new avenue for empowering stakeholders and people affected by machine learning systems to have a say in the creation of these systems.

Returning to the foundational challenges of handling data, the fourth paper in the thesis, "Bias Begins with Data: The *FairGround* Corpus for Robust and Reproducible Research on Algorithmic Fairness" (Chapter 7), has been developed specifically with the goal of addressing some of the problematic data practices observed in the literature and described in Chapter 4. By manually collecting and annotating a corpus of 44 different datasets, this work aims to lower the barrier to utilizing diverse pools of datasets in research. It aims to improve transparency and reproducibility

of data processing by providing sensible defaults for dataset processing as well as the option to – explicitly – deviate from defaults.

Finally, the fifth and last contributing article, "occupationMeasurement: A Comprehensive Toolbox for Interactive Occupation Coding in Surveys" (Chapter 8), returns to the very beginning of the pipeline and explores novel approaches of data collection and encoding. While primarily a software contribution, the paper introduces a new method for coding occupational data into existing classification schemes. The interactive application in the R package enables respondents in surveys to self-identify their occupational category, providing notably higher agency than current practices in which expert labellers assign categories to respondents post-hoc based on free-text responses. Due to the high complexity of occupation classification schemes, the tool utilizes a pre-trained ML model to suggest potential occupational categories based on free-text input.

In addition to academic articles, I have worked on several software projects while conducting the research that forms the basis of this thesis. These software projects have often been developed in conjunction with or as follow-up work to academic publications. In many cases, the software projects have also played a critical role in enabling the publications presented in this thesis. While some of the software projects are released alongside academic publications, e.g., the `occupationMeasurement` R Package (Simson et al., 2023) and the `fairml-datasets` library (Simson et al., 2025b), others are not (yet) accompanied by publications. Two secondary software contributions, which are not directly part of the contributing publications but have been utilized in them, are listed in Part III, alongside brief descriptions and links to their source code. The Python package `multiversum` (Chapter 9) allows for specifying and conducting multiverse analyses in Python. It evolved from the multiverse analysis in Chapter 5 and has enabled the analyses in Chapters 6 and 7. Taking inspiration from Hartshorne et al. (2019), `World-Wide-Lab` (Chapter 10) is a platform for running large-scale online data collections with the option of readily scaling computing resources up and down to match traffic. It has powered the data collection for Chapter 6. All publications and software projects in this thesis are accompanied by public repositories, and code is made available under permissive licenses.

Lastly, I end this thesis in Part IV with a conclusion (Chapter 11), where I reflect on the different articles and their contributions within the broader context of the field, including an outlook on future research directions.

2. Background

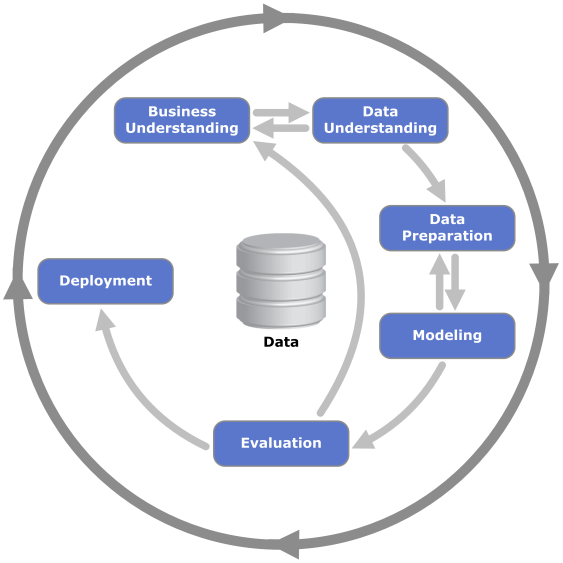
This section aims to provide the required background knowledge for the publications included in the thesis. It begins with a description of typical machine learning (ML) pipelines and data science processes, how they are described in the literature, and the decision space they create in Section 2.1. Next, multiverse analyses are introduced as a method for navigating this space in Section 2.2, followed by an introduction to the field of algorithmic fairness, including its metrics, criteria, and methods in Section 2.3. Afterwards, I discuss the importance and problematic state of data practices in Section 2.4 before moving on to describe the promises and potential perils of participatory design in Section 2.5. Lastly, Section 2.6 briefly describes the well-established ML metrics used in this work.

2.1. The Machine Learning Pipeline

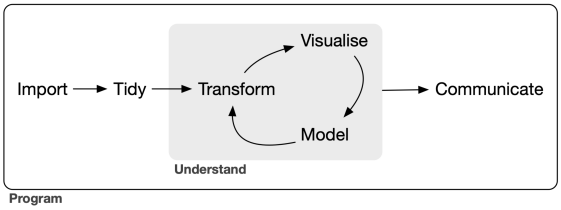
Several people have attempted to describe and classify the various steps of the machine learning pipeline from different perspectives and at varying levels of granularity. These levels can range from conceptual, high-level descriptions, e.g., Yu and Barter (2024), to more specific, low-level steps, often as part of practical implementations, e.g., Olson and Moore (2018). While no single dominant classification or perspective has emerged so far, approaches on a particular level usually share many similarities.

At the highest level, abstract process models typically describe not only a machine learning pipeline by itself, but rather embed machine learning/modeling of data as steps within a larger pipeline or life cycle. These approaches are often framed as data mining, data analysis, and data science life cycles.

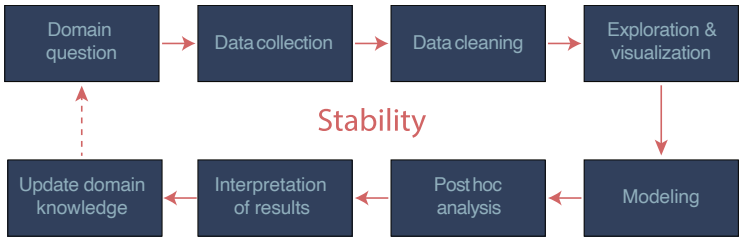
One of the earliest such life cycles to be formalized is the *CRoss-Industry Standard Process for Data Mining* (CRISP-DM, Figure 2.1a) (Chapman et al., 2000). Conceived in late 1996, it is a 76-page document aimed at organizing data-related projects and processes in (large) companies. It organizes the process of working with data into six different phases: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, and eventually *Deployment*. The order of different phases is not strict, and the approach emphasizes going back and forth between phases as well as the cyclical nature of the process, as illustrated in Figure 2.1a. CRISP-DM has been described as "by far the most widely-used analytics process standard" (Brown, 2015).Wickham (2023) describes a slightly more specific approach (Figure 2.1b): data is first *Imported*, then brought into a *Tidy* format (Wickham, 2014). In a loop termed *Understand*, the data is then *Transformed*, *Visualized*, and *Modeled* before results are eventually *Communicated*. In the *PCS Framework* (Figure 2.1c), Yu and Kumbier (2020) and Yu and Barter (2024) describe a more exhaustive approach. Their framework consists of three core principles of data science: *Predictability*, *Computability*, and *Stability*, with Stability being the most central. Predictability corresponds to results reemerging in future data or aligning with prior findings, a "reality check"; Computability



(a) CRISP-DM (Chapman et al., 2000)



(b) Wickham (2023)



(c) PCS Framework (Yu and Kumbier, 2020)

Figure 2.1.: **Illustrations of the data science/data mining life cycle.** (a) The different phases in the CRISP-DM (Chapman et al., 2000) data mining life cycle. Reproduction of illustration by Kenneth Jensen, licensed under CC BY-SA 3.0 (Jensen, 2012). (b) Data science life cycle described in Wickham (2023) and Wickham et al. (2019). Reproduction of figure from Wickham et al. (2019), licensed under CC BY 4.0. (c) Data science life cycle in the PCS Framework (Yu and Kumbier, 2020; Yu and Barter, 2024). Reproduction of illustration from Yu and Kumbier (2020) in accordance with the Standard PNAS License Terms.

corresponds to the real-world constraint of actually being able to compute results, and Stability corresponds to the notion that results are stable to reasonable changes in the data, algorithm, or other processing. They describe the *data science life cycle* as consisting of the following steps[1] (Yu and Kumbier, 2020): A *Domain question* informing *Data collection*, which then leads to *Data cleaning*, followed by *Exploration and visualization*. Afterwards, there is a *Modeling* step, followed by a *Post hoc analysis* and the *Interpretation of results*, which lead to an *Update of domain knowledge*. Updating the domain knowledge has the chance to loop back into the domain question. The importance of the three principles (Predictability, Computability, Stability) applies at every stage of this process.

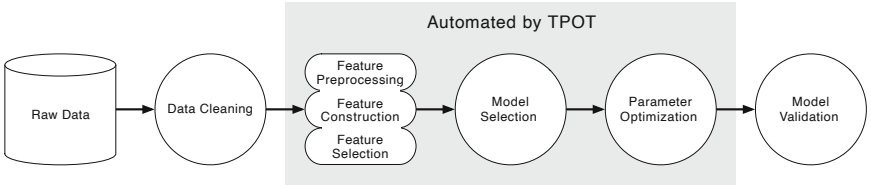
The three approaches were chosen to represent diverse and influental viewpoints: covering an established view from industry in CRISP-DM (Chapman et al., 2000), a practical perspective from the widely-used *tidyverse*[2] (Wickham et al., 2019) and a more recent, but influental view from research through the PCS Framework (Yu and Kumbier, 2020; Yu and Barter, 2024). Two important aspects which all of these three approaches have in common, are (1) their cyclical nature with data science as a continuous process where different phases inform each other, and (2) a processing of data where the original data is transformed (*Data Preparation*, Figure 2.1a; *Tidy* and *Transform*, Figure 2.1b; *Data cleaning*, Figure 2.1c).

Moving beyond high-level descriptions of data science life cycles, more specific examples and definitions of ML pipelines turn out to be hard to find in research. The most prominent field to discuss ML pipelines in this context is the field of *automated machine learning* (AutoML). In AutoML, the goal is to develop systems that automate the ML pipeline either partially or fully (Zöller and Huber, 2021; Pfisterer, 2022). To this end, the field is required to formalize and represent ML pipelines concretely. This is typically done in the form of directed acyclic graphs (DAGs). A directed acyclic graph is a graph consisting of nodes connected by directed edges, where any path along these edges never forms a circle, i.e., loops back to a prior node (a simple example of a DAG is e.g., Figure 2.2a). In the context of ML pipelines, each node represents a particular step in the pipeline, which is connected in a directed fashion with the next to indicate how steps follow onto one another. This graph can then be traversed to implement the full pipeline and make predictions.

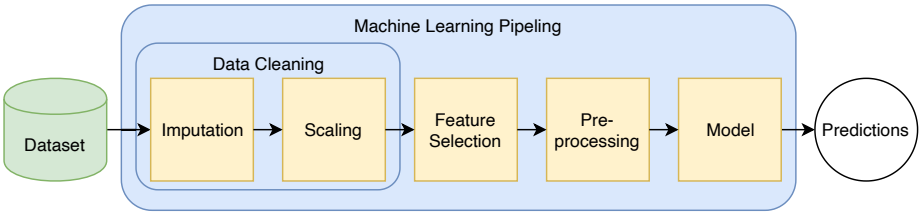
Several examples of (abstract) ML pipelines can be found in Figure 2.2. As context to their *Tree-based Pipeline Optimization Tool* (TPOT), Olson and Moore (2018) include a high-level description of the ML pipeline (Figure 2.2a). While they describe data cleaning only in abstract terms (e.g., formatting, imputation, and general preparation), they provide details and even an ordered list of building blocks for the processing steps within the scope of TPOT (e.g., feature scaling, see Eq. 2.1 and feature selection). Figure 2.2b displays the typical ML pipeline employed by most AutoML frameworks (Zöller and Huber, 2021). They describe the pipeline as a sequence of multiple steps within data cleaning and exactly one step for each of feature selection, variable preprocessing, and modeling. Data cleaning is described as offering a small degree of variability,

[1] Yu and Barter (2024) provide a slightly different variation of the data science life cycle with six grouped phases: *Domain problem formulation and data collection*; *Data cleaning, preprocessing, and exploratory data analysis*; *Exploration of intrinsic data structures*; *Predictive and/or inferential analysis*; *Evaluation of results*; *Communication of results*. I opted to include the variant described in Yu and Kumbier (2020) for licensing reasons.

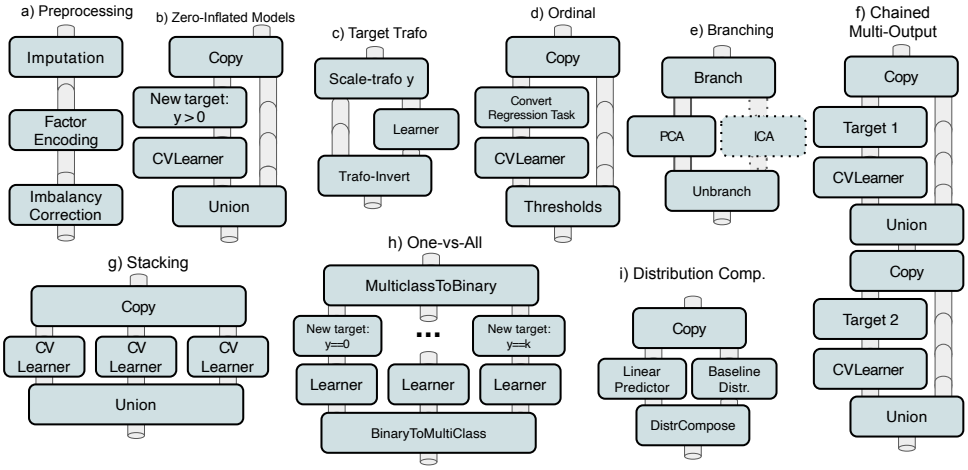
[2] The `tidyverse` package has been downloaded 90,405,084 times from the Comprehensive R Archive Network (CRAN) between the 15th of September 2016 (official release) and the 1st of November 2025; `https://cranlogs.r-pkg.org/downloads/total/2016-09-15:2025-11-01/tidyverse`.



(a) Olson and Moore (2018)



(b) Zöller and Huber (2021)



(c) Binder et al. (2021)

Figure 2.2.: **The diversity of ML pipelines.** Illustrations of the ML pipeline surrounding **(a)** a specific AutoML framework Olson and Moore (2018) and **(b)** a typical AutoML pipeline as implemented by several frameworks Zöller and Huber (2021). **(c)** Various ML pipelines supported by `mlr3pipelines` (Binder et al., 2021), nested lettering inside the subfigure corresponds to the following pipelines: a) classical preprocessing pipeline, b) preprocessing pipeline to deal with zero-inflated data (Zuur et al., 2009), c) transforming a target to be within 0 and 1 (and reversing it for predictions), d) ordinal regression via thresholding of a continuous target, e) support for forking/branching with different options, f) chaining multiple models for multi class prediction (Read et al., 2011), g) stacking models multiple models (Wolpert, 1992), h) multi-class prediction by combination of several binary models, i) distribution composition in the context of survival analysis (Sonabend et al., 2021). Reproduction of figures from Olson and Moore (2018) with permission from the publisher, Zöller and Huber (2021) licensed under the JAIR License 1.0, and Binder et al. (2021) licensed under Creative Commons CC BY 4.0.

with most frameworks implementing some form of imputation and scaling (Eq. 2.1). As Zöller and Huber (2021) note, real-world ML pipelines are often even more complex than the ones presented here, however. Some of this variation can be seen in Figure 2.2c, where several ML pipelines are demonstrated. An analysis of 3,000 ML pipelines also highlights the complexity of real-world ML pipelines and the importance of pipeline steps beyond model training, e.g., different forms of preprocessing (Xin et al., 2021).

At the most concrete level, ML pipelines also correspond to specific implementations and application programming interfaces (APIs) offered by ML software: The popular Python package `scikit-learn` (Pedregosa et al., 2011) provides a `Pipeline()` interface, which allows for multiple processing steps to be chained together. Each step in this pipeline takes in a dataset and typically implements an optional *fit* and *transform* step. The *fit* step corresponds to setting parameters based on the passed data, and *transform* corresponds to a transformation of the passed data. A classical processing step in such a pipeline would be scaling or z-scoring of data, where mean ($\bar{x}$) and standard deviation (σ_x) are calculated during fitting and then used for scaling (Eq. 2.1) during transformation. This design enables fitting the pipeline during model training and reusing it during inference. Comparable implementations exist for R in the `mlr3pipelines` (Binder et al., 2021) and `recipes` (Kuhn et al., 2025) packages.

$$z(x) = \frac{x - \bar{x}}{\sigma_x} \tag{2.1}$$

Similarly, end-to-end ML systems (Baylor et al., 2019; Zaharia et al., 2018; Liberty et al., 2020) allow for the creation of complex pipelines, including tracking of model provenance, as ML pipelines are often continuously executed and ever evolving (Baylor et al., 2019).

Notably, each step in the machine learning process is a choice. Oftentimes, whether or not to include a particular processing step in the pipeline is optional (e.g., scaling), and even in cases where it is not (e.g., having to address missingness in data due to model restrictions), there are usually several alternative options available (Simson et al., 2024b; Ganesh et al., 2025). Each step along the pipeline, therefore, broadens the space of potential alternatives and increases a data scientist's degrees of freedom. While this large space comes with exciting opportunities, it also allows for high subjectivity in how it is navigated. While there is awareness of subjectivity in data generation and labeling (Díaz et al., 2022; Perikleous et al., 2022; Schumann et al., 2023; Beck et al., 2024), it is less present regarding the influence of ML pipeline or design decisions (Hopkins and Booth, 2021). Recent work explores heightening awareness by showing notifications to practitioners as they are traversing and implementing the ML pipeline (Harrison et al., 2024).

2.2. Multiverse Analysis

Concerns have been voiced regarding the validity and reproducibility of research across several disciplines (Ioannidis, 2005; Laraway et al., 2019), including ML and AI (Hutson, 2018; Semmelrock et al., 2023; Gundersen and Kjensmo, 2018; Herrmann et al., 2024). *Multiverse analyses* (Steegen et al., 2016) have been proposed in response to these issues and, in particular, low reproducibility (OPEN SCIENCE COLLABORATION, 2015), implausible findings (Wagenmakers et al., 2011), and questionable research practices (John et al., 2012) in Psychology.

Similar to the decisions one is confronted with when creating or designing an ML pipeline, one encounters a multitude of decisions when analyzing a dataset to, for example, test hypotheses using null-hypothesis significance tests which lie at the heart of the replication crisis in empirical research (Simmons et al., 2011). The decisions one encounters can be explicit and implicit in nature, both during data analysis and ML pipeline design (Simson et al., 2024b; Ganesh et al., 2025).

As there are usually multiple decisions to be made in any data-related project, the process of navigating the decision space has been likened to walking a *garden of forking paths* (Gelman and Loken, 2014). In this garden, each decision corresponds to a fork in the road where one has to choose which arm to proceed on. For explicitly made decisions, one is consciously picking a particular arm and maybe even weighing the different options at hand; however, oftentimes this garden is navigated absent-mindedly, and one may not be aware of the forks one is traversing (Simmons et al., 2011; Simson et al., 2024b; Ganesh et al., 2025). Examples of this are implicitly-made decisions, where one is, for example, adopting preprocessing approaches based on "what one has done in the past" or "what comes to mind" without reflecting on potential alternative choices.

Different people choose different pathways through this garden of forking paths and end up at different endpoints, as demonstrated by *many-analyst studies* (Breznau et al., 2022; Silberzahn et al., 2018; Coretta et al., 2023; Silberzahn and Uhlmann, 2015). In many-analyst studies, multiple researchers (or groups of researchers) each conduct an analysis of the same shared dataset to answer the same shared research question. In a classic example by Silberzahn et al. (2018), 61 different analysts found odds ratios between 0.89 and 2.93 when answering the question of whether players with dark skin tones are more likely to receive a red card than players with light skin tones. Similarly, a study by Breznau et al. (2022), involving 161 researchers (in 73 teams), found 16.9% of statistically significant positive effect sizes, 25.4% of statistically significant negative ones, and 57.7% including zero in their 95% confidence interval (across $n = 1,253$ converged models) when analyzing the question of whether greater immigration reduces the support for social policies. These many-analyst studies expose the *hidden multiverse* of pathways and endpoints in the garden of forking paths.

The goal of a *multiverse analysis* is to make the *garden of forking paths* transparent and assess the variation within it (Steegen et al., 2016). This is done by examining the different choices (i.e., forks) one encounters as part of a data project and collecting all plausible options within a given choice (i.e., arms or paths leading away from a fork). The different options can then be combined so that all unique combinations of options are created to form the full *multiverse* of plausible outcomes. In this multiverse, each unique combination of options corresponds to a single potential and plausible *universe* of how one might have conducted the project. While one usually creates all the unique combinations of different options, it may also be the case that certain decisions are dependent on each other and only plausible in particular combinations. In these instances, only plausible combinations should be included in the final multiverse.

One complication of multiverse analyses is that, as decisions are usually independent of each other, the size of the multiverse tends to grow exponentially with each additional decision (Del Giudice and Gangestad, 2021), as illustrated in Figure 2.3. This means that even for simple (and computationally inexpensive) analyses, a multiverse analysis can quickly be constrained in size by computational cost and feasibility. This is related to a second difficulty of multiverse analyses: The subjectivity of choosing which decisions to include in the analysis and which not. Due to the open-endedness of data analysis and processing (Simmons et al., 2011), there are near-infinite choices and options one may consider for inclusion in a multiverse analysis. Conducting a multiverse analysis, therefore,

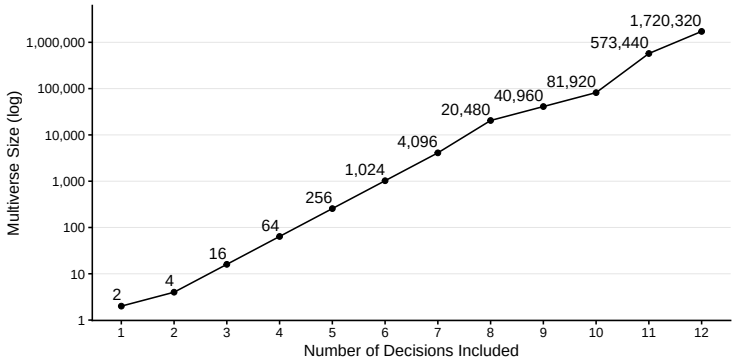


Figure 2.3.: **The multiverse grows exponentially in size with each additional decision.** Illustration of multiverse size based on the number of included decisions for the multiverse analysis conducted in Chapter 5 (Simson et al., 2024b). The Y-axis is on a logarithmic scale so exponential growth appears linear.

necessarily requires making judgment calls as to which decisions to include (Del Giudice and Gangestad, 2021). Computational limitations can (at least partially) be addressed by means of software packages which optimize multiverse computation, through, e.g., parallel exploration of universes (Liu et al., 2021; Sarma et al., 2023), see also Chapter 9.

Multiverse analysis is related to the notion of *Stability* in the PCS Framework (Yu and Kumbier, 2020; Yu and Barter, 2024), *perturbation analysis* (Geisser, 2017), *sensitivity analysis* (Leamer, 1985), *vibration of effects* (Patel et al., 2015) and most closely to *specification curve analysis* (Simonsohn et al., 2020). Specification curve analysis emerged around the same time, and both specification curve and multiverse analysis share the same goals and approaches, i.e., increasing transparency by traversing the garden of forking paths. The main difference between the two approaches (beyond terminology) is that specification curve analysis is characterized by opting for a specific visualization to report results. A specification curve orders different specifications (i.e., universes) by a given outcome metric (e.g., effect size) and plots them on a curve with the outcome metric on the Y-axis. Below, it shows the different decisions and their options in a binarized form, thus visually representing each particular combination of options for each point of the curve. This allows for any sufficiently clear patterns in the relationship between choices and the outcome metric to be visually examined, although specification curves struggle to represent large multiverses well (Del Giudice and Gangestad, 2021). Multiverse analysis has also been used to re-evaluate results from the many analyst study by Silberzahn et al. (2018) (discussed above), arguing for a smaller spread of results once more precise "estimands" are used (Auspurg and Brüderl, 2021).

2.2.1. Connections to Hyperparameter Optimization

Machine learning research examining the influence of decisions is largely situated in the field of *hyperparameter-optimization* (HPO) (Bischl et al., 2023). Work in HPO tries to optimize the search over a hyperparameter space to find optimal configurations as efficiently as possible (Bischl et al., 2023; Feurer and Hutter, 2019). What constitutes the hyperparameter space can be quite flexible and is not necessarily strictly limited to algorithm tuning parameters. In that light, a grid search (Bergstra and Bengio, 2012) over the hyperparameter space can be likened to a multiverse analysis (Bell et al., 2022). Unlike multiverse analyses, however, the explicit goal of HPO is to find a single optimal parameter configuration and to develop search algorithms that allow examining *as*

little as possible of the hyperparameter space/multiverse via, e.g., random search (Bergstra and Bengio, 2012) or Bayesian optimization (Snoek et al., 2012).

When considering the choice of algorithm as a decision to be optimized, *combined algorithm selection and hyperparameter optimization* (CASH) describes the goal of optimizing both algorithm choice and hyperparameters jointly (Thornton et al., 2013; Kotthoff et al., 2017; Zöller and Huber, 2021). CASH, in turn, is at the foundation of AutoML, a movement trying to develop systems to automate the creation (and design) of ML pipelines, including processing, algorithm selection, and optimization (Weerts et al., 2024; Feurer et al., 2015; Erickson et al., 2020; Pfisterer, 2022; Hutter et al., 2019).

In this thesis, I make use of the *functional analysis of variance* (FANOVA) (Hooker, 2007; Hutter et al., 2014) from the HPO literature to quantify the relative importance of decisions in a multiverse analysis (Chapter 5) and the AutoML library `autogluon` (Erickson et al., 2020) as a set of realistic and reasonable models to mirror a typical model selection process in Chapter 4.

2.2.2. Connections to Multiplicity

Akin to a multiverse analysis, several studies have examined sets of plausible machine learning models, though usually smaller in scope and only along a single dimension of variation, e.g., random seeds for train-test splits (Cooper et al., 2024). An interesting finding demonstrated by such studies is the existence of a set of models with equal or similar performance (Rodolfa et al., 2020; D'Amour et al., 2022; Chen et al., 2018; Xin et al., 2022; Cooper et al., 2024; Meyer et al., 2023). This set is commonly referred to as the *Rashomon Set* (Semenova et al., 2022; Xin et al., 2022) – and its existence the *Rashomon Effect* (Breiman, 2001), both named after the acclaimed 1950 movie *Rashomon* (Akira Kurosawa, 1950), which portrays a single event from multiple perspectives. Rashomon sets have been reported within the machine learning multiverse of design decisions (Simson et al., 2024b), sparse decision trees (Xin et al., 2022), definitions of a target variable (Watson-Daniels et al., 2023), data-generating processes (Meyer et al., 2023), and random seeds (Cooper et al., 2024). Ganesh et al. (2025) distinguishes between intentional, conventional, and arbitrary choices in this light. Rashomon sets offer promising opportunities (Rudin et al., 2024) – and potentially legal duties (Black et al., 2024b) – such as selecting for secondary objectives like algorithmic fairness (Section 2.3) within the Rashomon set (Black et al., 2022; Islam et al., 2021).

A commonly observed consequence of the Rashomon Set is that the set consists of models that provide different predictions (or algorithmic decisions) for a given individual. This observation has been described in different fields and under different names, including *predictive multiplicity* (Black et al., 2022), *individual arbitrariness* (Long et al., 2023), and *predictive inconsistency* (Greene et al., 2022). As hinted by the name "individual arbitrariness", this phenomenon can be problematic when the choice of model is both arbitrary for developers, yet consequential for individuals and predictions, and thereby a model's decisions become difficult to justify (Black et al., 2022; Cooper et al., 2024; Long et al., 2023). Several measures have been proposed to quantify the phenomenon (Cooper et al., 2024; Hsu and Calmon, 2022; Dong and Rudin, 2020; Ganesh et al., 2025) and potential solutions to manage it (Black et al., 2022).

2.3. Algorithmic Fairness

The rise of ADM and the widespread deployment of AI and ML systems across various domains, including healthcare, education, finance, and hiring, has led to an increased influence of algorithms on our lives. Although algorithmic fairness predates AI and ML (Cole and Zieky, 2001), their increasing influence has led to a growing interest in fairness in AI and ML (FairML). Given the influence AI and ML systems can have on people's lives, it is crucial that systems not only perform well but are also fair in both their decisions and the decision-making process. While switching from human to algorithmic decision making comes with great opportunities for fairer and more transparent decision making – if done well – it also comes with the risk of introducing discriminatory black-box systems which do not live up to these promises (Barocas and Selbst, 2016).

Salient examples of when systems have not been designed well and lead to serious real-world harms, often to the most vulnerable in society, include the Dutch child care benefits scandal (Persoonsgegevens and Belastingdienst, 2021; Amnesty International, 2021; Henley, 2021), the Australian robodebt system (Henriques-Gomes, 2023), or the United Health Care court case (Mello and Rose, 2024; Atherton, 2023; United States District Court, District of Minnesota, 2023). The field of algorithmic fairness seeks to prevent these types of issues from occurring by providing criteria to empirically assess fairness, by building methodologies to create fairer models, and by bringing awareness and attention to these issues – among other things.

A central concept in algorithmic fairness, which originates in anti-discrimination law, is that certain (demographic) attributes are considered *protected* (or *sensitive*) (Barocas et al., 2023; Mitchell et al., 2021). Which attributes are considered protected by law is often dependent on a particular context and legislation (Simson et al., 2024a), e.g., US Fair Lending law (Chen et al., 2019) or EU Fair Hiring law (Fabris et al., 2025), but more general charters exist as well, such as the Universal Declaration of Human Rights (United Nations, 1948). As an example, Article 21 of the Charter of Fundamental Rights of the European Union states:

> "Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited." — European Parliament, Council, and Commission (2007)

2.3.1. Criteria

While any given person might usually have a readily available sense of which things they deem to be fair – or often more clearly which not – formalizing these intuitive notions of fairness is not always that straightforward, yet required if we want to measure or even optimize for them. On top of this, a steadily growing body of research demonstrates that notions of fairness vary between people and can (even within a single person) be dependent on context (Starke et al., 2022; Binns et al., 2018; Kern et al., 2022).

There exists a large array of different definitions to operationalize algorithmic fairness (Barocas et al., 2023; Mehrabi et al., 2021; Kusner et al., 2017; Chouldechova and Roth, 2020). So much so that the abstract and numerical focus of the field has been a point of critique (Birhane et al., 2022b; Lee et al., 2021). In this section, I will only discuss criteria of algorithmic fairness in

relation to tabular classification tasks, the most common task in the field (Fabris et al., 2022), and the most salient in the context of algorithmic decision making. A rich field of task and domain-specific criteria exists outside of tabular classification, e.g., fairness metrics for image upsampling (Laszkiewicz et al., 2024) and bias evaluations for generative text-to-image models (Cho et al., 2023). While many of these notions of fairness are applicable to decision-making processes in general (irrespective of whether they are automated or made by a human), I discuss them in the context of algorithmic decision-making as it matches the context of the thesis more closely. For illustrative purposes, I will assume a scenario with a binary outcome where an algorithm's prediction is equal to the actual decision, which is not a given (Kuppler et al., 2022). Moreover, I assume that a positive decision $\hat{Y} = 1$, i.e., acceptance, is the favorable outcome, and a negative one $\hat{Y} = 0$ is not. We have a protected (demographic) attribute S with groups g (e.g. g_1 and g_2) and knowledge of the ground truth Y, where $Y = 1$ denotes that acceptance was the *correct* decision for a particular individual.

Notions of algorithmic fairness are typically examined on one of two levels: the individual level or the group level (Mehrabi et al., 2021; Pessach and Shmueli, 2022). As an exception to this rule, subgroup fairness lies between individual and group fairness, trying to get the best from both worlds (Kearns et al., 2018, 2019; Hebert-Johnson et al., 2018) by focusing on finer groups, e.g., intersections.

On the individual level, fairness through awareness (Dwork et al., 2012) aims to give similar predictions to similar individuals: specifically, it posits that an algorithm is fair, if its predictions are similar for individuals i and j who are similar under a distance metric $d(i, j)$ (Kusner et al., 2017; Castelnovo et al., 2022). Naturally, the distance metric is highly important (Dwork et al., 2012). Counterfactual fairness, a different notion on the individual level, posits that a system is to be considered fair if an individual's prediction would be the same in a counterfactual world where their membership of a group g of the protected attribute S would be different (Kusner et al., 2017). Last, there exists the notion of fairness through unawareness, where a system is considered fair if it does not make use of protected attributes, e.g., S for its predictions (Grgic-Hlaca et al., 2016). This notion has been criticized, however, as proxy variables have been shown to still be able to affect predictions even when protected attributes are not used directly (Pessach and Shmueli, 2022; Barocas et al., 2023; Kleinberg et al., 2018).

Group fairness, on the other hand, asks that different groups g on a protected attribute S are treated equally (Mehrabi et al., 2021; Dwork et al., 2012; Castelnovo et al., 2022). Barocas et al. (2023) discuss three criteria underlying the more applied statistical criteria of algorithmic fairness: *Independence*, *Separation*, and *Sufficiency*.

Independence corresponds to the notion that, for a protected attribute S, model decisions are independent of or unrelated to S. Concretely, this means that for different groups g of S, acceptance rates have to be equal (Eq. 2.2). This criterion is also referred to as statistical or demographic parity, group fairness (before the term was used as a category), or disparate impact (Dwork et al., 2012).

$$\mathbb{P}\{\hat{Y} = 1 \mid S = g_1\} = \mathbb{P}\{\hat{Y} = 1 \mid S = g_2\} \tag{2.2}$$

One of the benefits of Independence – as illustrated in Eq. 2.2 – is that it does not rely on the existence of a ground truth Y. This has the benefit that it can be applied to data and

scenarios where *existing decision data* might be biased. The word decision is important here, as the notion that acceptance rates should be equal across groups is highly context-, legislation- and sample-dependent.

Separation corresponds to the notion that error rates should be equal across different groups g of the protected attribute S. The way this is commonly formulated is that the *true positive rate* (TPR) and the *false positive rate* (FPR) should be equal for all groups g of S (Eqs. 2.3 and 2.4). This criterion is most commonly known as equalized odds, but has also been referred to as disparate mistreatment, error rate balance, and conditional procedure accuracy equality (Pessach and Shmueli, 2022).

$$\mathbb{P}\{\hat{Y} = 1 \mid Y = 1, S = g_1\} = \mathbb{P}\{\hat{Y} = 1 \mid Y = 1, S = g_2\} \tag{2.3}$$

$$\mathbb{P}\{\hat{Y} = 1 \mid Y = 0, S = g_1\} = \mathbb{P}\{\hat{Y} = 1 \mid Y = 0, S = g_2\} \tag{2.4}$$

A relaxed variant of this criterion requiring only equality of the *true positive rate* (i.e., only Eq. 2.3) is known as *equal opportunity* or as false negative error rate balance (as the false negative rate is the inverse of the true positive rate, $FNR = 1 - TPR$) (Verma and Rubin, 2018).

Sufficiency corresponds to the notion that the true outcome Y is independent of the protected attribute S given the model's prediction $\hat{Y}$ (Eq. 2.5; or its predicted score R, in that case Eq. 2.6), i.e., $\hat{Y}$ is *sufficient* to predict Y and there is no additional information in S.

$$\mathbb{P}\{Y = 1 \mid \hat{Y} = y, S = g_1\} = \mathbb{P}\{Y = 1 \mid \hat{Y} = y, S = g_2\} \tag{2.5}$$

$$\mathbb{P}\{Y = 1 \mid R = r, S = g_1\} = \mathbb{P}\{Y = 1 \mid R = r, S = g_2\} \tag{2.6}$$

Sufficiency (and related criteria) are less commonly used than Independence and Separation (Pessach and Shmueli, 2022). Moreover, the three criteria of Independence, Separation, and Sufficiency have been shown to be in conflict with each other, such that one can not fulfill them at the same time (Barocas et al., 2023; Pessach and Shmueli, 2022).

2.3.2. Metrics

Metrics of algorithmic fairness are usually defined as deviations from the operationalized fairness criteria discussed in the prior section, i.e., to what degree a criterion is fulfilled or not. Here, I will describe two commonly used fairness metrics (Pessach and Shmueli, 2022; Weerts et al., 2023) employed in publications within this work: Equalized Odds Difference (EOD; Eq. 2.7) and Demographic Parity Difference (DPD; Eq. 2.8). The two metrics were chosen as they represent the two popular criteria of Independence (DPD) and Separation (EOD).

For a protected attribute S with groups g, the two fairness metrics are defined as follows (additional context on metrics in Section 2.6).

$$\text{EOD} = \max_{y \in \{0,1\}} \left(\max_g \mathbb{P}(\hat{Y} = 1 | Y = y, S = g) - \min_g \mathbb{P}(\hat{Y} = 1 | Y = y, S = g) \right) \tag{2.7}$$

$$\text{DPD} = \max_g \mathbb{P}(\hat{Y} = 1|S = g) - \min_g \mathbb{P}(\hat{Y} = 1|S = g) \tag{2.8}$$

As both metrics quantify the deviation from their respective criteria, a score of 0 corresponds to the criterion being fulfilled and a score of 1 to the criterion not being fulfilled at all. Different operationalizations for the same criteria exist, e.g., calculating a relative ratio compared to the absolute difference (Weerts et al., 2023).

2.3.3. Methods and Algorithms

Several techniques, methods, and algorithms have been developed in FairML to create models that satisfy the statistical criteria for non-discrimination (Friedler et al., 2019; Defrance et al., 2024; Pessach and Shmueli, 2022; Bellamy et al., 2018). These algorithms are generally classified into one of three categories, based on where in relation to the model fitting step in the ML pipeline they are situated (Barocas et al., 2023; Pessach and Shmueli, 2022; Bellamy et al., 2018). These three categories are *Pre-*, *In-*, and *Post-Processing*. Each category is briefly described below, alongside relevant algorithms that have been used in the publications of this thesis.

Pre-Processing Methods are applied to and modify the training data directly rather than the model itself. They are applied prior to model training to reduce bias in the training data before a regular machine learning model is trained on the data. The following two preprocessing methods are used in this thesis. *Learning Fair Representations* is an approach by Zemel et al. (2013) where fairness is treated as an optimization problem of encoding data as well as possible, while at the same time obfuscating any information about the protected attribute. It optimizes for independence and individual fairness. *Disparate Impact Remover* by Feldman et al. (2015) focuses on the notion of independence (i.e., disparate impact, see above) and updates the data to reduce disparate impact while preserving rank order.

In-Processing Methods are applied during or as part of the model training step, or they are themselves the model training step. They include constraints from fairness notions into the training process itself to produce a model in line with the specified notion(s). *Adversarial Debiasing* (Zhang et al., 2018) trains both a predictor and an adversary model. The predictor tries to predict the target Y from the features X, while the adversary tries to predict the protected attribute S from the predictor's predictions $\hat{Y}$. The goal is to maximize the predictor's ability to predict Y while simultaneously reducing any bias in its predictions in regards to the protected attribute S. The *Meta-Algorithm* developed by Celis et al. (2019) allows for specification of a fairness constraint and is then able to provide a classification model with provable guarantees. *Rich Subgroup Fairness / GerryFair* (Kearns et al., 2018, 2019) supports several prediction algorithms as well as fairness notions and has been developed to optimize these for subgroup fairness. *Grid Search Reduction* (Agarwal et al., 2018) reduces the problem of fair classification down to multiple cost-sensitive classification problems (Elkan, 2001), where different types of errors differ in how costly they are, and existing algorithms are available. A similar algorithm is available for regression tasks (Agarwal et al., 2019).

Post-Processing Methods are applied after a regular ML model has been trained. The "initial" predictions of a model are taken in by a method and modified/overwritten to align with a fairness notion. *Group-Specific Thresholds* by Hardt et al. (2016) provides such an algorithm, which allows for different classification thresholds between different groups of the protected attribute S to optimize for the fairness notion, e.g., separation (i.e., equalized odds).

Recent work has argued that when accounting for algorithm complexity, post-processing methods are consistently superior or comparable to other processing approaches on the Pareto-frontier of fairness and performance (Cruz and Hardt, 2024).

2.3.4. Manipulating Fairness

Different ways have been described of how notions and metrics of algorithmic fairness can be vulnerable to manipulation. Meding and Hagendorff (2024) first coined the term *fairness hacking* in this context. They demonstrate one instance of fairness hacking, showing that it is possible to favorably describe a particular model by calculating a large array of metrics and only reporting the most favorable metric(s). Similarly, in Chapter 5, we demonstrate how a bad actor could search over the space of plausible evaluation strategies of a model to find and report only the most favorable one without modifying the model itself (Simson et al., 2024b). Thankfully, conducting and reporting a multiverse analysis over the evaluation space is also a promising solution to protect against such practices. Similar observations have been described under the terms *d-hacking* (Black et al., 2024a), although they are more closely related to the concept of *predictive multiplicity* (Section 2.2.2), given that the model itself changes.

2.4. Data Practices

Data is at the heart of ML and AI (Gebru et al., 2021; Hardt, 2025; Whang et al., 2023). The availability and usage of datasets can impact the trajectory of the field greatly. One example of this is the ImageNet dataset (Deng et al., 2009), which was foundational in the rise of deep learning methods in the field (Fei-Fei and Krishna, 2022).

This foundation is brittle, however, as datasets are often used without critical reflection, being treated as a simple source of optimization problems without real-world context. The reality is a very different one, however: Datasets are rich in context, they are sourced by, annotated by, and include or represent people with goals, convictions, and identities.

In reality, either the datasets themselves or the way they are used can be quite problematic. Popular datasets in the FairML literature (Fabris et al., 2022), such as *Adult* (Kohavi, 1996), *COMPAS* (Angwin et al., 2016), and *German Credit* (Hofmann, 1994) all come with issues (Ding et al., 2021; Grömping, 2019; Bao et al., 2022), as do datasets in other domains (Birhane and Prabhu, 2021; Luccioni et al., 2022). These issues can range from methodological ones in usage (e.g., deducing gender from insufficient information (Grömping, 2019)) to problems as severe as the unintentional distribution of child abuse material (Schuhmann et al., 2022; Thiel and Hancock, 2023). In several instances, this has led to the depreciation of datasets[3] (Luccioni et al., 2022).

[3] https://neurips.cc/public/deprecated-datasets

Which and how populations are represented in data is also an important issue. *Western, Educated, Industrialized, Rich and Democratic* (WEIRD) (Henrich et al., 2010) populations are much more likely to be represented in data (Septiandri et al., 2023; Simson et al., 2025b; Urman et al., 2025), especially so from the United States of America, whereas populations from the Global South are not (De et al., 2025; Okolo et al., 2022). Complex tensions along geography and power are also at play when it comes to the annotation or labeling of data (Díaz et al., 2022; Casilli et al., 2024).

These issues are further complicated by a lack of transparency in *how* a dataset is used. Only a small fraction of research publishes code alongside their research, and even if it does, it is often only snippets of example code that describe a new methodology, rather than reproducible code of experiments (Hutson, 2018; Semmelrock et al., 2023; Gundersen and Kjensmo, 2018). Paired with a lack of clear reporting in papers themselves, this makes it difficult to understand which features X, protected attributes S (in the context of FairML), and even targets Y were used (Simson et al., 2024a). In certain cases, it is not even clear *which dataset* is used, as there are often competing variants of datasets (e.g., in the case of Bank Marketing (Moro et al., 2014)) which are not clearly identified all the while the literature suggests identifying datasets by short – often unofficial and inconsistently chosen – names is sufficient to clearly identify them (Simson et al., 2025b).

2.5. Participatory Design

Originating in Scandinavia during the 1970s and 1980s, *participatory design* has a rich history of questioning existing conventions and empowering people (Muller and Kuhn, 1993; Spinuzzi, 2005). Following revelations of harmful systems across several domains and a subsequent shift in public attitudes towards AI and ML, participatory design has also been applied to AI and ML. *Participatory AI* and *participatory ML* (usually used synonymously) are the result of this and aim to address the power differential between developers of AI/ML systems and those affected by them (Kulynych et al., 2020). The goal here is to create systems that are more fair, inclusive, just, accountable, and transparent by incorporating stakeholder input (Birhane et al., 2022a; Feffer et al., 2023b; Delgado et al., 2023).

Whether participatory design – and by extension participatory AI and ML – actually lives up to its promises of empowerment depends on the degree of *actual participation* within any application of it. Figure 2.4 illustrates the risk of unempowered forms of participation. In this light, Arnstein (1969) provides a "ladder of participation" to assess different degrees of participation (in a simplified manner). Ranging from manipulation (lowest) to citizen control (highest), the ladder provides eight different levels of increasingly more empowering participation. At its lower end, *manipulation* (1) and *therapy* (2) are classified as "nonparticipation", essentially aiming for the opposite of participation by pushing viewpoints onto participants. Next, *informing* (3), *consultation* (4), and *placation* (5) fall under "degrees of tokenism" as they offer the chance for participants to be heard but do not offer any such guarantees. Lastly, *partnership* (6), *delegated power* (7), and *citizen control* (8) are grouped under the category of "degrees of citizen power." Here, participants hold real power, ranging from the ability to negotiate to making decisions. The participatory ladder purposefully contrasts those in power with those not in power – with a clear relation towards socioeconomic status – into just two groups, something the original work already acknowledges as an oversimplification. Nevertheless, it continues to offer a helpful framework to identify problematic forms of participation, in line with the sentiment expressed in Figure 2.4, and retains its relevance to this day (Delgado et al., 2023; Birhane et al., 2022a).

Figure 2.4.: **Imbalances in power can be problematic for participation and by extension participatory design, AI and ML.** A political poster from the French student uprising in May 1968, now located in the Bibliothèque nationale de France (1968) reading: "I participate, you participate; he participates; we participate; you all participate; they profit" (translated to English). Inclusion inspired by Arnstein (1969).

The underlying concepts behind the "ladder of participation" designed within the context of city planning and public governance, also apply within the context of participatory ML and AI. While ample examples exist in the space at this point (Halfaker and Geiger, 2020; Denecke et al., 2019; Brown et al., 2019; Lee et al., 2019; Huang et al., 2024), the majority of work does not engage in participation beyond a "consultation" phase (Corbett et al., 2023). It is also relatively rare for participatory AI / ML to include choices regarding the design or evaluation of models (Delgado et al., 2023). The work included in Chapter 6 of this thesis moves beyond this view to explicitly include participants in design and evaluation decisions across the ML pipeline (Simson et al., 2025a).

Even when participants are solely *consulted* as respondents in a survey, there is often room for more agency, as they are commonly assigned to categories without a chance of confirmation. Examples of this include classification of skin color by interviewers (Telles and Steele, 2012) or occupation after an interview based on open text responses (Schierholz et al., 2018; Simson et al., 2023). Inferring attributes without confirmation carries risks of misassignment and harm, especially when it happens for attributes that may be considered protected (Andrews et al., 2023).

Despite its promises, participatory AI/ML comes with limitations and risks (Birhane et al., 2022a; Corbett et al., 2023). A central risk of participatory design in general is "performative participation" (Corbett et al., 2023) or co-optation (Birhane et al., 2022a), where input is gathered solely for performative reasons, but not taken into account for actual decision-making (i.e., nonparticipation under Arnstein (1969)). This risk can be mitigated by formal and binding a priori commitments and established pathways for integrating participatory input. Moreover, enabling participants to

co-design the participatory process itself offers a powerful avenue for higher degrees of participation as well as novel discoveries and learnings for developers of how processes may be organized. Moreover, participatory design comes with the risk of a tyranny of the majority (Feffer et al., 2023a), where decisions are made by – and more importantly, to the benefit of – a majority and to the detriment of any minorities.

Altogether, the principles of participatory design, especially in their relation to power and empowerment, offer valuable information to inform how we design, deploy, and train ML and AI systems as well as collect their training data.

2.6. Metrics

Several metrics are used across the publications in this thesis. Metrics of algorithmic fairness are introduced and specified in Section 2.3.2. Performance metrics are introduced and specified below, along with the underlying measures used to calculate both performance and fairness metrics.

Table 2.1.: Confusion matrix for a classification task.

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

The different metrics used in the thesis all build upon the classic confusion matrix (Table 2.1), which is created by comparing a model's predictions (columns) with an existing ground truth (rows) for a holdout or test dataset that the model has not been trained on. Several measures can be calculated from this confusion matrix.

$$\text{Precision} = \Pr(Y = 1|\hat{Y} = 1) = \frac{TP}{TP + FP}$$
$$\text{Recall / Sensitivity} = \Pr(\hat{Y} = 1|Y = 1) = \frac{TP}{TP + FN}$$
$$\text{Specificity} = \Pr(\hat{Y} = 0|Y = 0) = \frac{TN}{TN + FP}$$

We use *Accuracy* (Acc; Eq. 2.9), *Balanced Accuracy* (bAcc; Eq. 2.10), and *F1 Score* (Eq. 2.11) as measures of performance. The performance metrics are defined as follows:

$$\text{Acc} = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.9}$$

$$\text{bACC} = \frac{\text{Specificity} + \text{Recall}}{2} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{2.10}$$

$$\text{F1 Score} = \frac{2}{\text{Precision}^{-1} + \text{Recall}^{-1}} = \frac{2TP}{2TP + FP + FN} \tag{2.11}$$

3. Contributing Publications

The following chapter contains a brief introduction to each publication included in this thesis. The chapter situates each work and its contributions in relation to the other papers in the thesis and the broader literature. This includes how different works in the thesis build on each other, oftentimes addressing problematic practices in earlier work and then working towards a solution in later works.

3.1. Lazy Data Practices Harm Fairness Research

The first paper in this thesis, "Lazy Data Practices Harm Fairness Research", highlights a series of problematic practices observed in the literature on fairness in ML. The paper focuses particularly on data practices and highlights the central role of data (and how it is used) within research. It is situated at the early stages of the ML pipeline, focusing on both how and which datasets are created and used. The paper identifies several problematic practices that are picked up and (partially) addressed in later work in this thesis. It identifies in particular the following three "lazy" data practices: (1) the neglect of certain identities, (2) the omission of (sub) populations, and (3) opaqueness in (pre-)processing of data.

The first lazy data practice discussed in the paper, *neglected identities*, refers to the mismatch between identities/attributes that are deemed protected under anti-discrimination legislation and those protected attributes examined in the FairML literature. To demonstrate this, the paper compiles a list of protected attributes under several legislations and compares it with both the availability in datasets and the usage within empirical experiments. Both datasets and experiments are sourced from the FairML literature. This comparison reveals a pronounced discrepancy, as several legally protected attributes are rarely utilized in experiments and often not available in datasets to begin with. Salient examples of this are *religion*, which is considered protected under all eight examined legislations, but which is not represented in a single one of the examined datasets, and *disability*, considered protected under five legislations and available in three datasets, but not used in a single empirical experiment. The reasons for this lack of representation in datasets can be complex, and there are difficult trade-offs to navigate, such as the one between representation and privacy. The second lazy data practice, *omitted populations*, refers to the common practice of excluding or aggregating smaller groups of a protected attribute. The paper demonstrates this by systematically annotating the processing of the protected attribute *race* in the popular *COMPAS* dataset (Angwin et al., 2016). Only *one* out of 59 publications chose to use the full protected attribute as is, whereas 35 retained only the two biggest groups and 18 aggregated the protected attribute into only two groups. The third and final practice, *opaque preprocessing*, refers to the observation that there is a lack of transparency on how datasets are used in the literature. This ambiguity is the product of two separate observations: (1) a large part of the examined literature does not describe its usage and processing of data in sufficient detail, and (2) in instances

where there is sufficient information, there is high diversity in the usage and processing of data. In a case study using the popular *Bank* dataset (Moro et al., 2014), we demonstrate that the observed variation in processing affects FairML metrics and may influence a practitioner's choice of algorithm.

The paper concludes with a series of recommendations for the field aimed at (partially) addressing the lazy data practices described. While these recommendations are important, the paper's primary contribution lies in highlighting the problematic nature of such practices. Subsequent works in this thesis build on this foundation by focusing on addressing and improving these issues.

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

The second paper in the thesis, "One Model Many Scores: Using Multiverse Analysis to Prevent Fairness Hacking and Evaluate the Influence of Model Design Decisions", focuses in particular on the influence of those decisions along the machine learning pipeline that tend to be overlooked. As the previous paper has shown, there is often a high degree of variation as well as opaqueness in how a dataset is preprocessed. In this paper, the influence of various decisions along the ML pipeline is systematically examined by introducing multiverse analyses into algorithmic fairness research. Several plausible decisions regarding model design and evaluation are combined to form a large multiverse, which is then traversed to assess the influence and importance of individual decisions. This allows directing attention towards the most impactful decisions. The study analyzes this multiverse in two steps: First, using a fixed evaluation strategy to examine the influence of design decisions, a multiverse of $N = 61,440$ models is examined. Second, a larger analysis examining 28 potential evaluation strategies for any given model creates a multiverse of $N = 1,720,320$ different scores.

A noteworthy finding of the paper is the high degree of variability in fairness metrics resulting from changes in the evaluation protocol. The vast majority of models (94.51%) displayed a spread ($\Delta = max(x) - min(x)$) along the fairness metric of $\Delta \geq 0.9$, solely by varying how the model is evaluated. This opens the door for fairness hacking, where a malicious actor could explore the multiverse of evaluation strategies to find and report only the most favorable evaluation strategy. This allows a flawed ML model to be presented as "fair" to regulators, stakeholders, and the public. Thankfully, multiverse analyses also offer a potential solution for this problem: by reporting a multiverse analysis across different evaluation protocols, one can counteract practices of fairness hacking and increase the robustness of reported scores.

3.3. Preventing Harmful Data Practices by using Participatory Input to Navigate the Machine Learning Multiverse

The third paper, "Preventing Harmful Data Practices by using Participatory Input to Navigate the Machine Learning Multiverse", builds on the foundations laid in the previous two papers by

introducing a novel approach to navigate the machine learning multiverse that helps to improve data practices.

Building on and expanding beyond the case study presented in the second paper, this paper compiles a set of decisions along the machine learning pipeline and rephrases them to be accessible to a lay audience. These decisions are then presented to the general public to vote on, allowing them to choose which options they deem appropriate and which ones they do not. The decisions are further used to construct a multiverse and conduct a multiverse analysis. The results from this multiverse analysis are then linked and weighted based on the respondents' input to produce a more compelling space of models.

While respondents provided generally high-quality responses, their stance on lazy data practices in particular stands out: in contrast to practices in the literature, the vast majority of respondents voted against the omission of subpopulations. This makes participatory input a promising avenue in navigating degrees of freedom during the design of ML systems.

3.4. Bias Begins with Data: The FairGround Corpus for Robust and Reproducible Research on Algorithmic Fairness

The fourth paper in this thesis, "Bias Begins with Data: The FairGround Corpus for Robust and Reproducible Research on Algorithmic Fairness", directly aims to improve practices around data usage and sourcing. The paper focuses on the crucial role of datasets in research, highlighting how they can shape and influence the research itself.

The paper introduces *FairGround*, a corpus of richly annotated datasets as well as a supporting framework including a Python package and multiple dataset collections. The corpus comprises 44 distinct datasets, which can be easily loaded and preprocessed using the package. The project aims to address opaqueness and inconsistency in data preprocessing by offering a default processing pipeline with sensible defaults. The processing pipeline in the package is easily modifiable to allow for diversity in processing choices, while requiring any deviations from defaults to be made with more explicit choices. The dataset annotations in conjunction with the Python package also make it easier to conduct analyses with multiple datasets, with the goal of increasing the number of datasets with which new methods, algorithms, and approaches are evaluated. By quantifying dataset metadata, such as geographic representation, FairGround highlights gaps in currently available data and aims to inspire the sourcing of new datasets from underrepresented regions. At the same time, the creation of several dataset collections makes it easy to use collections of datasets that are highly diverse in how different fairness-aware algorithms perform on them.

Among the different works in this thesis, FairGround is the most clearly solution-oriented, directly aiming to improve some of the problematic data practices from the first paper. By offering a comprehensive framework, the project's goal is not only to provide and support the current dataset corpus but also to provide a blueprint for application in other disciplines or data types.

3.5. occupationMeasurement: A Comprehensive Toolbox for Interactive Occupation Coding in Surveys

The final paper in the thesis, "occupationMeasurement: A Comprehensive Toolbox for Interactive Occupation Coding in Surveys", introduces a new R package for the interactive coding of occupational data into official classifications. In line with other works in this thesis, this paper questions existing practices regarding the deployment of ML models and the sourcing of data. Rather than applying ML-based suggestions post-hoc, it explores the opportunity of enabling respondents to be in the loop.

Coding a person's occupation into an official classification scheme, such as the *ISCO-08* (ILO, 2012) or the *KldB-2010* (Bundesagentur für Arbeit, 2011), is a difficult task. Current practice is to collect free-text responses from respondents and then have expert annotators code these into classifications post-hoc. However, even for expert annotators, this task is challenging; agreement among annotators is often low, and annotations are costly. In this project, we implement and release an R package to allow for interactive occupation coding *within* surveys. A pre-trained machine learning model parses responses and generates likely options for participants to choose a category from. Using an auxiliary coding scheme (Schierholz, 2018), follow-up questions are used to further refine categories.

By allowing respondents to choose their own category in the classification – rather than be assigned to one without any say – the paper shifts power to respondents. Similar to the work on participatory design, the paper works towards empowering respondents in how their data is encoded.

Part II.

Publications

4. Lazy Data Practices Harm Fairness Research

Contributing article

Simson, J., Fabris, A., & Kern, C. (2024). Lazy Data Practices Harm Fairness Research. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 642–659. doi: 10.1145/3630106.3658931 URL https://doi.org/10.1145/3630106.3658931

Code repository

https://github.com/reliable-ai/lazy-data-practices

Copyright information

Author contributions

The idea for the project naturally emerged during early work on the contribution titled "Bias Begins with Data: The FairGround Corpus for Robust and Reproducible Research on Algorithmic Fairness". J. Simson provided the initial idea for the collaboration. A. Fabris provided access to initial annotation data and background information on them. J. Simson developed the annotation process in close collaboration with the other authors. A. Fabris reviewed and provided feedback on additional annotations. J. Simson implemented the empirical experiments in the work; analyzed the data and created all figures. A. Fabris played a key role in structuring and naming the article and produced the annotations for Table 1. J. Simson lead the writing, submission and revision process of the work. All authors contributed through fruitful comments, proofreading and revisions of the manuscript.

F. Weber and A. Kreider, mentioned in the acknowledgements of the paper provided help with the annotation process.

Lazy Data Practices Harm Fairness Research

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

Alessandro Fabris
Max Planck Institute for Security and Privacy
Bochum, Germany
alessandro.fabris@mpi-sp.org

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

Data practices shape research and practice on fairness in machine learning (fair ML). Critical data studies offer important reflections and critiques for the responsible advancement of the field by highlighting shortcomings and proposing recommendations for improvement. In this work, we present a comprehensive analysis of fair ML datasets, demonstrating how unreflective yet common practices hinder the reach and reliability of algorithmic fairness findings. We systematically study protected information encoded in tabular datasets and their usage in 280 experiments across 142 publications.

Our analyses identify three main areas of concern: (1) a **lack of representation for certain protected attributes** in both data and evaluations; (2) the widespread **exclusion of minorities** during data preprocessing; and (3) **opaque data processing** threatening the generalization of fairness research. By conducting exemplary analyses on the utilization of prominent datasets, we demonstrate how unreflective data decisions disproportionately affect minority groups, fairness metrics, and resultant model comparisons. Additionally, we identify supplementary factors such as limitations in publicly available data, privacy considerations, and a general lack of awareness, which exacerbate these challenges. To address these issues, we propose a set of recommendations for data usage in fairness research centered on transparency and responsible inclusion. This study underscores the need for a critical reevaluation of data practices in fair ML and offers directions to improve both the sourcing and usage of datasets.

CCS CONCEPTS

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

KEYWORDS

critical data studies, protected groups, fair ML generalization, reproducibility

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658931

ACM Reference Format:
Jan Simson, Alessandro Fabris, and Christoph Kern. 2024. Lazy Data Practices Harm Fairness Research. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3630106.3658931

1 INTRODUCTION

The identification and mitigation of harms against vulnerable individuals and groups embedded in data-driven algorithms lies at the core of fairness in machine learning (fair ML) research. Discriminatory practices take on various forms, affect a multitude of social groups in different contexts, and are often targeted against (intersecting) minority populations. Investigating discrimination in sociotechnical systems requires adequate and nuanced data sources as well as careful operationalizations of vulnerable groups. Data is highly influential in fair ML research. On the one hand, novel fairness methodology is typically developed and "benchmarked" in empirical applications, and thus the underlying data can be used to support the argument in favor of a specific technique. On the other hand, the information that is encoded and readily accessible in fairness data defines the scope of what can be tested empirically, priming fairness research to e.g. focus on those protected attributes that are most easily accessible. Practices concerning *which* data is used in published research, and *how* it is used, further set a standard for both practitioners and future research.

In this work, we study data practices in fairness research and identify common shortcuts that undermine its reach and reliability. Particularly, we study which protected groups are represented in datasets commonly used in fair ML and how the available data is utilized in the literature, identifying blindspots such as neglected identities and omitted subpopulations in data usage. We argue that through their wide range of applications, fairness datasets and their uses play a pivotal role in fairness research as they can be both drivers and barriers for sound methodological and empirical research.

More specifically, we study the *content* of fairness datasets in interaction with their *uses* in empirical research. This dual view is motivated by the concern that limitations inherent to the datasets themselves can be exacerbated by unreflective choices made in the processing and handling of these data. Both factors can jointly accumulate to the risk of neglecting "uncommon" protected attributes or specific subpopulations and contribute to normalize this practice, leading to a vicious cycle of canonical fairness research which

focuses on a limited set of social groups and the same standard datasets [42].

Related work. Critical studies have challenged research practices in fair ML on various grounds. Concerns have been raised regarding its narrow and too granular focus, tendencies of insularity [65], inconsistent notions of race [1], and a predominance of shallow discussions of specific negative impacts that neglect structural and social factors [14]. Critical data studies [16, 59] view these questions from a data-centric lens. Selected challenges have been tied to the empirical foundation of fair ML research, such as its overreliance on WEIRD (Western, Educated, Industrialized, Rich, and Democratic) samples [86] and a large share of fairness publications drawing on the same datasets, namely Adult, COMPAS and German Credit [42]. As these data come with considerable limitations [10, 34], there is a risk of self-perpetuating practices that steer empirical fairness research away from the social realities and diversity its data is supposed to represent.

Contributions. Against this background, we focus on both the scope of fairness datasets and their uses in empirical research to understand the interaction between limitations in datasets and the choices that are made in the handling of these data. We study 280 experiments across 142 fair ML publications and identify gaps in collective data practices hindering the reach and reliability of the field. Our study makes the following contributions:

- We present an inclusive list of attributes protected by anti-discrimination legislation across multiple continents and study their (under)representation in fairness datasets, as well as discrepancies between protected attribute availability and usage in fair ML research.
- We outline exclusionary patterns in empirical studies and demonstrate how a lack of transparency and unreflective processing choices normalize the omission of minorities and lead to ambiguous results in fairness research.
- We provide actionable recommendations to remedy existing limitations and pave a path forward towards more thoughtful and nuanced data practices in fair ML.

We start by outlining our selection and annotation process of fairness datasets and publications in Section 2. In Section 3, we contrast the availability and usage of protected attributes in fairness data with the salience of protected attributes in legislation across the globe. In Section 4, we demonstrate exclusionary data practices against minorities with a case study on COMPAS data. In Section 5, we focus on transparency and generalization, showing opaque design decisions affecting fairness evaluations with a second case study on the Bank dataset. We summarize our findings in Section 6, providing a list of recommendations towards better data practices in Section 7, and concluding remarks in Section 8.

2 METHODOLOGY

For this work, we collected and annotated tabular dataset usage for fair classification tasks. To create this corpus, we built on top of a comprehensive survey of fairness datasets [42], leveraging the same inclusion criteria for publications. We focus on tabular datasets and fair classification for their prominent role in the fairness literature [42, 43, 68]. We study the use of tabular datasets ($N = 36$) across 142 articles. Since many datasets appear in multiple publications and most publications use multiple datasets, the total number of dataset and publication combinations annotated was $N = 280$.

Information regarding the usage of different datasets was collected for each combination of dataset and publication. This information includes which variant of a dataset was used, which attributes were considered protected and whether sufficient information was available to reconstruct this, as well as the target variable and features used for prediction. To collect this information, the publications, their supplementary materials, and appendices were consulted for information regarding each dataset usage. Moreover, each publication was searched for mentions of source code; if unsuccessful, we searched on the internet for code repositories mentioning the publication's title. Detailed information on the annotation process and corpus selection is available in Appendix A.

The collected data on dataset usage as well as the code for all analyses presented in this work are publicly available at https://github.com/reliable-ai/lazy-data-practices. Analyses were conducted and visualizations created using Python version 3.9 [104], R version 4.2.2 [81] and RawGraphs version 2.0 [67].

3 NEGLECTED IDENTITIES

Acknowledging the diversity of vulnerability in fair ML is critical as the social impacts of prediction algorithms and the effectiveness of bias mitigation strategies can vary greatly between different protected groups. Vulnerable identities will not benefit from fairness research unless explicitly considered by it. This section studies the availability and usage of protected attributes in fair ML, which we introduce in the following subsections and summarize in Figure 1.

3.1 Protected Attributes Globally

To define protected attributes, we draw from domain-specific legislation and human rights law. We define as *protected* all socially salient attributes explicitly mentioned as prohibited drivers of discrimination and inequality. For example, Article 2 of the Universal Declaration of Human Rights states "Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status" [95].

On the one hand, we try to mitigate the *Global North bias* in AI ethics research [76, 83, 86] by covering international human rights instruments from around the globe, including the Universal Declaration of Human Rights [95], the African Charter on Human and Peoples' Rights [77], the Arab Charter on Human Rights [29], the ASEAN Declaration of Human Rights [9], the American Declaration of the Rights and Duties of Man [78], and the Charter of Fundamental Rights of the European Union [38]. On the other hand, we align with this bias, including a regional perspective on anti-discrimination in hiring and lending based on US and EU legislation [25, 40], covering, for example, the Fair Housing Act [96], the Equal Credit Opportunity Act [97], the Racial Equality Directive [27], and the Employment Equality Directive [28]. There are two mutually reinforcing reasons for this, namely the convenient availability of summary articles on the topic and the influence of these regions on anti-discrimination and fairness research.

Table 1: Protected attributes in global anti-discrimination law. Protected attributes are found in international human rights instruments and domain-specific anti-discrimination law. We report a tick (✓) when the literal phrasing (in the original law or in official clarifications) matches the row header. We report the literal phrasing otherwise.

	UN Charter [95]	African Charter [77]	Arab Charter [29]	ASEAN Declaration [9]	American Declaration [78]	EU Charter [38]	US Fair Lending [25]	EU Fair Hiring [40]
Gender and Sexual Identity								
Sex	✓	✓	✓		✓	✓	✓	✓
Sexual orientation						✓	✓	✓
Gender				✓			Gender identity	Gender; gender reassignment
Racial and Ethnic Origin								
Race	✓	✓	✓	✓	✓	✓	✓	Racial origin
Color	✓	✓	✓			✓	✓	
Ethnic origin	Territory to which person belongs	Ethnic group				✓		✓
National origin	✓	✓	✓	✓		Nationality	✓	
Language	✓	✓	✓	✓	✓	✓		
National minority						✓		
Socioeconomic Status								
Social origin	✓	✓	✓	✓		✓		
Property	✓	Fortune	Wealth	Economic status		✓		
Recipient of public assistance							✓	
Religion, Belief and Opinion								
Religion	✓	✓	Religious belief	✓	Creed	Religion or belief	✓	Religion or belief
Political opinion	✓	✓		✓		✓		
Other opinion	✓	✓	Opinion; thought	✓		✓		
Family								
Birth	Birth status	Birth status	✓	✓		✓		
Familial status							✓	
Marital status							✓	
Disability and Health Conditions								
Disability			✓	✓		✓	✓	✓
Genetic features						✓		
Age								
Age				✓		✓	✓	✓

Drawing from this literature, we provide a shallow categorization of protected attributes, reported in Table 1. We identify seven main categories for protected attributes: (1) gender and sexual identity, (2) racial and ethnic origin, (3) socioeconomic status, (4) religion, belief and opinion, (5) family, (6) disability and health conditions, and (7) age. Most protected attributes fall into at least one of these categories. We categorize attributes potentially relevant to more than one category, such as "genetic features", based on specialized literature [31]. It is worth noting **this is not a complete categorization** of all protected attributes around the globe and across sectors.[1] This categorization aims to guide an inclusive discussion of algorithmic fairness research through the lens of protected attributes.

3.2 Who is Missing

Incentives against the collection and use of protected data are well documented in the literature [5], motivating the line of work on fairness under unawareness [25, 41], which aims to measure and improve fairness with no access to protected attributes. In this section, we demonstrate that this effect is not uniform across all protected attributes. The left bar chart in Figure 1 depicts protected attributes available in popular fairness datasets. Attributes about *religion, belief and opinion* are entirely missing. Variables describing *disability and health conditions* are very infrequent ($n = 3$) and never used in the surveyed literature (right bar chart in Figure 1). *Socioeconomic status* descriptors are more commonly available yet frequently neglected.[2]

Some protected attributes are particularly sensitive and safeguarded by data protection law. The GDPR (General Data Protection Regulation [39]) bans the use of special categories of personal data, including religion and health data, making it more difficult to collect and use these data to audit or train algorithmic systems [102]. The Americans with Disabilities Act [100] imposes strict regulations to disability-related questions that employers can ask [99]. Data protection, however, does not fully explain the availability and usage of protected attributes in fairness research. In the following, we detail the causes and effects of neglecting protected identities.

[1] For example, veteran status does not appear in Table 1, despite being protected in certain countries and industry sectors. Moreover, we neglected the right to non-discrimination for exercising CCPA rights under the California Consumer Privacy Act [25] since it applies in a single country.

[2] For completeness, we also encountered a small number of protected attributes used in the literature but not referenced in legislation, including employment status, alcohol consumption, neighborhood, body-mass index, and profession.

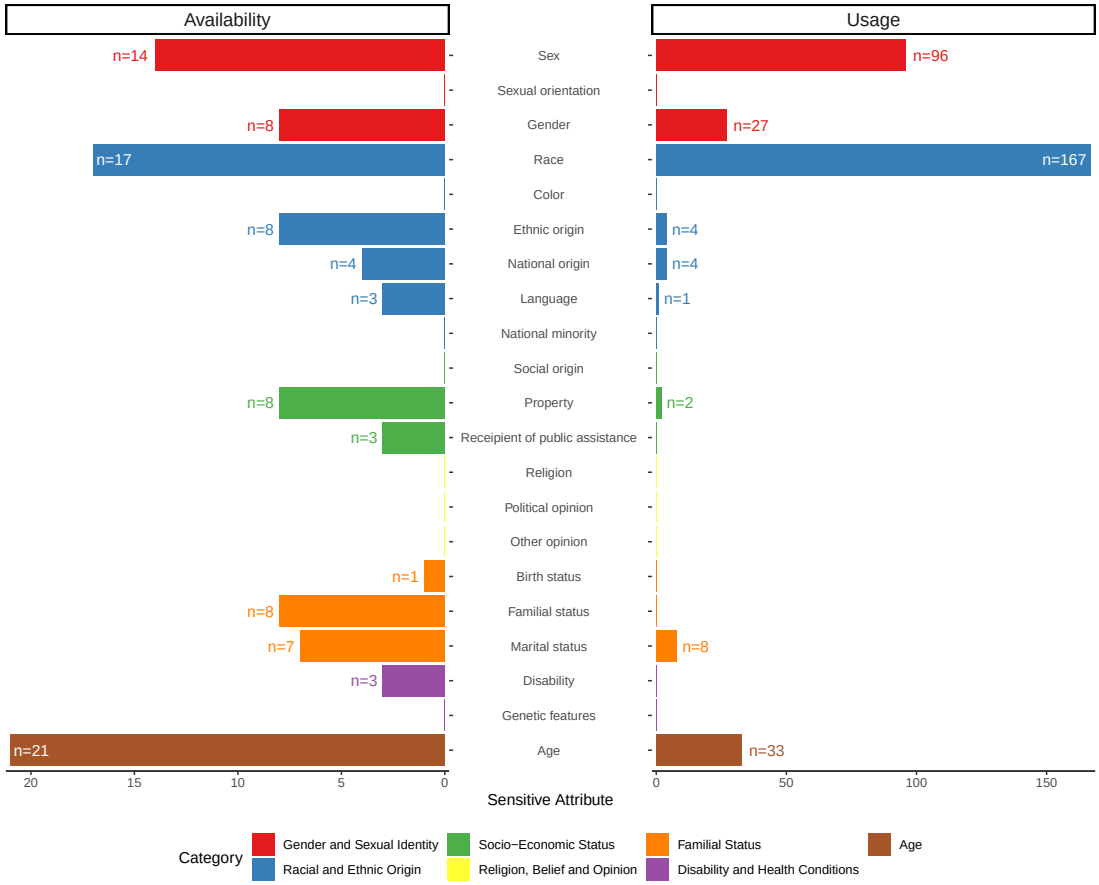


Figure 1: There is a large discrepancy between the list of attributes considered protected under international legislation and their availability or usage in datasets. Bar chart displaying the availability (left) and usage (right) of protected attributes in the literature for all categories of protected attributes in Table 1. Availability based on a total of $N = 36$ datasets; usage based on a total of $N = 233$ experiments with enough information available to reconstruct (or at least make an educated guess about) protected attribute usage (see Section 5 regarding a lack of available information).

Disability is a highly diverse, nuanced, and dynamic construct [93]. Technological ableism is pervasive [87]; algorithmic fairness is insufficient to counter it as it tends to oversimplify and flatten disability. Indeed, there have been multiple calls to move beyond simplistic notions of fairness and towards disability justice [11, 92]. However, this fundamental recognition of nuance may act as a double-edged sword. Even in specific contexts where disability can be treated more narrowly, such as speech recognition for people with speech disorders, data is sparsely available [79]. Research highlighting biases across speech impairments [51, 58] has not gained traction in algorithmic fairness venues [20, 94]. Overall, it seems plausible that other protected attributes have been prioritized, to the detriment of disabled identities, due to difficulties in handling a diverse spectrum of conditions, complex data ethics, and concerns of oversimplification. Acknowledging its limitations, we believe that fair ML research can benefit people with disabilities, especially for bias detection and analyses of its root causes.

Religion and creed are protected by all surveyed legislations. They are a strong driver of identity, bias, and prejudice; in the extreme, they can lead to violence [4, 26]. Religion is highly salient in specific contexts, for example materializing as anti-Muslim discrimination in Western societies [2, 3, 45]. Data collection, however, remains contingent on political will [49, 85]. It is often unavailable in census data [54, 101] and laws mandating data collection for anti-discrimination, such as the HMDA (Home Mortgage Disclosure Act

[98]), do not include religion [6]. Indeed the effectiveness of Western anti-discrimination law in protecting religious minorities such as Muslim identities has been called into question [15]. Negative stereotypes of Muslims have been documented in different regions of the world [17, 88, 103]. While fairness research has been able to study Muslim bias in language models [2, 32, 72], so far it has neglected allocative harms against Muslim people. It could be argued that a lack of focus on religion is compensated by research on racial and ethnic discrimination, since religions have strong ethnic foundations, and congregations tend to be racially homogenous [24, 62]. However, religious and ethnic discrimination can compound rather than simply overlap [33]. Moreover, racial classifications are insufficient for Middle Eastern and North African people, who are classified as white by the US government [66]. Overall, fairness research has neglected this important axis of discrimination and its intersections with other vulnerable identities [45, 73, 85].

Property. High-tech tools can disempower poor people [37, 63]. Stakeholders of child protection systems are concerned about models automating biases against the poor [91]. Overall, poverty shows mutually reinforcing negative effects on health, education, and justice [47, 64, 80, 82]. Despite this fact, property and other socioeconomic variables are seldom used as protected attributes in algorithmic fairness research. This is partly due to data availability: poverty data from household surveys is coarse and sometimes unavailable, especially in the developing world [74]. In addition, and perhaps to a greater extent, it is due to data usage. Wealth is often the target variable of models, such as algorithmic social policies [55, 74], or one of their (unprotected) input features, as in creditworthiness estimators [30]. This seems especially true in fairness research, where the most popular task is income prediction with the Adult dataset [42]. Among formally protected attributes, property is uniquely associated with a perception of mutability and merit: people tend to associate wealth and poverty with individual merit rather than structural constraints [18, 57]. This perception fuels the discourse on deservingness, seeking to distinguish between deserving and undeserving poor people, which determines the boundaries of admissible redistribution policies [8, 106]. In turn, this impacts algorithmic fairness research, not only discouraging bias mitigation based on wealth, but also constraining measurement along this protected axis.

This section highlights blindspots in fairness research, neglecting vulnerable and globally salient identities. It is worth noting that this trend extends to fairness research more broadly, including qualitative studies, and to more protected attributes, including sexual orientation. As a prevalent practice in the field, it has a tendency to self-reinforcement, further incentivizing future research to conform. Indeed recent articles published at fairness conferences, such as *FAccT* (the ACM Conference on Fairness, Accountability, and Transparency) and *AIES* (the AAAI/ACM Conference on Artificial Intelligence, Ethics and Society), mention race and gender more frequently (by one order of magnitude) than religion, disability, socioeconomic status, and sexual orientation [14]. Taking stock of a complex social, legal, and technical landscape, we argue for a move towards an ambitious research roadmap to tackle this complexity (as advocated, for example, in Guo et al. [53]); avoiding it will only prevent us from noticing and remedying existing harms.

4 OMITTED POPULATIONS

A lack of accurate and proper representation is at the heart of many issues the fairness community tries to address. Oftentimes minority groups are neglected in data, leading to discriminatory behavior of systems leveraging this data [68]. Neglect is nuanced and takes many forms. It can materialize as a lack of consideration for specific protected attributes, as discussed in the previous section. It can also derive from the underrepresentation of certain groups in the population during data collection, who are not easy to reach. As we will demonstrate in this section, the issue of underrepresentation gets exacerbated due to the common practice of excluding information about smaller groups during data processing. This is often done out of convenience, to turn a multi-group problem into a binary one, or in some cases, for privacy reasons. In tabular data, this exclusion can either take the form of outright removal of minority groups from the data or aggregation of multiple minority groups into one big "other" group.

These exclusionary data practices are surprisingly common in the examined literature and even more concerning is that they often apply to protected attributes. As protected attributes are, by definition, linked to vulnerability, this amounts to discarding data for disadvantaged minorities. Normalizing these practices sets a dangerous example and incentive for the adoption of such practices also outside of research within real-world systems, with great potential for harm, especially to the most vulnerable populations.

Case Study: Omitted Identities in COMPAS

To demonstrate this practice, we study the different processing strategies in publications using the COMPAS dataset [7], one of the most popular datasets in the fairness literature [42]. The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system is a risk assessment tool used in the US judicial system. The dataset, distributed under the same acronym, was constructed by ProPublica as part of a publication describing racial biases in the profiling system. It contains risk scores from the system for individuals in Broward County, Florida, US, generated during 2013–14. A datasheet [48] for the COMPAS dataset is available in the Appendix of Fabris et al. [42]. The attribute typically considered protected is *race* with a total of 6 categories: "African-American", "Asian", "Caucasian", "Hispanic", "Native American" and "Other".

Overall, we annotate $N = 69$ publications using the COMPAS dataset, with 85.5% (59) providing enough information to reconstruct whether and how the *race* attribute was processed. Although some publications considered additional attributes to be protected, we did not systematically annotate processing of other protected attributes. We identify a total of 8 different processing strategies with the frequency of their occurrence shown in Figure 2A. We sort processing strategies into three categories: (1) *none* if all data was retained as-is, (2) *aggregating* if all observations were retained, but subgroups were recoded and aggregated e.g. collapsing data into "African-American" and "Other", and (3) *filtering* if observations were discarded rather than recoded or aggregated, e.g. keeping only the groups "African-American" and "Caucasian" (the most common form of processing). We do not observe a combination of aggregating and filtering, although such a strategy could easily be

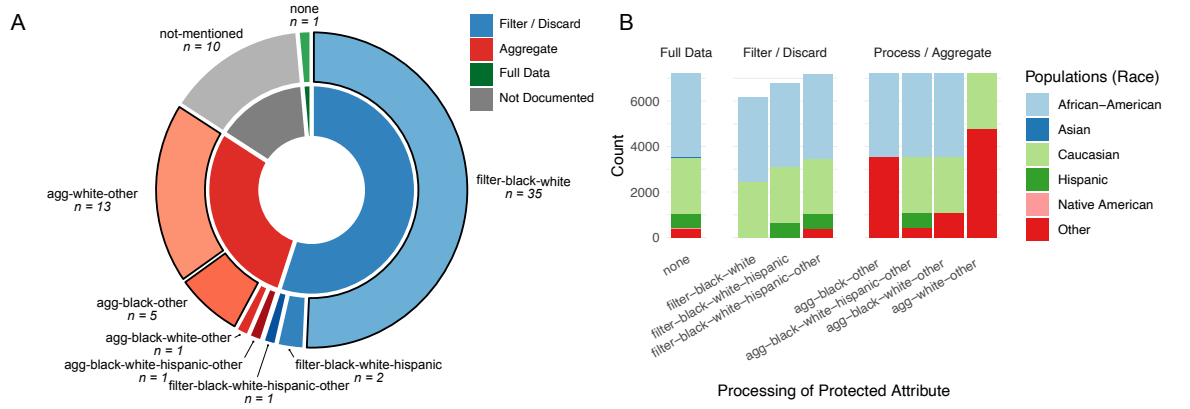


Figure 2: Data from smaller populations is almost *always* either discarded or aggregated within the annotated literature. (A) Prevalence of processing strategies for the COMPAS dataset within the annotated literature and (B) resulting base rates of the protected attribute from these different processing strategies. Due to the small sample sizes, the populations of Asians and Native Americans are difficult / impossible to see in the figure. Neither group is included as a category in any of the processing strategies except when using the *Full Data* ($n = 1$). Processing strategies binarising protected attributes (i.e. leaving a binary variable with only two groups) are highlighted with a black outline in A. The inner circle corresponds to the combined prevalence of processing strategies using a specific approach (e.g. filtering or aggregation).

conceived. Examining Figure 2A, we see that only a single publication examined the full data as-is. The overwhelming majority of publications either filter/discard (38) or aggregate (20) populations. The most extreme processing strategies, leaving only two groups, are the most common (53).

To highlight how processing strategies affect data, we apply each processing strategy on the COMPAS dataset and show the distribution of the resulting *race* attribute in Figure 2B. While we compare all processing strategies on the same version of the COMPAS dataset (compas-scores-two-years.csv), we observe different publications using different versions of the dataset. Figure 2 demonstrates how different strategies for data processing alter the composition and distribution of protected attributes. Many of the strategies leave only two groups, either discarding or aggregating minority groups; none of the actual processing strategies retain Asian or Native American populations as distinct groups. In general, few papers describe, and even fewer justify their choices when handling protected attributes [1].

This fact shows a tendency to simplification and binarization in fair ML empirical research, which seems at odds with the importance of diversity and socio-technical context broadly acknowledged in this field. We speculate that this is partly driven by methodological advances which are more practical under binary protected attributes, and partly by a tendency to algorithmic benchmarking, which is more straightforward in the binary setting. Binarization as an implicit norm in the literature sets a dangerous precedent for research and practice in the field. As a consequence, we see a risk of omission disproportionately affecting vulnerable minorities. Besides the dangerous precedent of normalizing the exclusion of vulnerable subgroups from the data, this also threatens the transparency and reproducibility of fairness research; Figure 2A demonstrates a large share of publications without enough information to reconstruct processing decisions. It is worth noting that, while different publications use different versions of the dataset, this section focuses on a single dataset for comparability and simplicity. Our results, therefore, give a lower bound on data processing variation. As the next section shows, these opaque and diverse choices can lead to very different outcomes during model evaluation and comparison.

5 OPAQUE PREPROCESSING

The previous section describes disparate practices for protected attribute processing that are often overlooked. This section discusses a broader lack of documentation on dataset usage and its consequences. This is a significant risk to the reproducibility and generalization of fairness research for a combination of two reasons: (1) many publications do not document their usage of a dataset sufficiently, assuming that merely the name of a dataset clearly identifies its usage and (2) publications that do document data usage or offer reproducible code vary greatly in their usage, disproving the idea that merely identifying a dataset by its name is sufficient information. These variations in usage, or preprocessing, are likely to affect fairness [70, 89]. Beyond the variation in the mere *usage* and processing of a dataset, we also observe many publications using different *variants* or versions of datasets, sometimes from the same official source and sometimes from undocumented sources. These variants often lack information regarding the processing that happened to create them.

For each dataset-publication combination experimenting with a prediction task ($N = 262$),[3] we annotated the level of documentation, including whether a publication included enough information to reconstruct dataset usage. In particular, we annotated the level of information regarding (1) the target variable that was being predicted y, (2) the features used for classification X, and (3) the protected attributes S. We graded each publication for each aspect into one of three levels: *Yes*, if there was sufficient information, *Guessable* if someone familiar with the dataset could reasonably make an educated guess, and *No* if there was insufficient information or none at all provided. For each publication, we looked for information in the main publication, the supplementary materials, and the source code. We annotated the availability of source code for every dataset-publication pair ($N = 280$). As source code was often not directly referenced in publications, we also searched for it explicitly for every annotated experiment. If source code was available with a certain publication but did not match the publication's analyses, we discarded it as *Not Available*. An example of this are articles presenting new methodologies and experiments, which provide an implementation of the new method but no code reproducing their experiments.

The resulting annotations are summarized in Figure 3, showing that the provided information was insufficient to reconstruct the target variable for 16% (41 out of 262) of annotated experiments and 9% (23) of experiments were lacking information regarding protected attributes. Regarding features, the situation is even worse, with *half* of the annotated experiments (132) containing either not enough information (98) or forcing one to guess (34) to reconstruct feature usage. As publications themselves seldom provide sufficient information to reconstruct dataset usage, this issue is also largely due to a lack of available source code, with just 39% (108 out of 280) of publications providing source code for their analyses. This lack of documentation is problematic for both the reproducibility of research and the generalization of findings in the field, as we will demonstrate in the following.

It is worth noting that proper documentation of preprocessing choices is not sufficient on its own. For example, 10 out of 22 publications using the "German Credit" dataset report extracting *gender* or *sex* information from the data. This is based on the widespread misbelief that this information can be extracted from a column in the dataset, when in fact the necessary information is not available [52]. Nonetheless, having this information explicitly available in the respective publications allows readers to evaluate essential aspects of their correctness and quality.

Case Study: Opaque Preprocessing of Bank

We demonstrate the extent and impact of the variation in dataset usage using the "Bank Marketing" dataset [69] (from here onwards: Bank). This dataset is quite relevant in fairness research (fifth most popular [42]) yet understudied in the literature. Bank describes telemarketing of long-term deposits at a Portuguese bank in the late 2010s. Instances represent telemarketing phone calls and include client-specific features (e.g. job and age), call-specific features (e.g. duration), and environmental features (e.g. euribor). The associated task is to predict whether clients subscribed to a term deposit after the call.

Disparate Preprocessing Choices. We compiled a short list of structured preprocessing choices for Bank across 9 scholarly articles in our corpus focusing on dataset version and protected attributes. First, we note which version of the dataset was used, as there are a total of four different versions available in the original source, two of which have been used in our corpus: *bank-full* and *bank-additional-full*, with the version marked as *additional* containing additional variables, but having slightly fewer observations than the other version. Second, we examine which attributes were considered protected, and third, how they were processed.

We find *age*, *job*, and *marital* to be considered protected, with one publication considering both *age* and *job* protected. While most examined publications consider *age* protected, they show variability in its preprocessing. We identify 3 different strategies to turn age into a binary column.[4] Overall, the 9 publications produce 7 distinct combinations of these three choices. An overview of these scenarios, alongside a visualization regarding the prevalence of each choice, is presented in Figure 4. Notice we are not considering additional choices in dataset processing, such as selection of non-protected features (X), thereby providing a lower bound on the variation in the usage of Bank.

Impact of Disparate Preprocessing. As shown in Figure 4, disparate data processing choices translate into variations in the base rates of the protected attributes, shown beside the identifying letter of each scenario. To quantify the impact of this variation on algorithmic fairness, we consider a fair classification task with the different scenarios in Figure 4. For each scenario, we fit and examine multiple models using the state-of-the-art automated machine learning library *autogluon* version 1.0 [35, 36, 50]. A total of $N = 13$ models are considered; 12 correspond to the default model/hyperparameter configurations in *autogluon* plus a logistic regression model, included for its popularity in the literature and its common use in practice. We use the variable y as a target, consider all non-protected columns as features, and evaluate fairness using the protected attributes as processed under each scenario. We evaluate the performance (F1 score) and fairness (equalized odds difference [56]) of each model, averaging across 10 train-test splits. The fairness and performance measures used in this work are defined in Appendix C. The within-scenario variations of both measures are sizeable with an average spread ($\bar{\delta} = mean(max(x) - min(x))$) of $\bar{\delta}_{EOD} = 0.10$ for equalized odds difference and $\bar{\delta}_{F1} = 0.20$ for F1 score across all scenarios and splits.

Within each scenario, we rank models based on their performance as well as their fairness scores, mimicking a model comparison and selection process based on accuracy and fairness evaluation. We compare model rankings across scenarios to estimate the impact

[3] 18 publications do not fit the typical paradigm of using features to predict a target variable and are therefore omitted. Experiments on synthetically generated datasets are coded as *Not Applicable*.

[4] Strictly speaking there are 4 different strategies, as we observe a single publication processing age as "age < 25 or age > 60" as opposed to "age >= 25 and age < 60" which was observed in two other publications. As these two strategies are equivalent to each other except in how they encode individuals who are exactly of age 60, we combine them under the more popular choice. Moreover, one publication does not mention processing the protected attribute, in which case we also use the most popular processing strategy, as keeping age unprocessed would have been unrealistic.

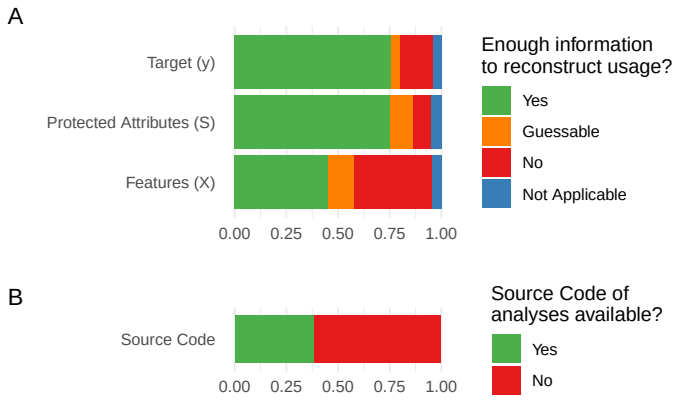


Figure 3: A large section of the annotated literature lacks sufficient information to reproduce analyses. Bar diagrams showing whether publications in the annotated literature contain (A) sufficient information to reconstruct usage of the predicted target variables y, the protected features S and the features used for prediction X and (B) source code to reproduce analyses. Only publications containing a prediction task are included in the figure.

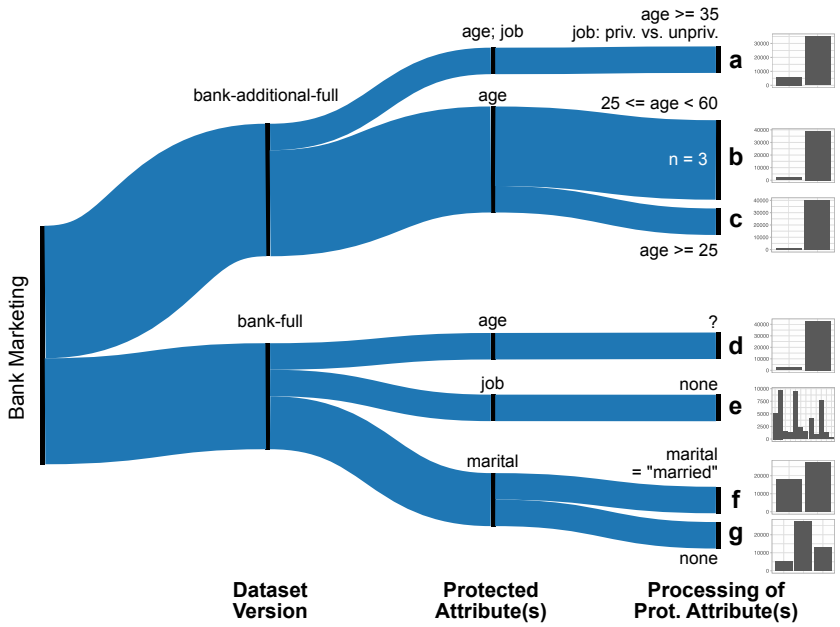


Figure 4: The "same" dataset is used in many different ways within the literature. Sankey diagram illustrating the usage of the Bank dataset within the annotated literature. Each split corresponds to a choice where differences were observed in the literature. Each unique combination of choices or scenario is identified by a unique letter, with the base rates of the protected attribute(s) displayed on the right. We constructed this figure to provide a conservative, lower-bound estimate regarding the variation in dataset usage.

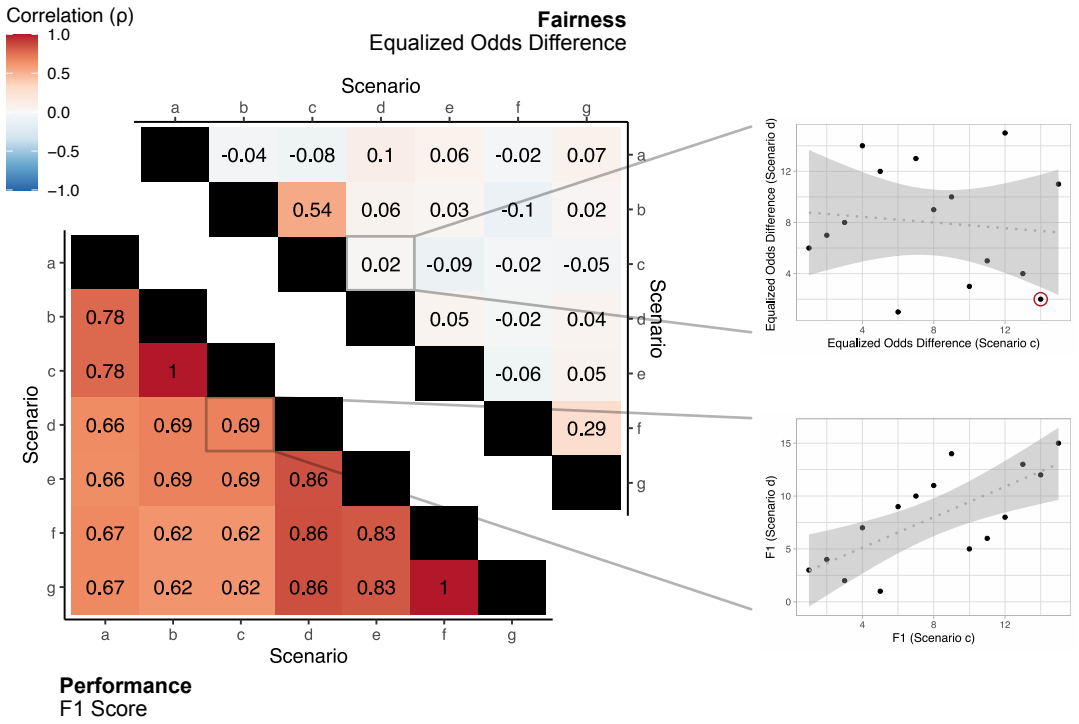


Figure 5: While a practitioner would choose roughly similar models based on performance across the different scenarios, they would choose very different ones based on fairness. Spearman's ρ correlations of model ranks on a measure of fairness (Equalized Odds Difference) and performance (F1 score) between different scenarios. Letters correspond to scenarios described in Figure 4.

of data processing choices. We compute Spearman rank correlations (ρ) on these rankings, reporting the full correlation matrices in Figure 5.

Correlations are high for performance measures (F1 score), with a mean of $\bar{\rho}_{F_1} = 0.747$. This means that model comparison and selection based on performance is stable and generalizes across different scenarios. When examining correlations based on fairness, we observe significantly lower and much more variable (sometimes even negative) correlations, with a mean $\bar{\rho}_{EOD} = 0.04$. This finding suggests that model comparisons based on equalized odds are strongly dependent on different data processing scenarios. The plots on the right in Figure 5 exemplify this fact, depicting model comparisons for a single run of the analysis under scenarios c and d based on F1 score (bottom) and equalized odds difference (top). A rank correlation close to zero for fairness-based rankings entails that the fairest model in scenario c may be among the least fair in scenario d. For example, the second-best model for equalized odds under scenario c (highlighted in red) is the second-worst performer under scenario d. Comparing model fairness under different data processing scenarios yields completely different results. Additional results of this analysis can be found in Appendix C, including correlation matrices using balanced accuracy and demographic parity [21] (Figure 8).

Additionally, we extend our analysis to algorithms designed specifically for fair ML by training and evaluating the methods in Friedler et al. [46] on the Bank data from each scenario. We used the exact same list of algorithms as the original work [22, 44, 46, 60, 107]. This experiment, reported in Appendix C (Figure 9), confirms the instability of fairness-based model comparisons under these preprocessing choices. Overall, the results demonstrate how variability in dataset usage translates into variability of fairness scores; fairness-aware experiments would choose very different models based on the different scenarios, despite working with the "same" Bank dataset.

6 DISCUSSION

In the present article, we demonstrate how common choices in algorithmic fairness datasets harm the quality and curb the impact of fair ML research. We identify multiple worrying aspects regarding prevalent data practices in the literature. First, we notice that **several protected attributes are neglected** (Section 3). This problem is partly due to privacy concerns and is exacerbated by how datasets are used in practice, with many publications focusing on a small fraction of protected attributes while relying on an even smaller number of datasets.

Moreover, we find that **smaller subpopulations are often excluded from analyses** (Section 4), either by aggregating all subpopulations into a single "Other" group or by just outright dropping their data. Therefore, rare identities, such as religious minorities or people with uncommon disabilities, have a double risk of being neglected: important protected attributes are often unavailable, and when they are, small minorities can be filtered out or aggregated for convenience. This is an exclusionary practice that fair ML work should not normalize, but rather counter. Ultimately, misrepresentation of minorities and careless processing choices have been identified as sources of biases in the first place [84], and thus represent practices that should not be reproduced by fairness research itself. We further note that neglecting minorities limits research on intersectionality as the identification of intersectional subgroups depends on the presence of (all) interacting attributes and their sufficient representation in data.

Last, we observe a large amount of variation in the practical usage of datasets which leads to very different model comparisons based on fairness properties. Paired with the lack of proper documentation, this poses a **threat to the reproducibility and generalization of experimental results** (Section 5), potentially misleading practitioners during model evaluation and selection.

Limitations. There are certain limitations to our results. First, work reflecting on the practices of the algorithmic fairness community should also study the industry perspective. This article focuses on fairness research since we were unable to conduct practitioner interviews or otherwise evaluate common practices in the industry. Although research differs significantly from industrial contexts, it certainly influences the prevalent methodologies and best practices in the field. Second, this work studies tabular datasets used for fair classification. We expect minor differences in the usage and availability of protected attributes in other data modalities and tasks, including e.g. the availability of skin type annotations in vision datasets [19]. Moreover, this work focuses on the corpus of publications studied in Fabris et al. [42], containing articles published up to and including 2021. While rather unlikely, data practices in the field may have significantly changed. We examine the robustness of our findings in Appendix B by considering manuscripts covering different fair ML tasks and data modalities published in 2023. Our results indicate that the analyzed data practices largely remain the same, with the exception of the recently introduced and rapidly adopted Folktables datasets [34].

7 RECOMMENDATIONS

The present results remain relevant and warrant addressing; we propose the following recommendations.

Careful inclusion of missing protected attributes in the data. Attributes such as religion and disability are uncommon in fairness research and, more broadly, in machine learning datasets. Strong incentives against their collection include concerns about privacy and consent. We call for dedicated initiatives, including data donation campaigns and citizen science initiatives, capable of filling this gap and responsibly handling the collected data [13]. Targeted data collection initiatives are certainly difficult to undertake, as they require ethical reviews, advertisement through trusted parties, meaningful consent elicitation, and proper data infrastructures with permission systems. By making this gap more visible, we hope to incentivize new work in this direction, including methods to build semi-synthetic datasets that can be used for fairness research without compromising sensitive information of data subjects [12, 90].

Handling multiple small subgroups. Discarding or aggregating data from protected subpopulations is a practice with a high potential for harm that should be countered, rather than normalized, especially by the fair ML community. If real-world data is complex, featuring multiple protected groups with skewed distributions, such complexity should be acknowledged and addressed directly. Pretending that these challenges do not exist by artificially making problems binary, harms the omitted populations immediately, as they are neglected in the present analysis, and in the long term by legitimizing exclusionary practices. First, we call for more explicit discussion about the practicality of proposed approaches beyond binary settings, as with works on intersectionality and rich subgroup fairness [61, 105]. Authors should be explicit (and reviewers demanding) about the applicability of techniques allegedly presented under a binary framing for "notational convenience". Second, the fact that omitted groups are always smaller points to an (often implicit) concern about the significance and stability of groupwise differences. Disaggregated analyses can be unstable for small groups; there is no easy way around this. We advocate the development of nuanced fairness evaluations for disaggregated analyses over small groups; such measures should convey information on uncertainty akin to confidence intervals and describe the statistical significance of differences.

Transparent data usage. Silent subgroup omission is an example of a broader issue of opaque data processing. We call for reflection and transparency in the usage of datasets. Researchers should clearly document how and why specific datasets are chosen and, even more importantly, how they are used. Publications should document which version of a dataset is used (if there are several) and how exactly the data was processed. If the setting is a prediction task, they should mention which variables were predicted, which features were used for prediction, and which attributes were considered protected. Authors can use appendices and supplementary materials when brevity is important. Ideally, they should also provide the source code of analyses, following best practices regarding reproducibility and open research [71, 75]. In this regard, we recommend including all the code used to preprocess data, even when preprocessed data is cached and made available, as it can be hard to reconstruct the origin of the data.

8 CONCLUSION

In this work, we demonstrated common data practices in algorithmic fairness research, including the unavailability of certain protected attributes, the frequent omission of minority groups, and the lack of documentation about preprocessing choices that influence fairness evaluations despite being overlooked. These practices harm fairness research by neglecting vulnerable identities, leading to undetected harms, and by threatening the reproducibility and generalization of findings. They are currently normalized in the literature, where they set a dangerous precedent unless countered with thoughtful data choices. Data is at the core of this field; we

hope the issues raised here will lead to better usage of existing datasets and inspire the careful curation of new resources.

RESEARCH ETHICS AND SOCIAL IMPACT

Ethics Statement

Our analyses hinge on a specific type of social data summarizing scholarly publications. In this context, authors of articles are data subjects whose interests should be considered and balanced against the need to keep community data practices in check. We believe that scientific critique of publicly available works is legitimate and that *negative citations* are unlikely to have a sizable effect on the popularity of an article [23] and the livelihood of its authors. Despite these facts, we decided that criticism of individual manuscripts would not add much utility to our work, while potentially leading to (limited) negative consequences for their authors. Therefore, we focused on aggregate analyses of data practices without singling out individual manuscripts.

Positionality Statement

All authors are affiliated with European organizations from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) countries, in line with a documented pattern in this research community [86]. We found this bias especially relevant when sourcing definitions of protected attributes, as we were initially more inclined to consult resources representing European and North American points of view. We tried to mitigate this bias by consulting international human rights declarations and conventions from around the globe, but our background and the prevalent points of view in the research community inevitably influenced this work.

Adverse Impact Statement

Our adverse impact concerns are threefold. First, we would like to reiterate that our categorization of protected attributes in Section 3 is incomplete and partial. We are unaware of other manuscripts providing a list of globally protected attributes and therefore caution readers against considering our work a comprehensive resource on the topic. Second, our call for transparent data usage, in Section 7, implies an additional documentation effort by researchers; we believe this individual effort will benefit the research community, leading to more careful and reflective data practices as well as more reliable findings. Third, we highlight the word **careful** in our recommendation to include missing protected attributes: the tension between fairness research and data protection is especially relevant for this problem and requires careful consideration; the former should not carelessly trump the latter.

ACKNOWLEDGMENTS

We would like to thank F. Weber and A. Kreider for their help in the annotation process.

Funding

This work is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research, the Munich Center for Machine Learning and the Federal Statistical Office of Germany. The work by A.F. is supported by the FINDHR project, Horizon Europe grant agreement ID: 101070212 and by the Alexander von Humboldt Foundation.

REFERENCES

[1] Amina A. Abdu, Irene V. Pasquetto, and Abigail Z. Jacobs. 2023. An Empirical Analysis of Racial Categories in the Algorithmic Fairness Literature. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023*. ACM, 1324–1333. https://doi.org/10.1145/3593013.3594083

[2] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. In *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, Marion Fourcade, Benjamin Kuipers, Seth Lazar, and Deirdre K. Mulligan (Eds.). ACM, 298–306. https://doi.org/10.1145/3461702.3462624

[3] Ali M Ahmed. 2010. Muslim discrimination: Evidence from two lost-letter experiments. *Journal of Applied Social Psychology* 40, 4 (2010), 888–898.

[4] Amarnath Amarasingam, Sanober Umar, and Shweta Desai. 2022. "Fight, Die, and If Required Kill": Hindu Nationalism, Misinformation, and Islamophobia in India. *Religions* 13, 5 (2022), 380.

[5] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, Madeleine Clare Elish, William Isaac, and Richard S. Zemel (Eds.). ACM, 249–260. https://doi.org/10.1145/3442188.3445888

[6] McKane Andrus and Sarah Villeneuve. 2022. Demographic-Reliant Algorithmic Fairness: Characterizing the Risks of Demographic Data Collection in the Pursuit of Fairness. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*. ACM, 1709–1721. https://doi.org/10.1145/3531146.3533226

[7] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *ProPublica* (May 2016), 254–264.

[8] Lauren D Applebaum. 2001. The influence of perceived deservingness on policy decisions regarding aid to the poor. *Political psychology* 22, 3 (2001), 419–442.

[9] Association of Southeast Asian Nations. 2012. ASEAN Declaration of Human Rights. https://asean.org/asean-human-rights-declaration/.

[10] 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. https://doi.org/10.48550/arXiv.2106.05498 arXiv:2106.05498 [cs]

[11] Cynthia L. Bennett and Os Keyes. 2020. What is the Point of Fairness? Disability, AI and the Complexity of Justice. *SIGACCESS Access. Comput.* 125, Article 5 (mar 2020), 1 pages. https://doi.org/10.1145/3386296.3386301

[12] Karan Bhanot, Miao Qi, John S Erickson, Isabelle Guyon, and Kristin P Bennett. 2021. The problem of fairness in synthetic healthcare data. *Entropy* 23, 9 (2021), 1165.

[13] Matthew Bietz, Kevin Patrick, and Cinnamon Bloss. 2019. Data donation as a model for citizen science health research. *Citizen Science: Theory and Practice* 4, 1 (2019).

[14] Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. 2022. The Forgotten Margins of AI Ethics. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*. ACM, 948–958. https://doi.org/10.1145/3531146.3533157

[15] Rachel AD Bloul. 2008. Anti-discrimination laws, Islamophobia, and ethnicization of Muslim identities in Europe and Australia. *Journal of Muslim minority affairs* 28, 1 (2008), 7–25.

[16] Danah Boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society* 15, 5 (2012), 662–679.

[17] Christia Spears Brown, Hadeel Ali, Ellen A Stone, and Jennifer A Jewell. 2017. US children's stereotypes and prejudicial attitudes toward Arab Muslims. *Analyses of Social Issues and Public Policy* 17, 1 (2017), 60–83.

[18] Mauricio Bucca. 2016. Merit and blame in unequal societies: Explaining Latin Americans' beliefs about wealth and poverty. *Research in Social Stratification and Mobility* 44 (2016), 98–112.

[19] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[20] Maarten Buyl, Christina Cociancig, Cristina Frattone, and Nele Roekens. 2022. Tackling Algorithmic Disability Discrimination in the Hiring Process: An Ethical,

Legal and Technical Analysis. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*. ACM, 1071–1082. https://doi.org/10.1145/3531146.3533169
[21] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops*. IEEE, 13–18.
[22] Toon Calders and Sicco Verwer. 2010. Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery* 21 (2010), 277–292.
[23] Christian Catalini, Nicola Lacetera, and Alexander Oettl. 2015. The incidence and role of negative citations in science. *Proceedings of the National Academy of Sciences* 112, 45 (2015), 13823–13826.
[24] Mark Chaves. 1998. National Congregations Study. https://web.stanford.edu/group/ssds/dewidocs/icpsr3471/cb3471.pdf.
[25] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, danah boyd and Jamie H. Morgenstern (Eds.). ACM, 339–348. https://doi.org/10.1145/3287560.3287594
[26] Swee Hoon Chuah, Simon Gächter, Robert Hoffmann, and Jonathan HW Tan. 2016. Religion, discrimination and trust across three cultures. *European Economic Review* 90 (2016), 280–301.
[27] Council of the European Union. 2000. Council Directive 2000/43/EC Implementing the Principle of Equal Treatment Between Persons Irrespective of Racial or Ethnic Origin. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32000L0043.
[28] Council of the European Union. 2000. Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32000L0078.
[29] Council of the League of Arab States. 2004. Arab Charter on Human Rights.
[30] Sanjiv Das, Richard Stanton, and Nancy Wallace. 2023. Algorithmic Fairness. *Annual Review of Financial Economics* 15 (2023), 565–593.
[31] Aisling De Paor and Delia Ferri. 2015. Regulating genetic discrimination in the European Union. *Eur. JL Reform* 17 (2015), 14.
[32] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, Madeleine Clare Elish, William Isaac, and Richard S. Zemel (Eds.). ACM, 862–872. https://doi.org/10.1145/3442188.3445924
[33] Valentina Di Stasio, Bram Lancee, Susanne Veit, and Ruta Yemane. 2021. Muslim by default or religious discrimination? Results from a cross-national field experiment on hiring discrimination. *Journal of Ethnic and Migration Studies* 47, 6 (2021), 1305–1326.
[34] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 6478–6490. https://proceedings.neurips.cc/paper/2021/hash/32e54441e6382a7fbacbbbaf3c450059-Abstract.html
[35] Nick Erickson. [n. d.]. Autogluon-Benchmark/V1_results at Master · Innixma/Autogluon-Benchmark. https://github.com/Innixma/autogluon-benchmark/tree/master/v1_results.
[36] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. 2020. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. *arXiv preprint arXiv:2003.06505* (2020).
[37] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.
[38] Euopeam Union. 2000. Charter of Fundamental Rights of the European Union C-364/01. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32000X1218%2801%29.
[39] European Parliament. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). https://eur-lex.europa.eu/eli/reg/2016/679/oj.
[40] Alessandro Fabris, Nina Baranowska, Matthew J Dennis, David Graus, Philipp Hacker, Jorge Saldivar, Frederik Zuiderveen Borgesius, and Asia J Biega. 2024. Fairness and Bias in Algorithmic Hiring: a Multidisciplinary Survey. (2024).
[41] Alessandro Fabris, Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani. 2023. Measuring Fairness Under Unawareness of Sensitive Attributes: A Quantification-Based Approach. *J. Artif. Intell. Res.* 76 (2023), 1117–1180. https://doi.org/10.1613/JAIR.1.14033
[42] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Algorithmic fairness datasets: the story so far. *Data Min. Knowl. Discov.* 36, 6 (2022), 2074–2152. https://doi.org/10.1007/S10618-022-00854-Z
[43] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Tackling Documentation Debt: A Survey on Algorithmic Fairness Datasets. In *Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO 2022, Arlington, VA, USA, October 6-9, 2022*. ACM, 2:1–2:13. https://doi.org/10.1145/3551624.3555286
[44] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.
[45] Mariña Fernández-Reino, Valentina Di Stasio, and Susanne Veit. 2023. Discrimination unveiled: a field experiment on the barriers faced by Muslim women in Germany, the Netherlands, and Spain. *European Sociological Review* 39, 3 (2023), 479–497.
[46] 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.
[47] Thomas E Fuller-Rowell, Gary W Evans, and Anthony D Ong. 2012. Poverty and health: The mediating role of perceived discrimination. *Psychological science* 23, 7 (2012), 734–739.
[48] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.
[49] Sonia Ghumman, Ann Marie Ryan, Lizabeth A Barclay, and Karen S Markel. 2013. Religious discrimination in the workplace: A review and examination of current and future trends. *Journal of Business and Psychology* 28 (2013), 439–454.
[50] Pieter Gijsbers, Marcos L. P. Bueno, Stefan Coors, Erin LeDell, Sébastien Poirier, Janek Thomas, Bernd Bischl, and Joaquin Vanschoren. 2023. AMLB: an AutoML Benchmark. arXiv:2207.12560 [cs.LG]
[51] Jordan R Green, Robert L MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn A Ladewig, Jimmy Tobin, Michael P Brenner, et al. 2021. Automatic Speech Recognition of Disordered Speech: Personalized Models Outperforming Human Listeners on Short Phrases.. In *Interspeech*. 4778–4782.
[52] Ulrike Groemping. 2019. South german credit data: Correcting a widely used data set. *Rep. Math., Phys. Chem., Berlin, Germany, Tech. Rep* 4 (2019), 2019.
[53] Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hanna M. Wallach, and Meredith Ringel Morris. 2020. Toward fairness in AI for people with disabilities SBG@a research roadmap. *ACM SIGACCESS Access. Comput.* 125 (2020), 2. https://doi.org/10.1145/3386296.3386298
[54] Cristina Gutiérrez Zúñiga and Renée De La Torre Castellanos. 2017. Census data is never enough: How to make visible the religious diversity in Mexico. *Social Compass* 64, 2 (2017), 247–261.
[55] Rema Hanna and Benjamin A Olken. 2018. Universal basic incomes versus targeted transfers: Anti-poverty programs in developing countries. *Journal of Economic Perspectives* 32, 4 (2018), 201–226.
[56] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).
[57] Nicholas Heiserman and Brent Simpson. 2017. Higher inequality increases the gap in the perceived merit of the rich and poor. *Social Psychology Quarterly* 80, 3 (2017), 243–253.
[58] Julio C Hidalgo Lopez, Shelly Sandeep, MaKayla Wright, Grace M Wandell, and Anthony B Law. 2023. Quantifying and Improving the Performance of Speech Recognition Systems on Dysphonic Speech. *Otolaryngology–Head and Neck Surgery* 168, 5 (2023), 1130–1138.
[59] Andrew Iliadis and Federica Russo. 2016. Critical data studies: An introduction. *Big Data & Society* 3, 2 (2016), 2053951716674238.
[60] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*. Springer, 35–50.
[61] Michael J. Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. An Empirical Study of Rich Subgroup Fairness for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, danah boyd and Jamie H. Morgenstern (Eds.). ACM, 100–109. https://doi.org/10.1145/3287560.3287592
[62] Rebecca Y Kim. 2011. Religion and ethnicity: Theoretical connections. *Religions* 2, 3 (2011), 312–329.
[63] Keith Kirkpatrick. 2021. Algorithmic poverty. *Commun. ACM* 64, 10 (2021), 11–12. https://doi.org/10.1145/3479977
[64] Helen F Ladd. 2012. Education and poverty: Confronting the evidence. *Journal of Policy Analysis and Management* 31, 2 (2012), 203–227.
[65] Benjamin Laufer, Sameer Jain, A. Feder Cooper, Jon M. Kleinberg, and Hoda Heidari. 2022. Four Years of FAccT: A Reflexive, Mixed-Methods Analysis of Research Contributions, Shortcomings, and Future Prospects. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of*

Korea, June 21 - 24, 2022. ACM, 401–426. https://doi.org/10.1145/3531146.3533107
[66] Neda Maghbouleh, Ariela Schachter, and René D Flores. 2022. Middle Eastern and North African Americans may not be perceived, nor perceive themselves, to be White. *Proceedings of the National Academy of Sciences* 119, 7 (2022), e2117940119.
[67] Michele Mauri, Tommaso Elli, Giorgio Caviglia, Giorgio Uboldi, and Matteo Azzi. 2017. RAWGraphs: a visualisation platform to create open outputs. In *Proceedings of the 12th biannual conference on Italian SIGCHI chapter*. 1–5.
[68] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6 (2022), 115:1–115:35. https://doi.org/10.1145/3457607
[69] Sérgio Moro, Paulo Cortez, and Paulo Rita. 2014. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems* 62 (2014), 22–31.
[70] Carlos Mougan, José Manuel Álvarez Colmenares, Salvatore Ruggieri, and Steffen Staab. 2023. Fairness Implications of Encoding Protected Categorical Attributes. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2023, Montréal, QC, Canada, August 8-10, 2023*, Francesca Rossi, Sanmay Das, Jenny Davis, Kay Firth-Butterfield, and Alex John (Eds.). ACM, 454–465. https://doi.org/10.1145/3600211.3604657
[71] Marcus R Munafò, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John Ioannidis. 2017. A manifesto for reproducible science. *Nature human behaviour* 1, 1 (2017), 1–9.
[72] Deepa Muralidhar. 2021. Examining Religion Bias in AI Text Generators. In *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, Marion Fourcade, Benjamin Kuipers, Seth Lazar, and Deirdre K. Mulligan (Eds.). ACM, 273–274. https://doi.org/10.1145/3461702.3462469
[73] Kevin L Nadal, Kristin C Davidoff, Lindsey S Davis, Yinglee Wong, David Marshall, and Victoria McKenzie. 2015. A qualitative approach to intersectional microaggressions: Understanding influences of race, ethnicity, gender, sexuality, and religion. *Qualitative psychology* 2, 2 (2015), 147.
[74] Alejandro Noriega-Campero, Bernardo Garcia-Bulle, Luis Fernando Cantu, Michiel A. Bakker, Luis Tejerina, and Alex Pentland. 2020. Algorithmic targeting of social policies: fairness, accuracy, and distributed governance. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna (Eds.). ACM, 241–251. https://doi.org/10.1145/3351095.3375784
[75] Brian A Nosek, George Alter, George C Banks, Denny Borsboom, Sara D Bowman, Steven J Breckler, Stuart Buck, Christopher D Chambers, Gilbert Chin, Garret Christensen, et al. 2015. Promoting an open research culture. *Science* 348, 6242 (2015), 1422–1425.
[76] Chinasa T Okolo, Nicola Dell, and Aditya Vashistha. 2022. Making AI explainable in the global south: A systematic review. In *ACM SIGCAS/SIGCHI Conf. on Computing and Sustainable Societies (COMPASS)*. 439–452.
[77] Organisation of African Unity. 1981. African Charter on Human and Peoples' Rights. https://au.int/sites/default/files/treaties/36390-treaty-0011_-_african_charter_on_human_and_peoples_rights_e.pdf.
[78] Organization of American States. 1948. American Declaration of the Rights and Duties of Man. https://www.oas.org/en/iachr/mandate/Basics/american-declaration-rights-duties-of-man.pdf.
[79] Orestis Papakyriakopoulos, Anna Seo Gyeong Choi, William Thong, Dora Zhao, Jerone Theodore Alexander Andrews, Rebecca Bourke, Alice Xiang, and Allison Koenecke. 2023. Augmented Datasheets for Speech Datasets and Ethical Decision-Making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023*. ACM, 881–904. https://doi.org/10.1145/3593013.3594049
[80] Zachary Parolin and Emma K Lee. 2022. The role of poverty and racial discrimination in exacerbating the health consequences of COVID-19. *The Lancet Regional Health–Americas* 7 (2022).
[81] R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
[82] Bernadette Rabuy and Daniel Kopf. 2016. Detaining the poor: How money bail perpetuates an endless cycle of poverty and jail time. *Prison Policy Initiative* 10 (2016), 1–20.
[83] Cathy Roche, Dave Lewis, and PJ Wall. 2021. Artificial Intelligence Ethics: An Inclusive Global Discourse? *arXiv preprint arXiv:2108.09959* (2021).
[84] Kit T Rodolfa, Pedro Saleiro, and Rayid Ghani. 2020. Bias and fairness. In *Big data and social science*. Chapman and Hall/CRC, 281–312.
[85] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining Algorithmic Fairness in India and Beyond. In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, Madeleine Clare Elish, William Isaac, and Richard S. Zemel (Eds.). ACM, 315–328. https://doi.org/10.1145/3442188.3445896
[86] 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
[87] Ashley Shew. 2020. Ableism, Technoableism, and Future AI. *IEEE Technol. Soc. Mag.* 39, 1 (2020), 40–85. https://doi.org/10.1109/MTS.2020.2967492
[88] John Sides and Kimberly Gross. 2013. Stereotypes of Muslims and Support for the War on Terror. *The Journal of Politics* 75, 3 (2013), 583–598.
[89] Jan Simson, Florian Pfisterer, and Christoph Kern. 2023. Everything, Everywhere All in One Evaluation: Using Multiverse Analysis to Evaluate the Influence of Model Design Decisions on Algorithmic Fairness. *arXiv preprint arXiv:2308.16681* (2023).
[90] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. 2022. Synthetic Data - Anonymisation Groundhog Day. In *31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022*, Kevin R. B. Butler and Kurt Thomas (Eds.). USENIX Association, 1451–1468. https://www.usenix.org/conference/usenixsecurity22/presentation/stadler
[91] Logan Stapleton, Min Hun Lee, Diana Qing, Marya Wright, Alexandra Chouldechova, Ken Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. Imagining new futures beyond predictive systems in child welfare: A qualitative study with impacted stakeholders. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*. ACM, 1162–1177. https://doi.org/10.1145/3531146.3533177
[92] Nicholas Tilmes. 2022. Disability, fairness, and algorithmic bias in AI recruitment. *Ethics and Information Technology* 24, 2 (2022), 21.
[93] Shari Trewin. 2018. AI fairness for people with disabilities: Point of view. *arXiv preprint arXiv:1811.10670* (2018).
[94] Shari Trewin, Sara H. Basson, Michael J. Muller, Stacy M. Branham, Jutta Treviranus, Daniel M. Gruen, Daniel Hebert, Natalia Lyckowski, and Erich Manser. 2019. Considerations for AI fairness for people with disabilities. *AI Matters* 5, 3 (2019), 40–63. https://doi.org/10.1145/3362077.3362086
[95] United Nations. 1948. Universal Declaration of Human Rights. https://www.un.org/en/about-us/universal-declaration-of-human-rights.
[96] United States Congress. 1968. An Act to prescribe penalties for certain acts of violence or intimidation, and for other purposes. https://www.govinfo.gov/content/pkg/COMPS-343/pdf/COMPS-343.pdf.
[97] United States Congress. 1974. Equal Credit Opportunity Act. https://www.govinfo.gov/content/pkg/STATUTE-88/pdf/STATUTE-88-Pg1500.pdf.
[98] United States Congress. 1975. An Act to extend the authority for the flexible regulation of interest rates on deposits and share accounts in depository institutions, to extend the National Commission on Electronic Fund Transfers, and to provide for home mortgage disclosure. https://www.govinfo.gov/content/pkg/STATUTE-89/pdf/STATUTE-89-Pg1124.pdf.
[99] United States Congress. 1990. An Act to establish a clear and comprehensive prohibition of discrimination on the basis of disability. https://www.govinfo.gov/content/pkg/STATUTE-104/pdf/STATUTE-104-Pg327.pdf.
[100] United States: National Archives and Records Administration: Office of the Federal Register and United States: Congress: Senate: Labor and Human Resources. 1990. Americans with Disabilities Act of 1990. Part 1: Public Laws. , 327–378 pages.
[101] US Census Bureau. 2022. Does the Census Bureau have data for religion?
[102] Marvin Van Bekkum and Frederik Zuiderveen Borgesius. 2023. Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception? *Computer Law & Security Review* 48 (2023), 105770.
[103] Margaretha A Van Es. 2019. Muslim women as 'ambassadors' of Islam: Breaking stereotypes in everyday life. *Identities* 26, 4 (2019), 375–392.
[104] Guido Van Rossum, Fred L Drake, et al. 1995. *Python reference manual*. Vol. 111. Centrum voor Wiskunde en Informatica Amsterdam.
[105] Angelina Wang, Vikram V. Ramaswamy, and Olga Russakovsky. 2022. Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*. ACM, 336–349. https://doi.org/10.1145/3531146.3533101
[106] Celeste Watkins-Hayes and Elyse Kovalsky. 2016. The discourse of deservingness. *The Oxford handbook of the social science of poverty* 1 (2016).
[107] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*. PMLR, 962–970.

A ANNOTATIONS

A.1 Corpus selection

The selection criteria for the corpus are the same as in Fabris et al. [42]. The overall scope of considered literature consists of all articles

that were published in either (1) the proceedings of fairness-related conferences such as the ACM Conference on Fairness, Accountability, and Transparency (*FAccT*) and the AAAI/ACM Conference on Artificial Intelligence, Ethics and Society (*AIES*), (2) the proceedings of major machine learning conferences, including the IEEE/CVF Conference on Computer Vision and Pattern Recognition (*CVPR*), the Conference on Neural Information Processing Systems (*NeurIPS*), the International Conference on Machine Learning (*ICML*), the International Conference on Learning Representations (*ICLR*), the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (*KDD*), or (3) the proceedings of any of the "Past Network Events" or "Older Workshops" as listed on the FAccT Network. Works from 2014 up to and including June 2021 were considered (including *FAccT*, *ICLR*, *AIES* and *CVPR* in 2021). This list of literature was narrowed down to fairness-related articles by a manual review, after first filtering for articles which included one of the following substrings in their titles (with * denoting wildcards): *fair*, *bias*, discriminat*, *equal*, *equit*, disparate, *parit*.

A.2 Annotation Process

Annotations were performed by the first author and two research assistants over the course of multiple months. Research assistants were fairly compensated for their work, following university guidelines at 12.00 EUR per hour without an academic degree and 14.00 EUR per hour with a Bachelor's degree. The annotation scheme and process were developed by the authors and research assistants received interactive training on the annotation process.

Annotations were made using Google Sheets using two tables: *Datasets* and *Datasets-x-Papers*, with each annotated column having an explanatory note regarding its annotation scheme. Datasets were randomly assigned to annotators, based on their internal unique identifiers. Dataset-x-Paper combinations were assigned based on assigned datasets, with a subset of them being reassigned based on annotator availability towards the end of the annotation process.

Throughout the annotation process there were weekly meetings to address any difficulties or ambiguities with annotations and the additional option for asynchronous discussion via chat software. Difficult annotations could be marked as requiring additional input. Additionally, annotation quality was checked on face validity by the first author for a subset of annotations.

A.3 Annotation Instructions

Besides in-person training on the annotation process, the following written instructions were made available to annotators:

Tables

- *Datasets*, which contains data on individual datasets, incl. any varieties
- *Datasets-x-Papers*, which contains an entry for every dataset and paper that makes use of said dataset.

Annotation process. Start by annotating the data for a dataset, then annotate the papers that use it. Update the entry of the dataset if changes become necessary. For every column, you can find information on how to annotate it by hovering over its title. Annotate each row from left to right. When you want to put multiple values in a single cell (e.g. multiple column names), separate them with semicolons. When any questions emerge or something is unclear, post in the slack channel. Please always use filter views when performing annotations, to only see the annotations assigned to you.

Standardized Process for Searching relevant sections. When annotating entries in *Datasets-x-Papers*, it's important we do our due diligence in searching for information about how a dataset was used. This is especially important in regards to a paper's code (as code is typically an external ressource, so easier to miss). Please always try at least the following 5 steps when searching information about how a dataset is used. You're also free to try additional ways of finding information about the dataset, but we want to make sure, that at least these steps have been performed for every paper.

Searching for Code

(1) Search for "github" and "gitlab" in the paper.
(2) Search for the paper's name on google. Sometimes there's an external repository with code that uses the paper's name, but is not referenced in the paper.
(3) Check in the official location of the paper whether it has supplementary material e.g. an appendix or zip files. These can contain code or a detailed description of datasets.

Finding relevant sections

(1) Search for the common names of the dataset itself to find information about it (if it has a common name)
(2) Search for "dataset" or "data" to find the relevant sections describing how data is used.

B ROBUSTNESS

In this appendix, we investigate the robustness of Section 3 findings across time, fairness tasks, and beyond tabular datasets. Additionally, we ensure that the tabular datasets we focused on remained central in the literature. Considering the most recent proceedings (2023) of two well-known machine learning and fairness conferences such as *ICML* and *FAccT*, we select all articles whose titles contain the string `fair`. We manually select articles that focus on quantitative analyses of group fairness, without any restriction based on task or data specification. For each of these manuscripts, we annotate dataset and protected attribute usage. Our findings are presented below.

Popular datasets remained popular. Our analysis in Section 3 is based on publications up to 2021, building on top of Fabris et al. [42]. We find that 8 out of 10 most popular datasets remain the same, with the key exception of the recently-introduced Folktables datasets [34] (10 usages), complementing but not *retiring* Adult (13 usages). All such datasets are tabular, confirming the centrality of this data modality in fair ML research.

Neglected identities remain neglected. Figure 6 compares protected attributes in fair ML experiments up to 2021 and in 2023. Although we find isolated experiments on sexual orientation, property, and disability, it is clear that these attributes, as well as religion ($n = 0$), remain understudied, especially in comparison with sex, gender, and race. It is worth noting that we follow the naming of manuscript authors and dataset creators for sex and gender; the drop of the former in favor of the latter is a consequence of this fact and may not reflect an actual focus shift.

Table 2: The usage of datasets remained highly similar in 2023. Usage of datasets in fairness-related articles published at FAccT and ICML 2023 compared to usage within the annotated literature. Only datasets which are used at least twice in 2023 are shown. Datasets are ordered by their usage in 2023.

	2023			Up to 2021		
Dataset Name	Rank	Fraction	N	Rank	Fraction	N
Adult	1	20.3%	13	1	30.0%	84
Folktables *(new dataset)*	2	15.6%	10	-	-	-
COMPAS	3	12.5%	8	2	24.6%	69
Communities; Communities and Crime	4	7.8%	5	4	4.3%	12
German; German Credit; Credit	5	6.2%	4	3	9.3%	26
Law_School	5	6.2%	4	4	4.3%	12
Bank; Bank Marketing; Marketing	7	4.7%	3	6	3.2%	9
default of credit card clients	8	3.1%	2	11	1.4%	4
Student; Student Performance	8	3.1%	2	21	0.4%	1

C ADDENDUM: OPAQUE PREPROCESSING OF BANK

Here we present supplementary figures and information for the analyses in Section 5. The performance metrics used in this work are accuracy (Eq 1), balanced accuracy (Eq 2), and F1 score (Eq 3).

$$\begin{aligned} \text{Precision} &= \Pr(y = 1|\hat{y} = 1) \\ \text{Recall} &= \Pr(\hat{y} = 1|y = 1) \\ \text{Specificity} &= \Pr(\hat{y} = 0|y = 0) \\ \text{Acc} &= \Pr(\hat{y} = y) \end{aligned} \tag{1}$$

$$\text{bACC} = \frac{\text{Specificity} + \text{Recall}}{2} \tag{2}$$

$$\text{F1 Score} = \frac{2}{\text{Precision}^{-1} + \text{Recall}^{-1}} \tag{3}$$

The fairness metrics used in this work are equalized odds difference (Eq 4), demographic parity difference (Eq 5), and disparate impact (Eq 6).

$$\text{EOD} = \max_g \Pr(\hat{y} = 1|y = 1, S = g) - \min_g \Pr(\hat{y} = 1|y = 1, S = g) \tag{4}$$

$$\text{DPD} = \max_g \Pr(\hat{y} = 1|S = g) - \min_g \Pr(\hat{y} = 1|S = g) \tag{5}$$

$$\text{DI} = \frac{\max_g \Pr(\hat{y} = 1|S = g)}{\min_g \Pr(\hat{y} = 1|S = g)} \tag{6}$$

The overall variation of different metrics for the first experiment in Section 5 is illustrated in Figure 7. As can be seen, there exists ample variation across the different metrics and variation is especially pronounced on metrics of algorithmic fairness.

Figure 8 depicts correlation matrices for the first experiment in Section 5, with different performance and fairness measures, namely *balanced accuracy* and *demographic parity difference*. Although we still note instability in fairness-based model comparison, comparisons based on demographic parity are more stable than for equalized odds difference. We interpret this as a consequence of a classifier's (groupwise) acceptance rate $\Pr(\hat{y} = 1)$ being more stable than its (groupwise) true positive rate $\Pr(\hat{y} = 1|y = 1)$ since the former is computed over all points in the test set, while the latter only on the positives ($y = 1$).

For the second experiment in the section, we aimed to repeat our analysis replicating a highly popular setting. We therefore used the same selection of (mainly) fairness-aware algorithms used in Friedler et al. [46] and applied their methodology on the differently processed versions of the Bank dataset in Figure 4. Specifically, we used the *numeric* variant of their analysis, as it works with a sufficiently large selection of algorithms and does not require the protected attribute to be binary. The correlation matrices for *accuracy* and *disparate impact* across scenarios are depicted in Figure 9. Both metrics were chosen following Friedler et al. [46]. Disparate impact is calculated using a binary version of the protected attribute, split into privileged and unprivileged groups. Using the non-binary, averaged version of disparate impact also discussed in the original paper, lead to similar and even more diverse results.

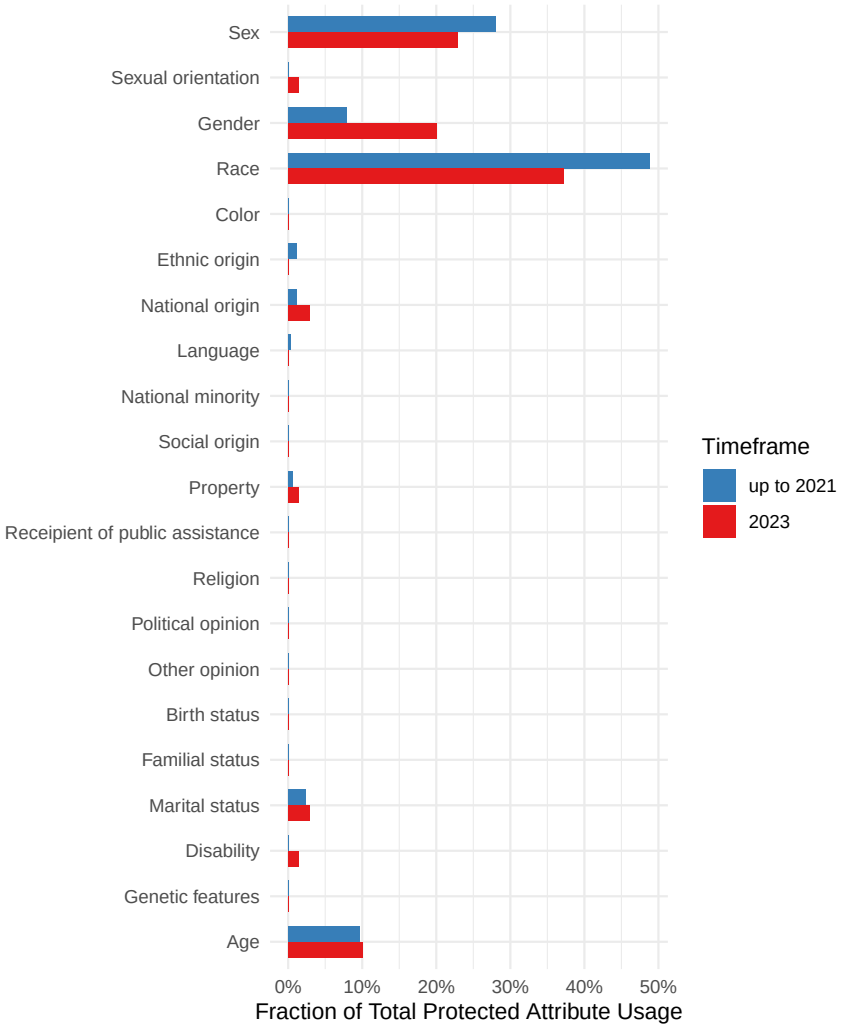


Figure 6: The usage of protected attributes remained similar in 2023. Relative usage of protected attributes in the annotated literature up to 2021 and within the subset of literature we examined in 2023. Usage within the annotated literature corresponds to the right half of Figure 1.

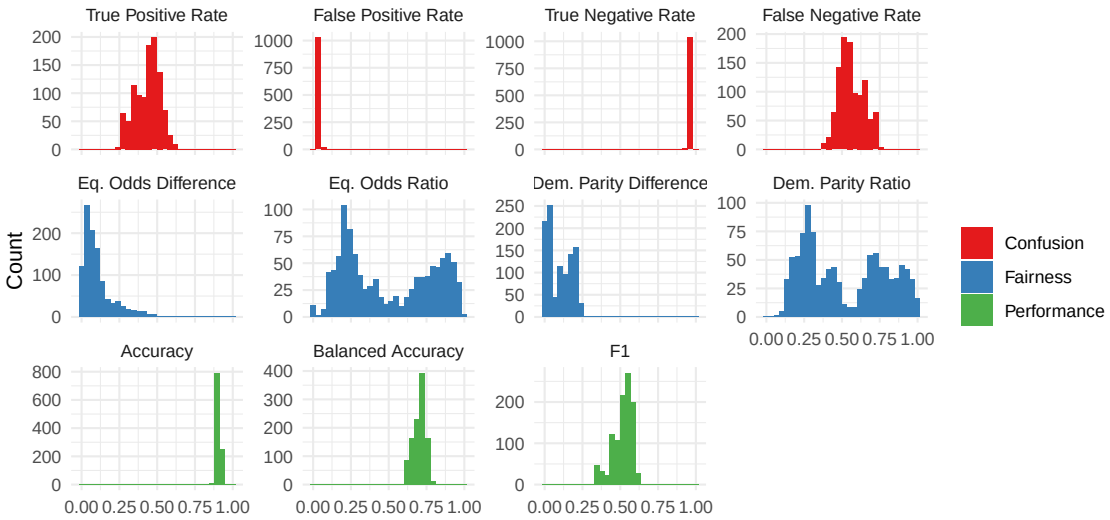


Figure 7: There is a large degree of overall variation, especially on fairness metrics. Histograms displaying the overall variation on different metrics within and across different scenarios and repetitions of the analysis.

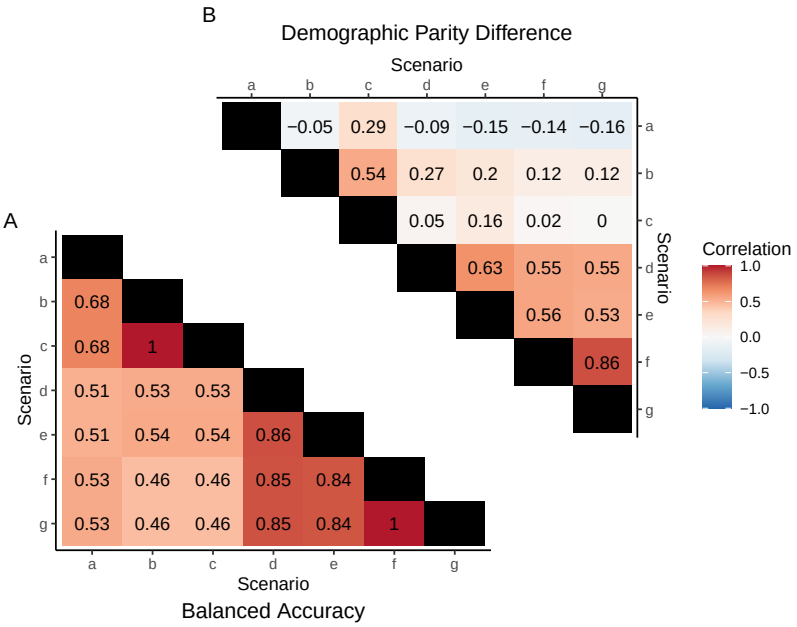


Figure 8: Spearman's ρ correlations of model ranks on (A) Balanced Accuracy and (B) Demographic Parity Difference between different scenarios. Letters correspond to the scenarios described in Figure 4.

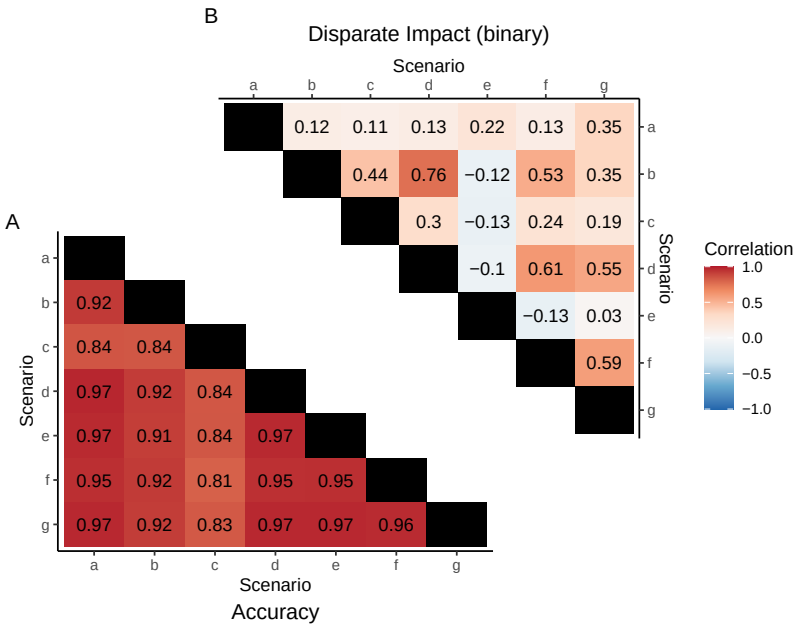


Figure 9: Spearman's ρ correlations of model ranks on (A) Raw Accuracy and (B) Disparate Impact (binary) between different scenarios when reproducing our analysis from Section 5 using an existing selection of fairness-aware algorithms and methodology [46]. Letters correspond to the scenarios described in Figure 4.

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

Contributing article

Simson, J., Pfisterer, F., & Kern, C. (2024). One Model Many Scores: Using Multiverse Analysis to Prevent Fairness Hacking and Evaluate the Influence of Model Design Decisions. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1305–1320. Association for Computing Machinery. doi: 10.1145/3630106.3658974 URL https://doi.org/10.1145/3630106.3658974

Code repository

https://github.com/reliable-ai/fairml-multiverse/

Copyright information

Author contributions

C. Kern provided the initial idea of examining the influence of decisions on machine learning models. F. Pfisterer provided the idea of using a FANOVA to assess the importance of decisions. All authors contributed to the final choice of decisions to evaluate. J. Simson implemented the empirical experiments in the work; analyzed the data and created all figures and tables. J. Simson lead the writing; submission and revision process of the work. J. Simson performed most of the writing of this work, F. Pfisterer contributed to Section 1.3, particularly the last paragraph regarding hyperparameter-optimization; C. Kern particularly contributed to the introduction. All authors contributed through fruitful comments, proofreading and revisions of the manuscript.

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**.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658974

KEYWORDS

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

ACM Reference Format:
Jan Simson, Florian Pfisterer, and Christoph Kern. 2024. One Model Many Scores: Using Multiverse Analysis to Prevent Fairness Hacking and Evaluate the Influence of Model Design Decisions. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3630106.3658974

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

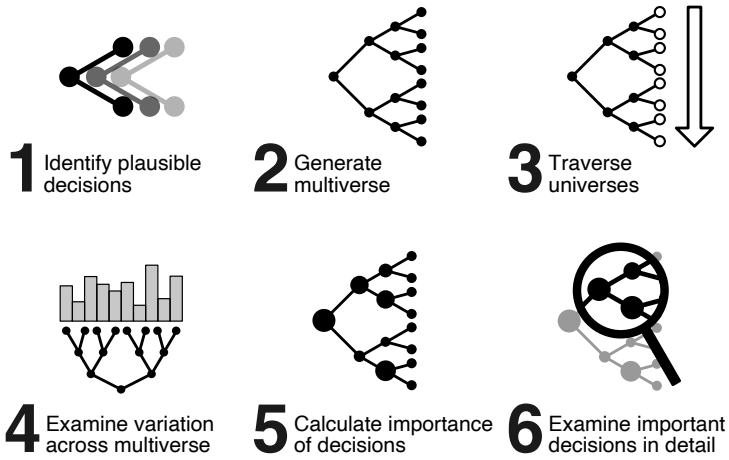


Figure 1: Steps to conduct a multiverse analysis for algorithmic fairness. Steps 1 - 4 apply to multiverse analyses in general, whereas steps 5 - 6 are unique to larger multiverse analyses for algorithmic fairness.

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.

	Decisions and Options Examined in Case Study	
Category	Decision	Options
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.

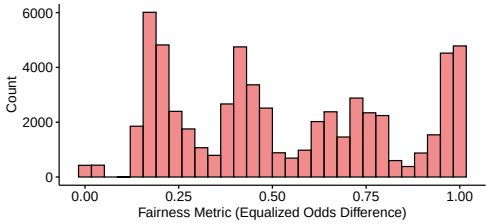


Figure 2: Variation in the multiverse spans the entirety of possible values of the fairness metric. Distribution of fairness metric (equalized odds difference) across universes. Lower values on the fairness metric indicate smaller *TPR* and *FPR* differences across groups.

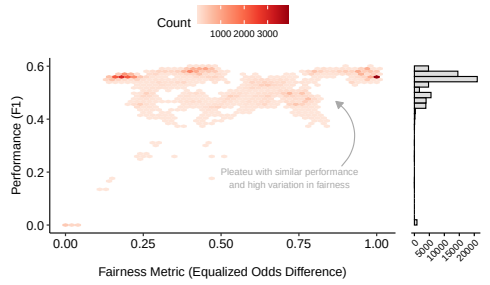


Figure 3: Performance and fairness are largely unrelated with plateaus of low variance in performance, but high variance in fairness. Distribution of overall performance as F_1 score and fairness metric (equalized odds difference) across all multiverses. Marginal histogram shows distribution of performance. 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

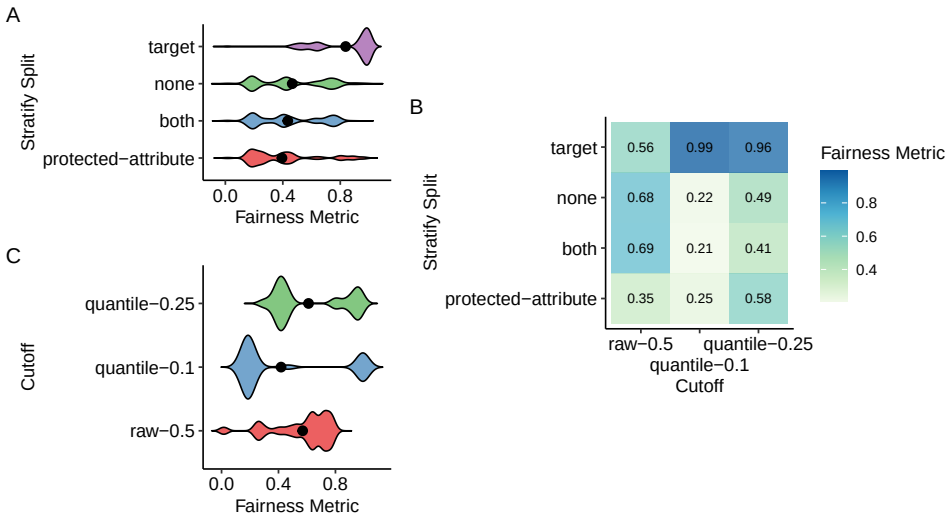


Figure 4: The influence of decisions on the fairness metric can only be understood when examining interactions on top of individual decisions. Visualization of the fairness metric depending on the three most important decision / decision combinations (from A - C by importance) and their respective options.

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.

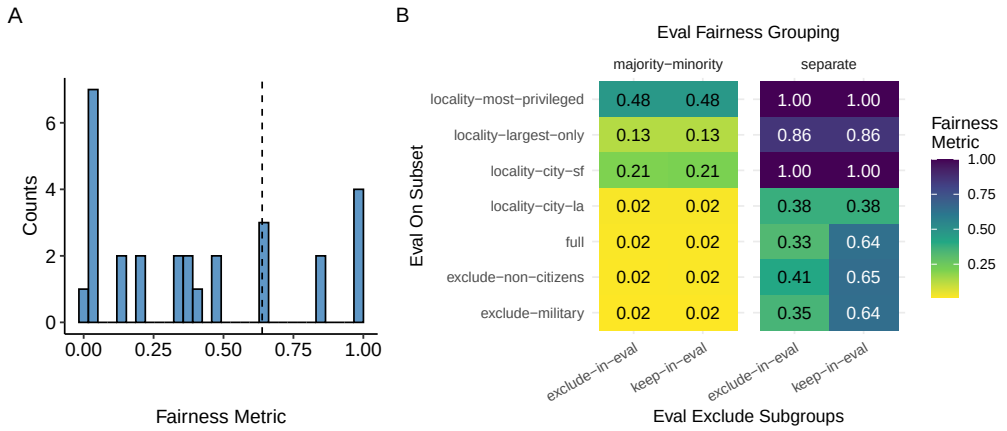


Figure 5: The fairness metric of the exact same model can be significantly altered by varying its evaluation strategy alone (A) and especially the interaction of different evaluation decisions leads to changes in the fairness metric (B). Overall distribution (A) and raw values (B) of fairness metric (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

This work is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research, the Munich Center for Machine Learning and the Federal Statistical Office of Germany.

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, Alejandro Ecker, Achim Edelmann, Maureen A. Eger, Simon Ellerbrock, Anna Forke, Andrea 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

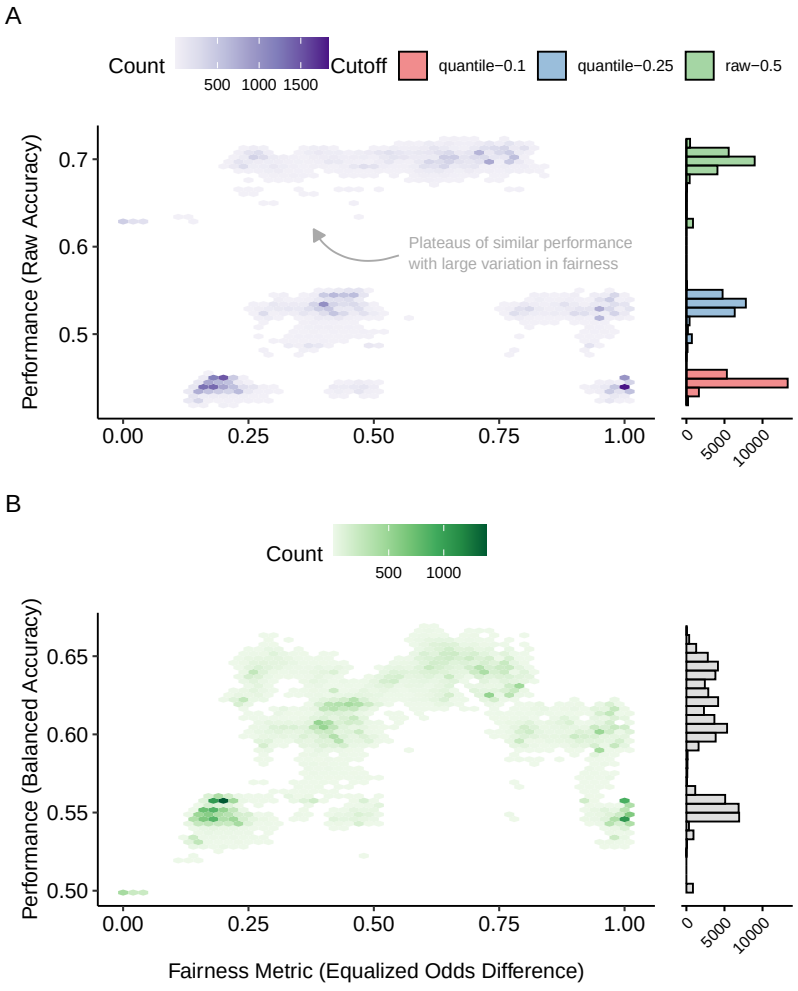


Figure A6: Performance and fairness are largely unrelated with clusters of low variance in performance, but high variance in fairness. Distribution of overall performance as raw (A) or balanced (B) accuracy and fairness metric (equalized odds difference) across all multiverses. Marginal histograms show 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.

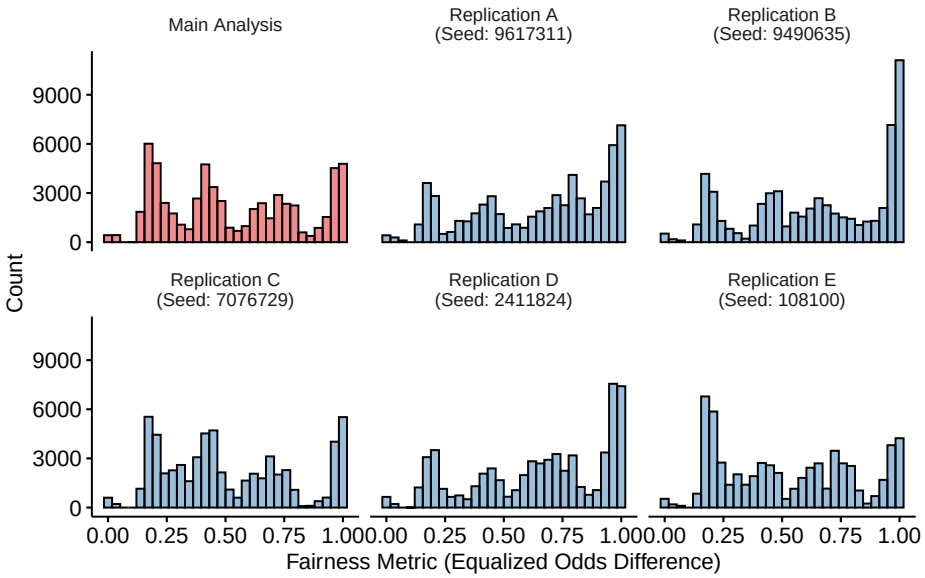


Figure A7: Overall variation in the multiverse is highly similar across different replications. Distribution of fairness metric (equalized odds difference) across universes in five different replications alongside the results reported in the main body of the paper. Lower values on the fairness metric indicate smaller *TPR* and *FPR* differences across groups.

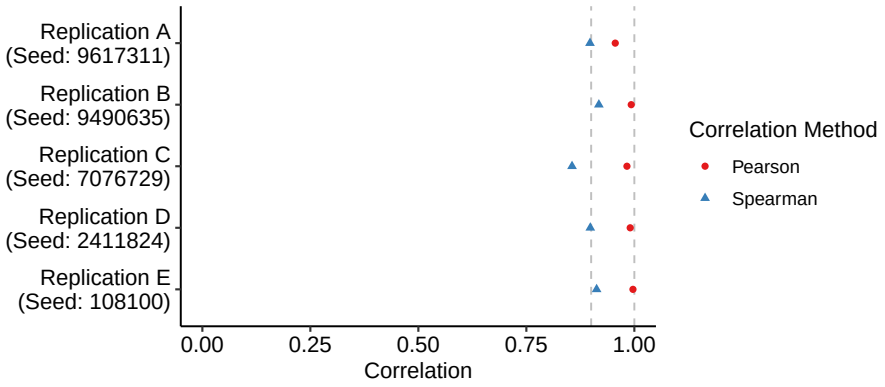


Figure A8: Estimates of decision importance are similar across replications of the analysis. Correlations of variance decomposition / importance estimates between the analysis reported in the main body of the paper and five replications. Pearson correlation coefficients are consistently higher than Spearman correlation coefficients, indicating better estimation of high-importance decisions. Dashed lines were inserted at 0.9 and 1.0 to indicate high correlation values.

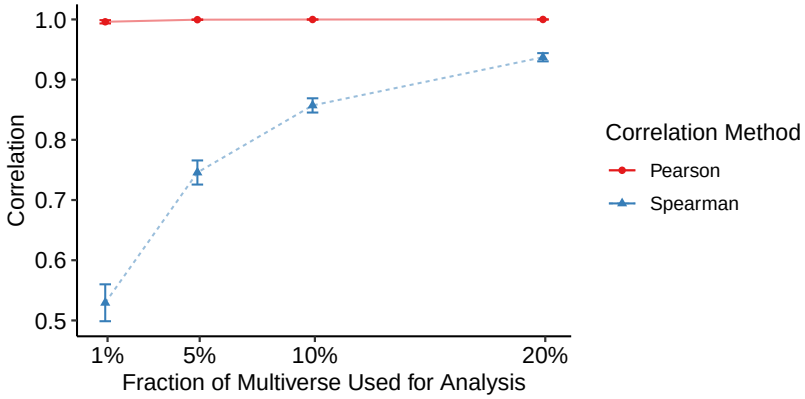


Figure A9: Conducting the analysis with smaller subsets of the complete multiverse leads to similar results. Correlations of variance decomposition / importance estimates between full dataset and random subsets of different sizes. Random subsets were drawn 50 times with points corresponding to mean correlations and lines to +/- 1 standard deviation. Pearson correlation coefficients are consistently higher than Spearman correlation coefficients.

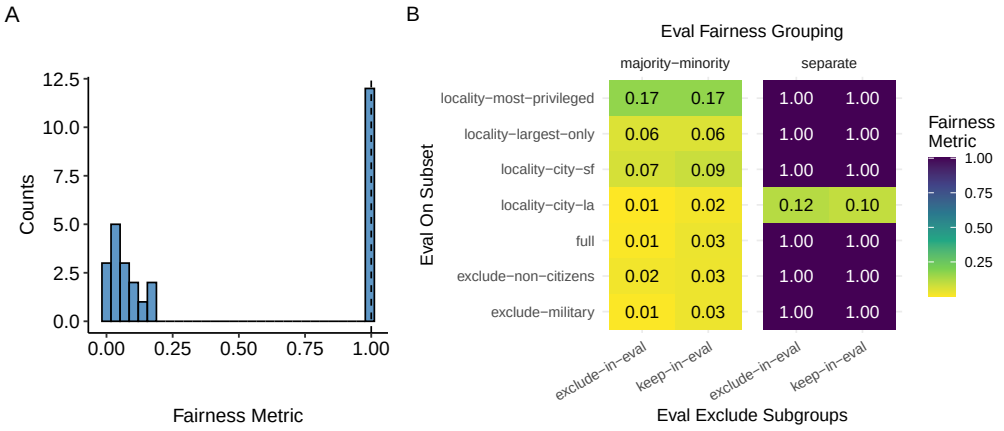


Figure A10: Evaluation decisions can strongly interact in their effect on the fairness metric. Overall distribution (A) and raw values (B) of the fairness metric for a single model exhibiting high variation over different decisions regarding its evaluation. This figure is analogous to Figure 5 in the main text.

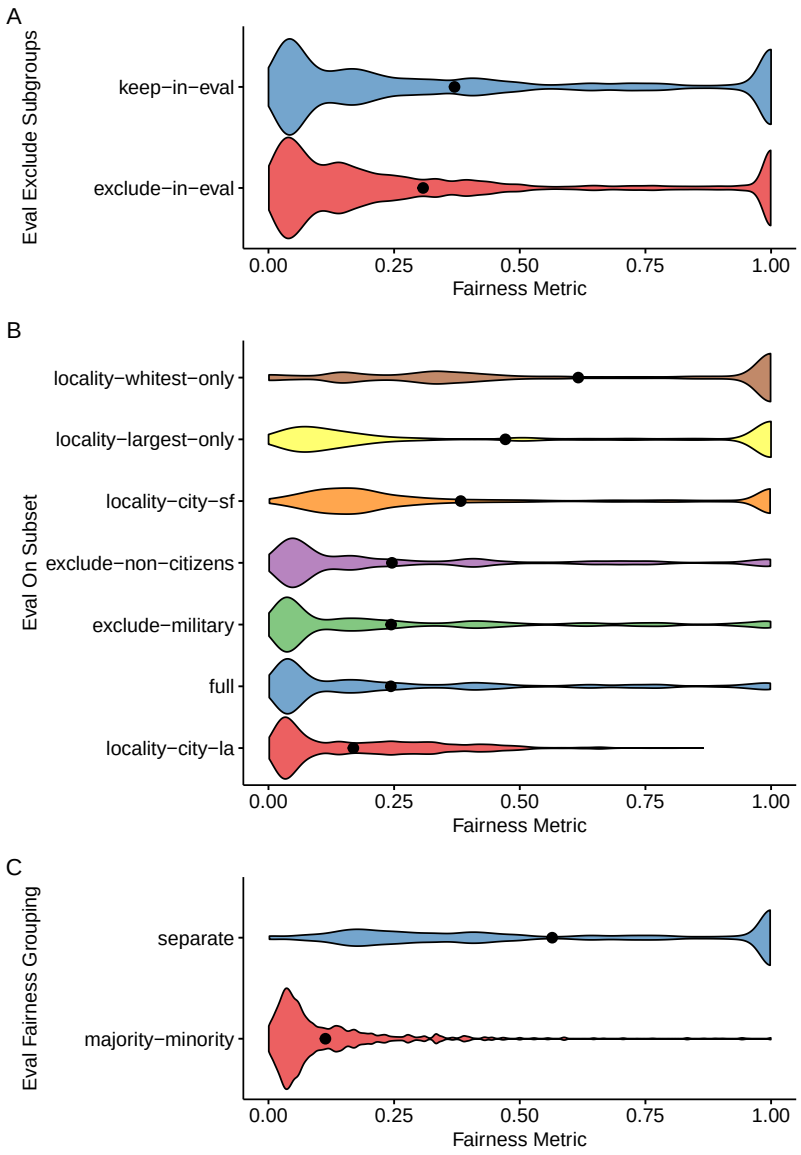


Figure A11: Despite strong interactions for the same model, evaluation decisions exhibit general tendencies in how they affect algorithmic fairness. Distribution of the fairness metric for different evaluation decisions across the complete multiverse of design decisions from studies 1 and 2.

6. Preventing Harmful Data Practices by using Participatory Input to Navigate the Machine Learning Multiverse

Contributing article

Simson, J., Draxler, F., Mehr, S., & Kern, C. (2025). Preventing Harmful Data Practices by using Participatory Input to Navigate the Machine Learning Multiverse. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pages 1–30, New York, NY, USA. Association for Computing Machinery. doi: 10.1145/3706598.3713482 URL https://doi.org/10.1145/3706598.3713482

Code repository

https://github.com/reliable-ai/participatory-multiverse/

Preregistration

https://aspredicted.org/dgyp-bs3b.pdf

Copyright information

Author contributions

J. Simson provided the initial idea for the collaboration and project. S. Mehr provided access to the data collection platform. J. Simson created the online experiment for empirical data collection. J. Simson implemented the empirical simulations in the work; analyzed the data and created all figures and tables. F. Draxler provided valuable information into the field and submission process. J. Simson lead the writing, submission and revision process of the work. C. Kern particularly contributed to the introduction (Section 1), F. Draxler to the background on participatory machine learning (Section 2.2). All authors contributed through fruitful comments, proofreading and revisions of the manuscript.

Preventing Harmful Data Practices by using Participatory Input to Navigate the Machine Learning Multiverse

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

Fiona Draxler
University of Mannheim
Mannheim, Germany
fiona.draxler@uni-mannheim.de

Samuel Mehr
School of Psychology
University of Auckland
Auckland, New Zealand
Child Study Center
Yale University
New Haven, Connecticut, USA
sam@auckland.ac.nz

Christoph Kern
Department of Statistics
LMU Munich
Munich, Germany
Munich Center for Machine Learning (MCML)
Munich, Germany
University of Mannheim
Mannheim, Germany
christoph.kern@stat.uni-muenchen.de

Abstract

In light of inherent trade-offs regarding fairness, privacy, interpretability and performance, as well as normative questions, the machine learning (ML) pipeline needs to be made accessible for public input, critical reflection and engagement of diverse stakeholders.

In this work, we introduce a participatory approach to gather input from the general public on the design of an ML pipeline. We show how people's input can be used to navigate and constrain the multiverse of decisions during both model development and evaluation. We highlight that central design decisions should be democratized rather than "optimized" to acknowledge their critical impact on the system's output downstream. We describe the iterative development of our approach and its exemplary implementation on a citizen science platform. Our results demonstrate how public participation can inform critical design decisions along the model-building pipeline and combat widespread lazy data practices.

CCS Concepts

• **Computing methodologies** → **Machine learning**; • **Human-centered computing** → **Collaborative and social computing**; **Human computer interaction (HCI)**; • **Social and professional topics** → *User characteristics.*

Keywords

Participatory Design, Machine Learning, Algorithmic Fairness, Multiverse Analysis, Citizen Science, Garden of Forking Paths

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713482

ACM Reference Format:
Jan Simson, Fiona Draxler, Samuel Mehr, and Christoph Kern. 2025. Preventing Harmful Data Practices by using Participatory Input to Navigate the Machine Learning Multiverse. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 30 pages. https://doi.org/10.1145/3706598.3713482

1 Introduction

Algorithmic decision-making (ADM) fueled by machine learning (ML) algorithms is becoming ubiquitous in many domains, affecting the lives of millions of individuals. Examples include jobseekers that are classified into different risk groups by profiling models [74], refugees that are re-allocated within their host country based on matching algorithms [7] and the denial or approval of health care coverage for patients [6, 89, 126]. While such systems are introduced with the aim of improving the effectiveness and efficiency of decision-making, there are also serious concerns that algorithmic decisions can treat the affected individuals unfairly [88, 89, 93]. Fairness implications of ADM ultimately depend on how the underlying models interact with biases and deficits in training data, and thus the design, implementation and evaluation of the ML system is of central concern [17, 117]. Addressing fairness and adverse impacts, therefore, does not only include technical measures but rather needs a broader public discourse where developers, stakeholders and affected individuals meet on an equal footing to design and evaluate algorithmic systems within their respective deployment context.

Despite increasing efforts to open up the ML pipeline and involve stakeholders in participatory designs, current research is lacking tools that enable public input where it matters most: the design of the ML model itself.

The ML pipeline includes a multitude of critical decision points – from data selection and curation to pre-processing, modeling and evaluation decisions. As each decision point allows for multiple

alternative choices, design decisions in ML resemble a *garden of forking paths* [50] where each fork corresponds to a decision which in turn leads to a set of further scenarios downstream. The full grid of decision combinations can also be understood as introducing a *multiverse* of (potential) ML models, in which each model is defined by a unique path through the set of design decisions upstream [117, 123]. Even with a handful of decisions, a multiverse can quickly grow extremely vast: For the example application we use in this paper, four design decisions lead to a multiverse with 16,352 endpoints (i.e., unique ML models) as illustrated in Figure 1, which grows to 784,896 combinations when four evaluation decisions are included — a detailed description of the use case is presented in Section 3.1.

In ADM contexts, the decision points in the ML multiverse often involve normative considerations and inherent trade-offs: should the model use sensitive attributes such as race and gender as features? Should a more complex or a more interpretable model be used for the task at hand? Which evaluation criterion (error metric) is most important when choosing the final model? Often, such decisions cannot (and should not) be "optimized" based on training data alone. A central result of the fair ML literature, for example, has shown that key fairness notions are incompatible with each other, and thus, it is essential to reason on substantive grounds which type of model error is viewed as most critical in a given application context [27]. Even for "technical" decisions, e.g., in the data pre-processing context, recent research has shown considerable design effects on model fairness downstream [24, 117]. At the same time, questionable design decisions can commonly be observed in data practice: In a review of 280 experiments in the field of fair ML, Simson et al. [116] identify harmful shortcuts such as filtering out members of ethnic minorities in data processing, which are commonly taken even by practitioners in fairness research. They group these shortcuts under the term *lazy data practices*. These observations jointly call for a democratization of the design process not only to provide critical public feedback but to engage diverse stakeholders and affected individuals (as domain experts) along the ML pipeline. As central decisions concerning data processing, evaluation and metrics need to be considered within the given application context. Active public engagement is critical to evaluate and tailor technical decisions according to the needs and perspectives of the communities in which a system is sought to be placed.

In this paper, we introduce a participatory approach to prune the garden of forking paths and help navigate the multiverse of design decisions in ML. While current work in participatory artificial intelligence (AI) rarely focuses on the collaborative shaping of the system's design, including technical decisions such as the type of model and features used [37], these decisions critically affect the eventual functioning of the system, its predictions and fairness properties. In light of inherent trade-offs (including fairness, privacy, interpretability versus performance considerations) and normative questions (e.g., which error notion – for which group – to prioritize), the ML pipeline needs to be made accessible for public input to foster inclusion and participation of diverse populations.

Using a case study of predicting public health care coverage, we demonstrate how participatory input can be implemented to produce meaningful data. Our results show how participants' choices can be better than common practices employed by practitioners and how this can be used to address harmful and lazy data practices in the field. We further demonstrate how results can be used to navigate the machine learning multiverse more efficiently, pruning pathways which are not deemed acceptable by a wide majority of participants.

Our contributions include a reusable workflow for preparing the ML pipeline for participatory input and a case study where this workflow is implemented. We further collect participatory input using a citizen science setup, successfully gathering a diverse sample of participants from across the world. We provide an empirical evaluation of this participatory data and put the results into context by simulating models in the machine learning multiverse.

2 Related Work

This paper links multiverses of decision points in ML engineering with participatory ML. Accordingly, this section introduces the multiverse concept, including application scenarios, its relation to fairness, and multiverse analyses, and it provides an overview of participatory methods in ML and related fields.

2.1 Multiverse Analysis

2.1.1 The Garden of Forking Paths. Whenever one conducts data analysis, there are many different decisions one has to make along the way [113], both explicit and implicit [117]. This has often been likened to a garden of forking paths [50], where each decision creates a fork in the path, creating a multiverse of pathways and destinations. *Many-analyst studies*, where multiple individuals or teams conduct an analysis using the same dataset and research question, show that such decisions can have large effects on the findings [21]. While awareness around this problem has mainly focused on statistical analyses, in particular on null hypothesis significance testing (NHST), it also applies within the context of machine learning and artificial intelligence [64, 90]. In many cases, practitioners may not even be aware of making decisions as they traverse the ML pipeline, although new solutions are being explored to bring potentially problematic decisions to light via notifications during development [60].

2.1.2 Hyperparameter Optimization. The classic scenario where a garden of forking paths is navigated within ML and AI is during hyperparameter optimization (HPO) [14, 47]. Here, a grid of different parameter configurations is created and traversed using either an efficient search algorithm (e.g., Bayesian Optimization [120]) or a full scan of parameter combinations. Research in this field tends to focus on efficiency and finding better search algorithms. However, it has also led to the development of new methods to better understand variance in the space, which can be adapted to the garden of forking paths or multiverse, such as efficient implementations of the functional analysis of variance [65, 68]. As the name suggests, the HPO literature focuses heavily on hyperparameters. However, it usually ignores other critical decision points in the ML pipeline, such as data selection, preprocessing or evaluation decisions. These decisions govern how the model interacts with (biases in) data and have been shown to affect fairness outcomes [24, 117].

2.1.3 Fairness & Multiplicity. Research on algorithmic fairness aims to address and reduce disparities in algorithmic systems. This

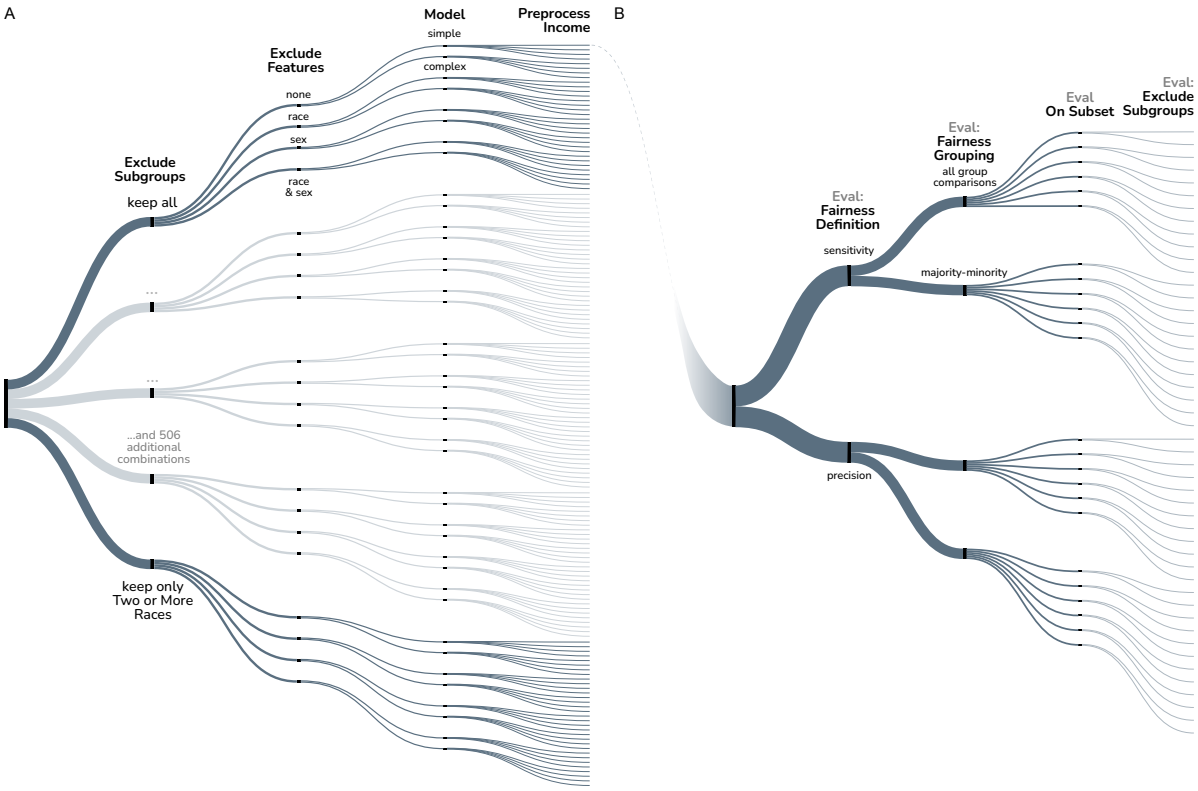


Figure 1: The full multiverse of different decisions is incredibly vast. Illustration of the multiverse of different ML models (A) and evaluation strategies (B). Due to practical limitations, only a small fraction of the multiverse of ML models is shown here, with 506 additional branches hidden, as illustrated by the reduced transparency and only 320 out of 16,352 endpoints visible (<1%). As the two multiverses are not mutually exclusive, their combined total are 784,896 combinations of model and evaluation.

includes formalizing fairness notions in a way that they can be applied to the predictions or scores of an ML system. Group fairness metrics typically compare (different types of) prediction error between groups defined by *protected attributes* [8, 93]. The selection of protected attributes may be based on anti-discrimination law, with *race* and *gender* [116] being common examples. Many specific fairness metrics, workflows, auditing and bias mitigation strategies exist, and the choice between them is neither trivial nor independent of context [8, 85, 109].

A consequence of the machine learning multiverse is that there are often different ML models with equal or comparable performance (examples in [25, 34, 107, 131]). This is termed the *Rashomon Effect* [20] and the set of comparable (best) models is commonly referred to as the *Rashomon set* [110, 131]. One benefit of the Rashomon Effect (among others [108]) is that it allows optimizing for secondary objectives such as fairness, e.g., by selecting the model with the optimal fairness metric from the Rashomon set [17, 69]. It has been argued that there may even be a legal duty to search through the Rashomon set for fairer alternative models [16].

A potentially more troubling result of the Rashomon Effect is that models in the Rashomon set can give very different predictions for a given individual [17, 29]. This phenomenon is commonly referred to as *individual arbitrariness* or *predictive multiplicity* [17, 82], although other names exist [54]. When the choice between different models and, thereby, individual predictions is arbitrary, multiplicity can lead to issues in the justifiability of predictions and decisions [17, 29].

The sources of multiplicity can be diverse, and multiplicity has been demonstrated across different dimensions, such as random seeds [29], target variables [128], different sparse decision trees [131] and the dataset generation process [91] as well as model design decisions [117]. Multiple measures have been proposed to quantify these effects [29, 41, 66]. Related work has examined the influence of different forms of imputation for missing data [24] and hyperparameter selection [42] on algorithmic fairness, albeit not

through the lens of multiplicity. A similar concept has also been demonstrated in the context of explainable ML through *fairwashing*, where equivalent models can be generated that show little dependence on sensitive features [2]. Thus, from a fairness angle, decisions along the model-building pipeline are critical: They can introduce individual arbitrariness, exacerbate or hide data deficits, and can result in a variety of models downstream, which differ strongly in their fairness properties.

2.1.4 Multiverse Analysis & Fairness Hacking. While the influence of a variety of decisions on algorithmic fairness has been well-documented in the field, it is usually only examined along a single dimension. As different decisions often interact [117], however, they should also be analyzed together. This can be done through a multiverse analysis, where all possible decision combinations – the forking paths – are evaluated.

Multiverse analyses [123] first emerged in response to the reproducibility crisis [28], where large amounts of research failed to be reproduced. A similar type of analysis, *specification curve analyses* [114], emerged around the same time. Specification curve analyses are characterized by their use of a particular type of visualization to show results from a multiverse analysis. The idea of multiverse analyses is also closely related to that of a *sensitivity analysis* [78], although multiverse analyses are typically bigger in scope, trying to include more decisions and especially more interactions between decisions.

Multiverse analyses have been successfully used in machine learning to study performance [10] and fairness [117] of models. Their most useful application in ML may be, however, as a potential solution for issues of fairness hacking [117]. *Fairness hacking* describes practices of presenting unfair models as fair [87]. This can be achieved by iterating over different definitions and metrics of algorithmic fairness [87] or evaluation strategies [117] and selecting the most favorable one while keeping the actual model fixed. A similar concept has been described under the name *d-hacking* [15].

In the ML context, a multiverse analysis allows us to explicate and structure important data processing, modeling and evaluation decisions and makes them visible and accessible. However, the decision points themselves often involve difficult trade-offs and choosing an option based on metrics alone is commonly not an option for ethical reasons. Even if a decision does not touch upon ethical issues, multiverse analyses can become unfeasible due to computational limitations, especially when costly and complex models are being fit. Participatory input has great potential in these cases to make more informed decisions and constrict the number of pathways in the multiverse.

2.2 Participatory Machine Learning

Participatory machine learning (and often used interchangeably, participatory AI) emerged as a response to power imbalances between system engineers and those affected by or those using a system [76], which can result in biases such as a disproportional impact on marginalized groups [72]. The core idea in participatory ML is to involve stakeholders in the design, development, and deployment of ML models or AI systems [13]. Thus, the aim is to empower stakeholders and to increase fairness, transparency, and accountability [37, 46]. In the context of this work, we specifically address participatory ML for *fairness*.

Application domains of participatory ML range from estimated content quality of Wikipedia edits [57] to various healthcare applications [38]. Contributors shape ML systems as participants in goal-setting or algorithm design workshops [22], model builders or adapters [57], evaluators in result assessments [79], etc. Past work has suggested a variety of tools that support participation: To name just a few examples, model cards help identify and discuss trade-offs in ML model design [112], visualizations and comparisons of model predictions for different inputs encourage reflection [26] and re-designed visualizations of prediction accuracy targeted specifically at non-experts enable them to better assess ML model performance [111].

In contrast to defining the goals of an ML project or eliciting user preferences, participatory approaches for decisions concerning the design and evaluation of ML models are relatively rare. Notably, in a review of 80 participatory AI projects, only 10% included stakeholders in design or specification tasks such as choosing models or features [37]. A related review reports that only a small share of participatory AI projects involved stakeholders beyond a "consultation" stage [31].

However, participatory model and evaluation design can be beneficial for multiple reasons. First, they increase the diversity of perspectives to take into account to avoid pitfalls that professionally blinkered ML engineers may run into. For example, past work highlights that developers often make implicit model-building and evaluation decisions that may impact model fairness [60, 117]. However, adjusting attribute weights in an ML model based on non-experts' fairness perception can actually make a system fairer – although some individual input may also worsen fairness ratings [94]. Second, many problems are only discovered when stakeholders are involved in the discussion [134], as highlighted in Criado-Perez's [33] book *Invisible Women: Exposing Data Bias in a World Designed for Men*. She argues that people must be asked to identify their needs and requirements, e.g., for reliably recognizing heart attack symptoms in women. Therefore, in our work, we look at the relationship between being a member of a minority group and the decisions that participants make.

Participatory ML also entails certain caveats. Notably, balancing power between participants and those asking for their participation is crucial to avoid exploitative practices [30, 31, 118]. There is a danger of co-optation, i.e., superficially acknowledging input and involvement, while the actual influence remains minimal [13]. It is also important to address aspects that non-experts may not be aware of. For example, when non-experts build ML models, they often optimize toward percentage accuracy and may overlook issues such as overfitting [132].

In this paper, we investigate the possibility of using participatory input for actual model decisions (design and evaluation) with a focus on fairness. Thus, we focus on specific steps of a full participatory ML pipeline for which participatory methods are currently underexplored. We rely on iteratively refined descriptions of decision choices to make them accessible for non-expert participants.

3 Methods

For our participatory approach, we chose an exemplary ML use case situated in the ADM context. We derive a set of relevant decisions on model design and model evaluation and set up an online experiment on a citizen science platform.

3.1 Decisions

We use a case study of predicting whether an individual is covered by public health care in the U.S. based on socio-demographic information with data from the American Community Survey Public Use Microdata Sample (ACS PUMS) [23]. In particular, we use the *ACSPublicCoverage* problem, one of a set of problems commonly referred to as "folktables" [40]. We opted for this problem, as it is prototypical for ADM scenarios, due to being a binary classification task which can be framed as a risk prediction problem with race defined as the protected attribute.[1] We also chose this particular problem due to its practical relevance, as healthcare is a highly important domain, with commonly reported fairness issues [6, 89, 96, 126]. In an ADM setting where decisions would be made based on the model's predictions, incorrect predictions of individual health care coverage could lead to communities falling under the radar of preventative measures or information campaigns that are allocated based on the predicted risk of non-coverage. This is particularly concerning for minority groups with historically low coverage rates. Specifically, incorrectly predicting that an individual is covered (i.e., a false positive) might exclude them from preventative measures or targeted campaigns, while incorrectly predicting non-coverage (i.e., a false negative) could lead to a misallocation of such measures. While these implications likely unfold differently if individuals are covered by private insurance plans instead, the *ACSPublicCoverage* task focuses on individuals with an income of less than 30,000 U.S. Dollar per year who may depend more strongly on public offers.

When determining the list of decisions to include in the study, we focused in particular on decisions which may be made "ad-hoc" (sometimes even without the awareness of making a decision), but which eventually introduce trade-offs and involve normative considerations. We selected, adapted and extended upon a list of decisions identified in prior work [117]. Implicit decision-making is particularly common for decisions regarding the evaluation of an ML system. Therefore, we ultimately decided to include four design decisions affecting the model itself and four affecting only its evaluation.

For each decision, we crafted a brief introductory text describing the trade-offs inherent to the decision, taking care not to present any one option as more favorable than others. We refined this introductory text as well as the descriptions of each option in the decision across multiple iterations to make them understandable without prior knowledge of machine learning or artificial intelligence concepts.

A brief explanation of each decision and its options can be found below. The actual wording of each decision, its introductory text and options can be found in Appendix B.3. An overview of all decisions and options can be found in Table 2, and the resulting multiverse is illustrated in Figure 1.

3.1.1 Model Design Decisions. We included four decisions on model design, covering preprocessing, data and model selection. Each decision includes between two and nine distinct options. However, as the decision *Exclude Subgroups* makes use of the combination of its nine options, it allows for a total of 511 unique combinations of these options. Together, the four model design decisions create a multiverse of 16,352 potential ML models. A subset of this multiverse is illustrated in Figure 1A.

Exclude Subgroups. Related work shows that the exclusion of certain subgroups before model training is common. For example, a recent review focusing on the popular *COMPAS* [5] dataset found that 38 out of 59 studies excluded data from subgroups of the protected attribute [116]. While certainly problematic, as it can lead to representation bias [88, 125], this does not have to be out of malicious intent: One may want to exclude data from certain subgroups to protect their privacy or to make analyses less complicated. Indeed, many commonly used fairness metrics are first and foremost designed with the assumption of only two protected groups, requiring adaptations to work with more nuanced protected attributes. However, we also want to clearly note that this is a problematic trend, and our inclusion of the decision in this study stems from the intent to highlight this issue rather than normalize the practice. We advise for careful deliberation and against the exclusion of subgroups in most real-world scenarios. In our experiment, we present all nine racial and ethnic groups available in the data of our modeling task to participants, using the order suggested by the ACS PUMS [23]. Participants have the option of combining the groups as they see fit to construct the list of included subgroups.

Participants were randomly assigned to one of two conditions for this decision: Either they were shown the percentages illustrating the relative size of each group next to the group name or not. We added this differentiation because overestimation of minority group sizes is common [3].

Exclude Features. It is common in fairness-related contexts to not include potentially sensitive attributes as predictive features in models due to privacy reasons or with the intent to produce less biased models [55, 77]. This practice does not necessarily produce fairer models, as fairness through unawareness has been shown not to work [8]. Nonetheless, we included this decision to represent the popular practice. We include four different options for this decision: (1) To exclude the protected attribute *race* as a feature for the model, (2) to exclude the potentially sensitive attribute *sex* as a feature for the model, (3) to exclude both *race* and *sex* as features or (4) to exclude neither of the two and use both as features for the model.

Preprocess Income. We included the preprocessing of the variable *income* as an example of a feature that is often binned into different categories. Income data is highly relevant to the outcome we are trying to predict, and the correct processing of income data is usually an arbitrary choice without a clear consensus. Preprocessing of income data was also shown to be an influential decision among several comparable preprocessing decisions in its effect on algorithmic fairness [117]. Indeed, the arbitrariness of thresholding income data (as a target) has been criticized [40] as one of the issues

[1] At the same time, datasets that feature other (more prominent) examples from the fairness literature (recidivism prediction, credit scoring) have been shown to suffer from considerable quality issues [43].

in *Adult* [9], a popular dataset based on U.S. Census data with an associated machine learning task of predicting whether an individual's income is above $50,000. We included four different options for processing income data: (1) Keeping it as is, (2) binning it into bins of $10,000 and binning it into (3) three or (4) four equally sized groups.

Model. Choosing the model type is a critical and consequential decision point in the ML pipeline. We thus wanted to include at least one decision on the type of ML model that is used. However, during the development of the decision and option descriptions, we quickly realized that explaining the nuances of different ML models is beyond the scope of a single study and may be challenging to understand for non-experts. We therefore opted to only present participants with the choice between (1) a simple and (2) a more complex model, highlighting the classic trade-off between performance and interpretability with increased complexity [51]. We decided to use a logistic regression [32] for the simple and a random forest [63] for the more complex model. Models were fit using default hyperparameters.

3.1.2 Evaluation Decisions. We included four distinct decisions on model evaluation, each with two to six options. Together, these decisions allow for 48 different strategies of how one might choose to evaluate a (fixed) ML model (Figure 1B). As they can be applied to each of the models created in the previous step, they increase the size of the multiverse from 16,352 unique models to 784,896 different evaluations.

Eval Fairness Definition. Deciding on a particular metric in fair ML is one of the most critical decisions, with multiple valid and conflicting options. While one may be able to narrow down the list of metrics using, for example, fairness-specific metric decision trees [85, 109], alternative choices and, thereby, metrics are usually also plausible. As it has been shown that one cannot optimize for all possible notions of fairness at the same time, this means that one will eventually have to prioritize some metrics over others [8]. One complicating factor with this is that a malicious actor can abuse this ambiguity to pick a definition that produces more favorable scores post-hoc, a practice termed *fairness hacking* [87] (cf. Section 2.1.4).

The full list of potential fairness metrics is large, with the nuances and trade-offs between different definitions often quite subtle and hard to judge, even for experienced practitioners. We, therefore, chose to focus on the important trade-offs between competing concepts of fairness in ML, which can be traced back to different ways of conceptualizing errors (and, conversely, prediction performance). The two options we included for this decision are (1) a focus on sensitivity, corresponding to the fairness concept *separation* and (2) a focus on precision, corresponding to the fairness concept *sufficiency* [85].

Eval Fairness Grouping. When identities from subgroups are not excluded in analyses, they are often aggregated away instead. The most common form of aggregation is to aggregate multiple different groups (e.g., several racial subgroups) into a majority (the biggest subgroup) and minority group (all other subgroups). This is often done when there are substantial imbalances in group size or for convenience reasons such as enabling simpler analyses.

The same study mentioned earlier [116] found that of the 21 studies (using *COMPAS*) that did not exclude data from subgroups, all but one aggregated data from different groups together. Across aggregation and exclusion, 53 out of 59 studies reduced the protected attribute to just two groups.

While we believe that the normalization of such practices is harmful, we still want to represent them in this study, with the hope that participatory input may help us to better understand public opinion of such practices. We, therefore, include two options for this decision: (1) To aggregate the protected attributes into two groups (majority, minority) when calculating fairness metrics or (2) to calculate fairness metrics as the maximum difference between all different combinations of subgroups of the protected attribute[2].

Eval On Subset. Machine learning systems are often evaluated using data that may not represent the eventual target population. This can be due to practical constraints such as limited resources or because only a certain smaller subset of the target population is reachable. A system may have also been developed with a certain target population in mind, but then the scope widened during or after development.

While one would generally want to evaluate a system using a sample that resembles the population eventually affected by the tool, we included this decision to also represent these more practical trade-offs between the cost of data collection and representation in the data. We included the following options for this decision: (1) Using data from the most populous area, (2) using data from the area where the most people have public health insurance, (3) using data from the closest major city, (4) using data from as many people as possible, but excluding military veterans, (5) using data of only U.S. citizens and (6) using data from the overall population in the United States. Note that in our case study, data from the overall population is available, but we create subsets for the decision options to represent and model different data collection practices.

Eval Exclude Subgroups. When data from a subgroup is excluded, it is typically excluded during both training and evaluation of a system. This is problematic, as it will hide any resulting issues affecting the excluded groups during evaluation, whether they be related to fairness or the quality of predictions. We decided to include this decision not due to inherent trade-offs but to see whether participants would pick up on this potential issue and whether their input could help address it.

There were two options available for this decision: (1) To exclude the same subgroups as in the training data or (2) to include all subgroups for evaluation. This decision was only shown to participants if it made logical sense based on their response to the decision *Exclude Subgroups*. When participants chose to exclude certain subgroups from the training data earlier, they were therefore shown this decision later on, asking whether they also wished to exclude data from groups when evaluating the system.

3.2 Procedure

The study was created using jsPsych version 7.3.1 [36], with data collected in World-Wide-Lab version 0.4.1 [130]. Our goal with the

[2]This corresponds to the default behavior in the commonly used Python library fairlearn [12]

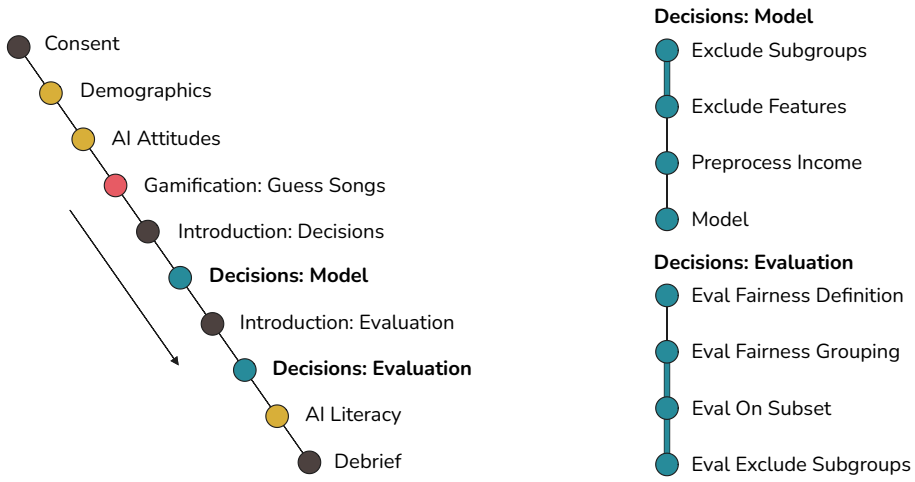


Figure 2: Diagram illustrating the different sections of the study and how they followed each other. The right-hand side of the graphic illustrates the different decisions which were presented in their respective sections. Decisions connected by a thicker line are considered to be within the same logical block during model design, and their order was randomized.

study was to recruit a diverse sample, as different peoples' identities and lived experiences can lend them unique forms of expertise [39] and shape their views on AI ethics [71, 72, 102, 122]. Since gamified citizen science studies have shown promise in recruiting diverse populations [62, 81], we opted for a similar approach in this study. We added a short game as an incentive and embedded the study on a citizen science website. We chose this particular website for its popularity and since its theme was unlikely to strongly bias recruitment in relation to the case study. As this website happened to be music-themed, the gamified section was about identifying whether or not a piece of music was generated by AI. To at least *partially* address the global north bias present in many fields of research (including AI ethics [106]), we chose to also make the study available for non-U.S. participants. Given the diversity of healthcare systems across the world, we believe that non-U.S. participants might indeed have valuable insights from living with different healthcare systems. Due to different notions of race across regions [70] we chose to slightly adapt demographics for U.S. versus non-U.S. participants (see below) and examine the data for differences in this regard later on.

Participants were first presented with a screen where the study was briefly explained, and they could then provide informed consent to participate. This was followed by a list of demographic items, in particular age, country of residence, primary and secondary language, race (U.S. only) and self-assessed membership of a minority (non-U.S. only). Afterwards, participants completed the ATTARI-12 [124], a 12-item questionnaire assessing their attitudes towards AI. Then, participants completed the recruitment game of the study, listening to a random collection of 6 AI-generated and 4 human-made 12-second music samples. After each song, participants were asked to guess whether a song was AI-generated, and they received feedback immediately afterwards. Next, participants were shown an introductory text illustrating the case study and why their input would be valuable. After the introduction, participants were shown the individual decisions, with decisions presented in the order they would be encountered during the design of an ML system. If there was no clear order between decisions, their order was randomized[3]. A second block of decisions related to the *evaluation* of an ML system was preceded by another brief introduction explaining basic concepts of fairness in ML. Each decision was presented with a brief introductory text explaining the decision, followed by a list of options. If there was no inherent order to options, their order was randomized. In addition to the actual decision options, each decision always presented the options of "I don't understand the description", "I prefer not to answer" and to "Suggest an alternative option". The evaluation decisions were followed by a short 16-item questionnaire assessing AI literacy [103]. For consistency, both attitudes towards AI [124] and AI literacy [103] were collected using 7-point response scales, labeled *strongly disagree* and *strongly agree* at their ends. Respondents were not required to select any option for the decision trials or AI attitudes / literacy response scales. At the end of the study, participants were shown a debrief with basic information about the study, their final score in the gamified section and a detailed list of the songs they listened to. Before the final screen, participants had the option to provide general feedback about the study and its design.

A high-level overview of the study procedure can be seen in Figure 2. The complete list of decisions and options presented to participants can be found in Appendix B. The study and its analyses were preregistered before any data were collected. The preregistration can be found at https://aspredicted.org/dgyp-bs3b.pdf.

[3]Due to an error in an earlier version of the experiment code, 10.67% of participants saw decisions in a fixed order.

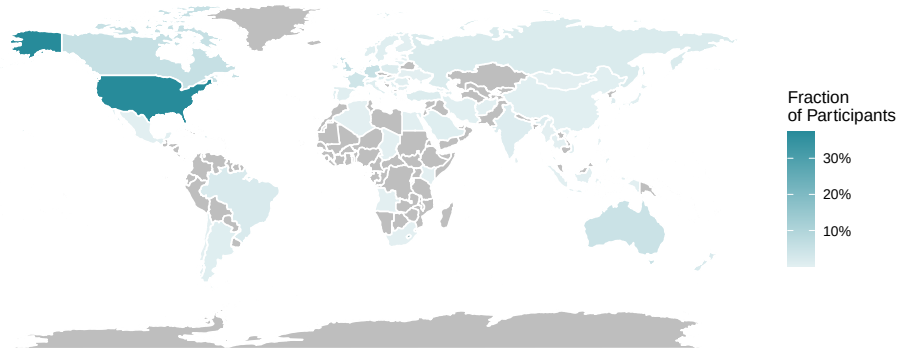


Figure 3: Participants were recruited from across the world with an over-representation of Western countries, especially the United States. Choropleth map of the world, shaded based on self-reported country of residence. A detailed breakdown of sample size per country can be found in Table 1.

3.3 Participants

Participants were mainly recruited through organic internet traffic to the music-focused citizen science website https://themusiclab.org. The study was also shared on a mailing list for auditory experiments and posted on social media platforms (Bluesky, Reddit). After a brief pilot, data for the study itself was collected over 20 days.

A total of 1403 sessions were recorded in the study, with 375 sessions completing the whole study[4]. As a new session is recorded every time someone navigates to the study or refreshes the webpage, this rate of completion (26.73%) is well within the expected range and slightly higher than the overall completion rate of other studies on the website during this time (22.10%).

We restricted the sample to sessions with data available for at least one of the decision trials. This left us with a final sample size of N = 534 individual sessions by n = 517 participants. As only 17 participants had multiple sessions with decision data, we decided to retain their data. Unless stated otherwise, the following analyses will use all the available data and will include every session with data available for a particular decision.

While there is a strong over-representation of both Western and especially English-speaking Western countries, there is also a significant number of non-Western countries present in the data. A graphical overview of participation rates by country can be seen in Figure 3, with detailed counts available in Table 1. Further information on the sample composition is available in Section A of the appendix.

The distribution of AI attitudes [124] ($M = 3.18$, $SD = 1.13$) and AI literacy [103] ($M = 3.20$, $SD = 1.22$) among participants displayed a high degree of variation, with the sample slightly leaning towards more positive AI attitudes and higher AI literacy (Figure 13).

4 Results

Detailed information on the software used for analyses is available in Section C. Code for the multiverse analysis simulation was adapted and extended from prior work [117] and implemented using an early version of the package multiversum [115]. The source code of simulations conducted in this study is available at https://github.com/reliable-ai/participatory-multiverse.

Below, we first address the validity of the survey and responses. Then we proceed to the core parts of the analysis: (1) The decision distributions including relevant differences between groups and (2) the resulting multiverse.

4.1 Quality of Decisions

Decisions were generally perceived as clear, with < 10% of participants checking that they did not understand a description across any decision (Figure 4). A sizable fraction of participants either did not check any answer option at all or indicated that they prefered not to answer a decision, an option we purposefully allowed them to do. This is not surprising, given that there were no monetary incentives and that the study was presented as a secondary objective to the gamified section to participants. While only a very small fraction of participants suggested alternative options during the piloting of the study, this number increased during data collection for the main study. In practice, the option to suggest an alternative was also used as a general feedback outlet by participants. In a small number of instances, participants also used this field to provide detailed and nuanced suggestions. Only $n = 83$ out of $N = 3,315$ responses were excluded for being unreasonably fast, at under two seconds, as the majority of people who did not bother to read descriptions would just not respond at all.

[4]A total of $n = 2$ sessions were excluded from analyses due to corruption in their data: In one case, this seems to have happened due to a particularly unreliable internet connection and in the other due to use of a non-standard in-browser translation feature which interfered with data collection.

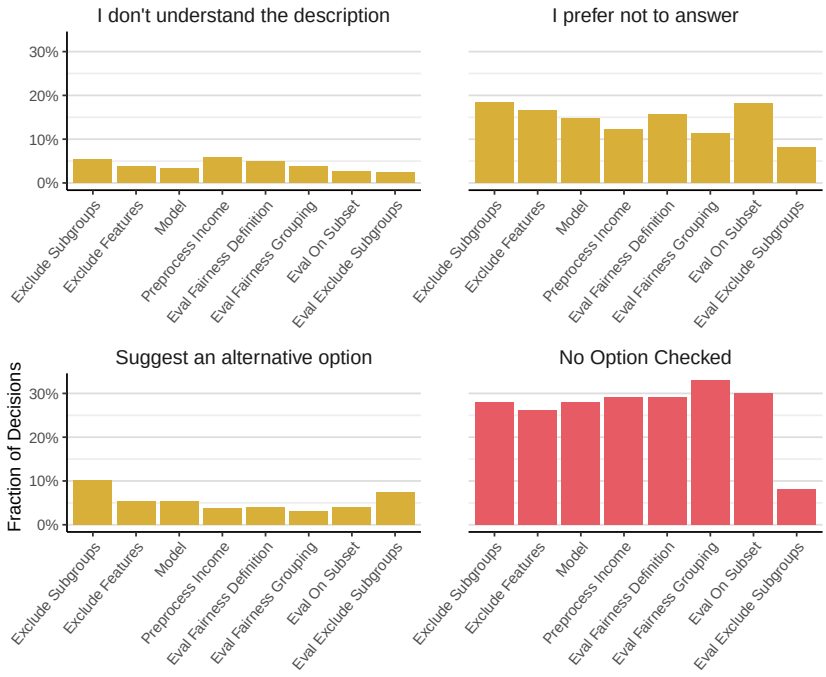


Figure 4: Decisions were generally well understood, and only a few additional options were suggested, but a large fraction of participants clicked through decisions without ticking any option or indicated that they prefer not to answer. Prevalence of different non-option answering trends across decisions in the study.

4.2 Distribution of Ratings

An overview of the overall distribution of ratings in the participatory multiverse can be seen in Figure 5. The figure shows an illustration of the different paths that participants took through the multiverse, weighted by how many participants chose a particular path and split into the multiverse of models (Figure 5A) and evaluations (Figure 5B). It becomes immediately visible that there are a few popular pathways and that if a participant does not respond to a decision or indicates that they prefer not to answer, they tend to do this for all decisions (red path "empty response"). The individual distributions of ratings for all countries with sufficient data are available as an interactive analysis at https://reliable-ai.github.io/participatory-multiverse/.

The degree of agreement between participants is generally high but varies by decision. Figure 6 illustrates this by showing the cumulative prevalence of different *combinations of options*. Naturally, decisions with a high number of different options generally have a lower prevalence per combination. Still, about half of all participants chose the same combination of options out of 511 theoretically possible combinations for *Exclude Subgroups*, indicating strong agreement across participants. The opposite is the case for the choice of *Eval Fairness Definition*: Here, participants had to chose one of two possible metrics (a combination was not possible), and choices are almost exactly equally distributed.

Figure 7 illustrates the different combinations of options observed in the data for the decisions *Exclude Subgroups* and *Eval On Subset*. It has to be noted here that for the decision *Exclude Subgroups*, the combination of choices is indeed what is used in the end, whereas for the decision *Eval On Subset*, each selected option is a separate valid strategy to be explored. Interestingly, there is sometimes little overlap between combinations of options that are of similar popularity. The results here also indicate that participatory results need to be taken with caution, as the second most popular option for the decision *Exclude Subgroups* was to use only data from people who identify as White.

4.2.1 Differences between Groups. We investigated whether participants' answers differed based on different characteristics and conditions. Due to the large number of possible combinations between participant characteristics and decisions, we chose to examine only a small subset of possible decisions of particular interest. We evaluated group differences based on individual votes, including partial data, but excluding responses where the respective grouping information was missing. We calculated comparisons for each option using Bonferroni correction to correct for multiple testing within

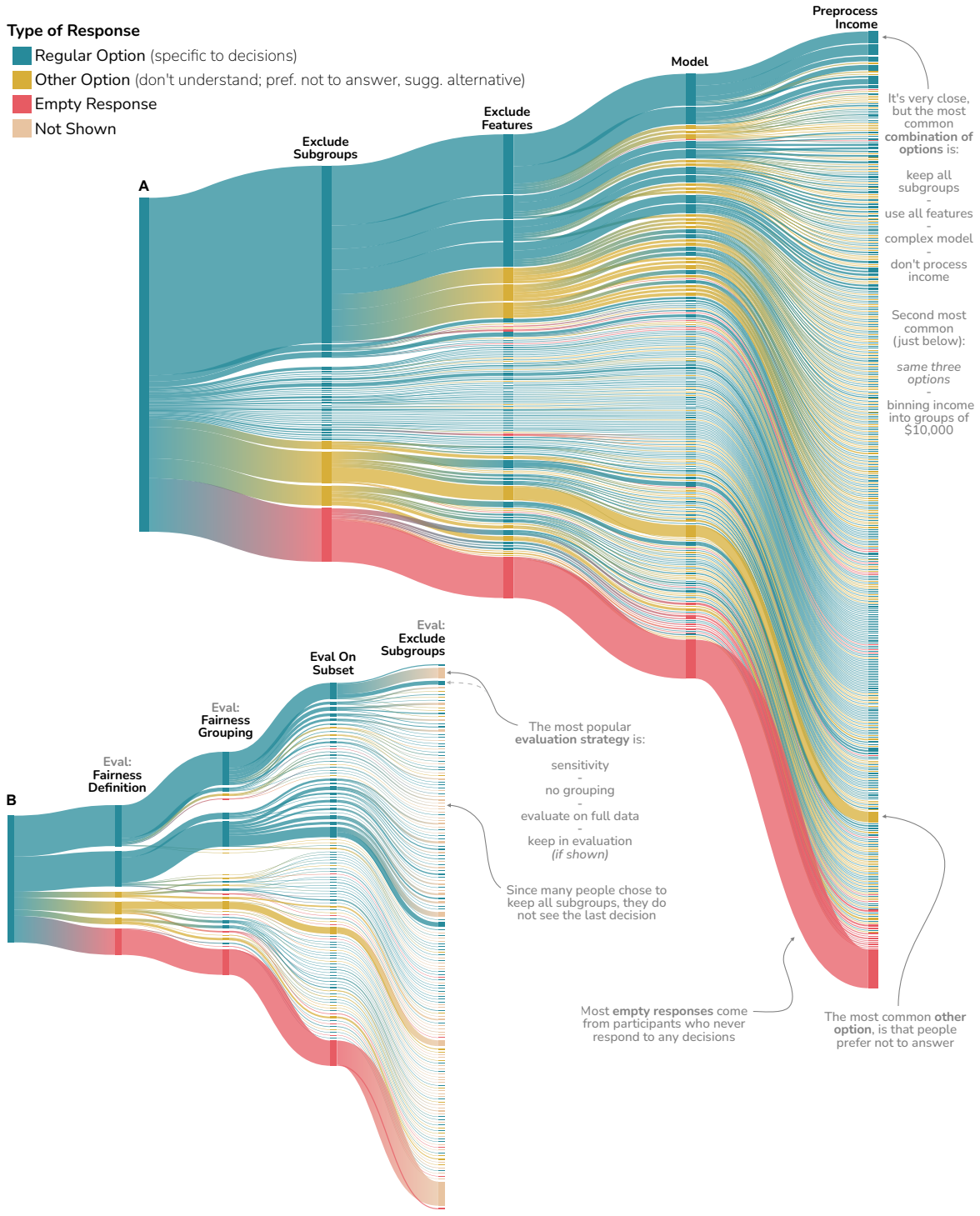


Figure 5: Specific paths in the participatory multiverse are significantly more popular than others, and if participants decide not to respond, they do this consistently. Weighted illustration of the multiverse of model design (A) and evaluation (B) decisions based on participants' votes. Each split corresponds to a decision taken by participants. Only data from participants with data available for all four decisions represented in each diagram were included.

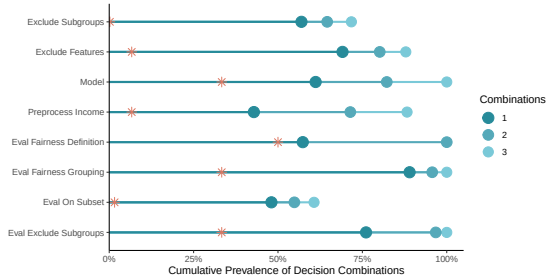


Figure 6: Agreement differs greatly across different decisions. Cumulative frequency of the three most common combinations of options across decisions. The minimum rate of agreement for each decision is highlighted with a star. The decision *Eval Fairness Definition* did not allow the selection of a combination of options.

each decision. For each comparison, frequencies of votes were compared using Fisher's exact frequency test [49] with a significance level $\alpha < 0.05$.

We examined three different comparisons for the decision *Exclude Subgroups*: First, whether participants were more likely to include a subgroup if they were also a member of it, second, whether responses differed based on the country of residency, and third, whether displaying percentages next to the different groups would have an effect on choices. Participants were indeed more likely to include subgroups if they were members of them, although the effect was small, especially in comparison to the overall tendency of participants to include many different subgroups ($p = 0.02$, $OR = 2.43$; Figure 8A). Responses also differed between data from the U.S. and other countries, with higher rates of inclusion in data from the United States (Figure 8B, Table 3). While the effect was only significant for the group "White", this could be due to an interaction of it being the biggest group and the previously described effect of higher rates of including one's own race. Whether or not percentages were shown next to the different groups had a negligible effect on participants' choices, as can be seen in Figure 15 (Table 4).

We examined whether participants who identify with a certain gender (Figure 16A, Table 5) or as being part of a minority (Figure 16B, Table 6) were more or less likely to exclude certain (corresponding) features from the model (*Exclude Features*). While we do see differences between the different groups here, results are hard to interpret due to significant imbalances between group sizes, and differences were not statistically significant.

In order to enable group comparisons between different levels of AI literacy and AI attitudes, we created three equally sized groups on each scale. The distribution of the three groups on either scale can be seen in Figure 14 and compared to the overall distributions in Figure 13.

The decision of which metric to prioritize (*Eval Fairness Metric*) is one of the most technical decisions encountered within the study. We therefore examined whether responses to it would be different based on self-reported AI literacy. Interestingly, we did not observe any significant differences on the regular response options. However, there were significant differences in the number of suggested alternatives and empty responses (Figure 9, Table 7).

We further examined whether there are differences in both explicit non-responses (checking "I prefer not to answer") or empty responses (checking none of the options) based on participants' self-reported AI attitudes. While most of these comparisons did not show significant differences, the observed frequencies show an interesting counter-play with both positive and negative AI attitudes generally showing a higher tendency to explicitly check "I prefer not to answer" (Figure 17, Table 8), whereas the middle group tends to not respond at all more often (Figure 10, Table 9). This is most likely related to overall response tendencies, with participants who are less engaged in the study opting for satisficing response strategies [75], such as responding closer towards the center of a scale and opting not to respond in the later sections of the study. A statistically significant difference in non-response between groups was found for the decisions *Exclude Features* and *Model* (Table 9).

4.3 Multiverse

Besides examining participants' ratings directly, we also conducted a multiverse analysis by traversing the complete multiverse of models, building and evaluating each of the models. In this section, we first present data from the complete multiverse of models before integrating it with participatory input and examining the intersection.

4.3.1 The Full Multiverse. We fit all 16,352 ML models in the multiverse and evaluated each using the 48 different evaluation strategies. This resulted in a total number of $N = 784,896$ different scores.

As our primary metrics of algorithmic fairness, we calculate the difference of either sensitivity (Eq. 1) or precision (Eq. 2) across groups of the protected attribute *race*. Across all combinations of any two racial groups (i, j), the maximum of the differences is recorded as the fairness score[5]. If there are only two groups due to aggregation (via *Eval Fairness Grouping*) or exclusion (via *Exclude Subgroups* and *Eval Exclude Subgroups*), only the difference between those two groups is used. Whether to use sensitivity or precision here corresponds to one of the decisions in the multiverse (*Eval Fairness Metric*). As a reference to these two metrics, we also use Equalized Odds Difference [1, 58], a commonly used fairness metric.

$$\text{Sensitivity} = \text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

$$\Delta\text{Sensitivity}_{\max} = \max_{i,j} \left| \text{Sensitivity}_i - \text{Sensitivity}_j \right| \quad (1)$$

$$\text{Precision} = \text{Positive Predictive Value} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

$$\Delta\text{Precision}_{\max} = \max_{i,j} \left| \text{Precision}_i - \text{Precision}_j \right| \quad (2)$$

[5]We note that this form of aggregation, which is commonly used in practice [12] and used here for descriptive purposes, can lead to an overestimation of performance differences between groups when the number of groups compared is large [83].

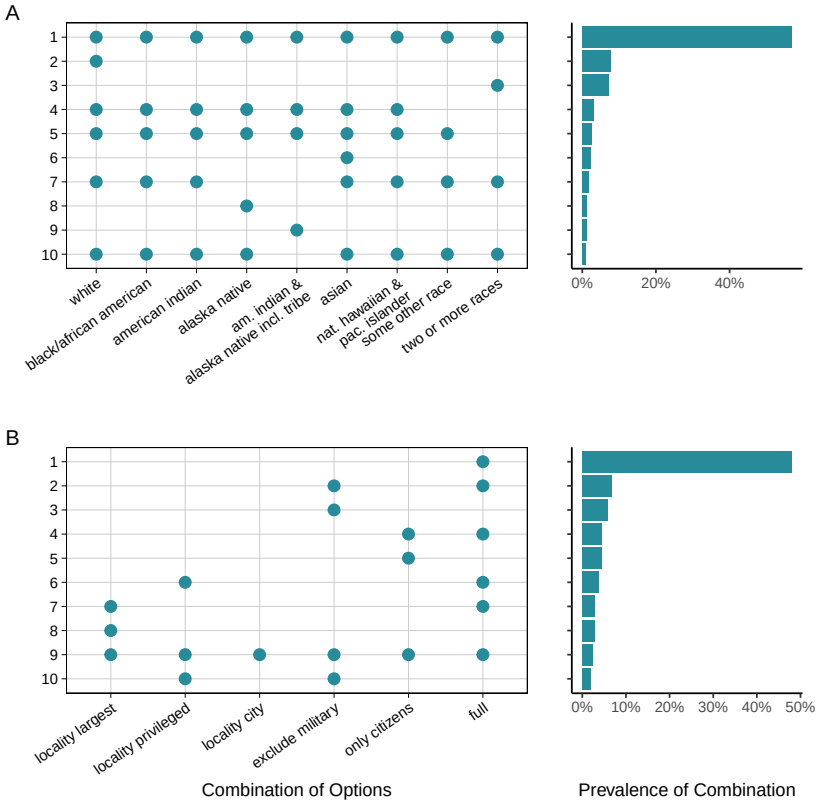


Figure 7: While participants often agree on a single combination of options, the degree of their agreement differs between decisions. The ten most common combinations of chosen options for the decisions *Exclude Subgroups* (A) and *Evaluate on Subset* (B), alongside their respective prevalence.

An overview of the complete multiverse of fairness scores for both metrics ($\Delta\text{Sensitivity}_{\text{max}}$ and $\Delta\text{Precision}_{\text{max}}$) can be seen in Figure 11. Variation within the multiverse is high, with values of both fairness metrics spanning their full possible range from 0 to 1, with a standard deviation of $SD = 0.347$ for $\Delta\text{Sensitivity}_{\text{max}}$ and $SD = 0.353$ for $\Delta\text{Precision}_{\text{max}}$. When examining the figure, the large spread of scores and a clustering towards the two ends of the scale become evident. A large degree of this extreme variation, however, can be attributed to different evaluation strategies.[6] Using a fixed evaluation strategy (see below), a more condensed distribution emerges (Figure 11, in red).

This brings up the question of which evaluation strategy one should use to evaluate the multiverse of models. In the present example, we suggest a strategy of more conservative choices, opting to not group the protected attribute, to use data from all subgroups during evaluation and to evaluate on the full subset of data. While we believe this to be a reasonable choice, it is by no means commonly used in the literature (as discussed in earlier sections, see also Simson et al. [116]). Luckily, this issue can be addressed well using the participatory data: Exactly this combination of evaluation choices is also the most popular in the present data. Therefore, we use this evaluation strategy to fix evaluations for the following analyses.

As there is no clear reason to favor one of the two definitions of fairness metrics over the other and since the decision did not impact scores in a very strong matter, we chose to not fix this decision, but rather continue to examine the two separately going forward.

4.3.2 The Participatory Multiverse. Participants' votes can be combined with data from the multiverse to create a participatory multiverse[7]. Participatory data can be combined with the theoretical full

[6]While some differences in fairness scores across, e.g., different data subsets may be expected, note that the comparisons made by the metrics (performance difference between racial groups) remain the same. The large spread of scores visible here highlights the susceptibility of fairness results of models trained for the same task to changes in the evaluation protocol, opening up opportunities for fairness hacking [15, 87] when evaluation strategies remain unchecked.

[7]As participatory data can contain missing data or non-responses, the combination is actually not a straightforward join. Rather, we assigned equal weights to all participants and spread these weights across all endpoints in the multiverse that match

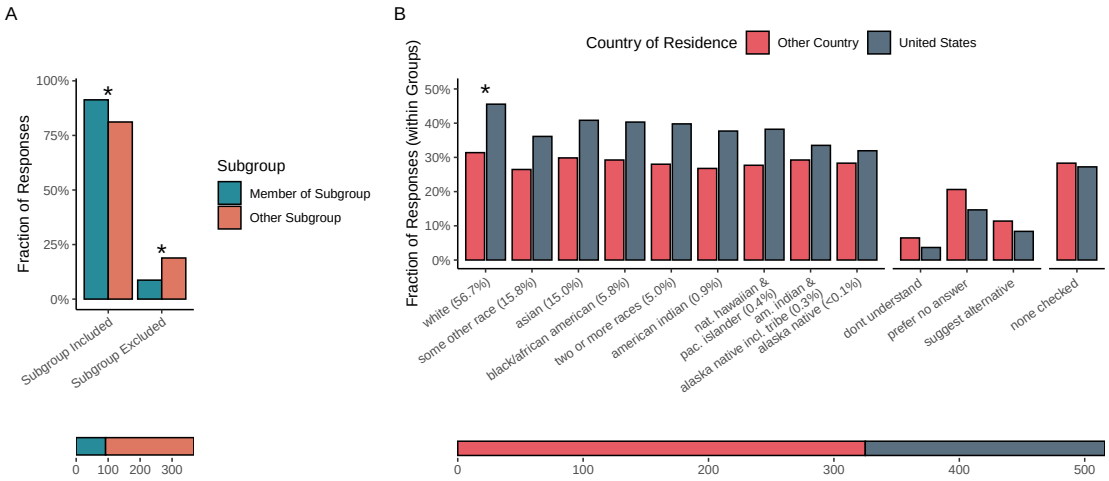


Figure 8: Participants were more likely to include a group that they are a member of (A) and more likely to include subgroups when from the United States (B). Inclusion of subgroups split by membership of participant (A) and country of residence (B) for the decision *Exclude Subgroups*. Bars below plots indicate the raw group distribution and number of votes.

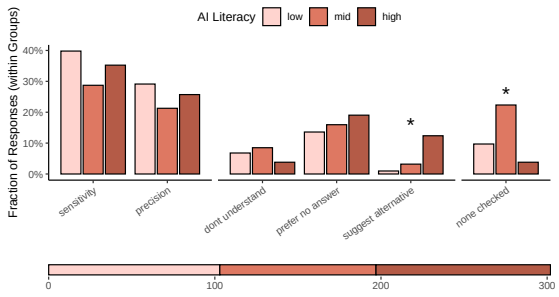


Figure 9: Participants with higher AI literacy opted to suggest alternatives for the fairness metric more often. Response to the decision *Eval Fairness Metric* split by self-reported AI literacy. Bar below plot indicates the raw group distribution and number of votes.

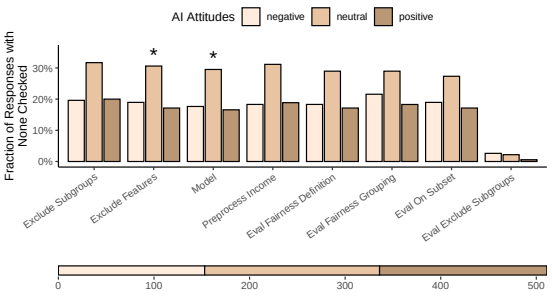


Figure 10: Participants with neutral AI attitudes showed a higher rate of non-response. Fraction of participants choosing *not to check any option* across decisions split by self-reported AI attitudes. Bars below plot indicate the raw group distribution and number of votes.

multiverse (Figure 1) to put variation in the participatory data into context. The resulting data can also be used to prune the full multiverse before its computation. This makes it possible to only examine and compute the most popular branches in the multiverse, greatly reducing computational costs. We discuss several approaches for this in Section 5.2.

Participatory data can also be combined with results from a multiverse analysis. This makes it possible to put the participatory multiverse of models and scores into context using actual metrics. This is exactly what we did here: We combined participants' votes with data from the multiverse analysis to investigate how a more narrow multiverse, weighted by participants' choices, compares to the one created by the complete multiverse analysis.

As can be seen in Figure 12, there are a significant number of models with equal or near-equal performance but significant variation in their fairness scores, illustrating the Rashomon Effect. However, besides areas of equal performance with "free" fairness gains, there also exists a Pareto front of models where any increase in either

their decisions. A very precise combination of responses will, therefore, assign more weight to its resulting endpoints than a very broad one. Multiple other algorithms for combination are conceivable.

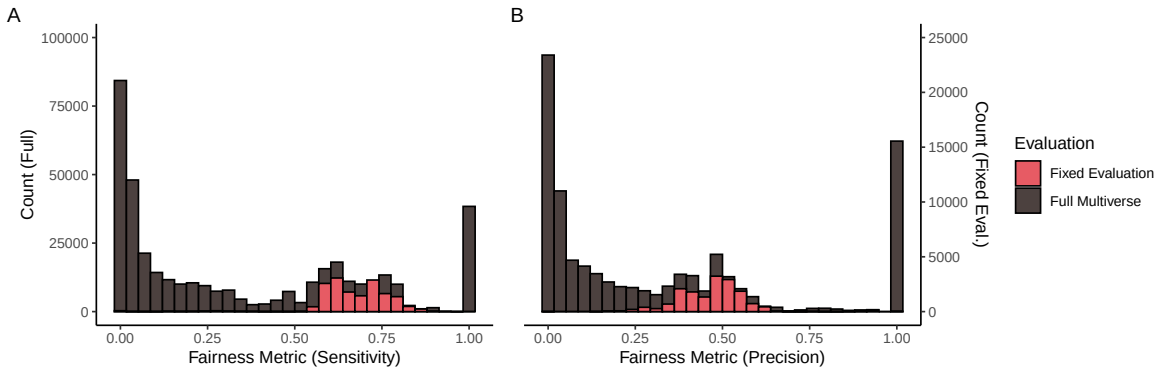


Figure 11: Examining the complete multiverse is unrealistic; rather one will want to use a fixed evaluation strategy to meaningfully compare different models. Histogram of evaluation scores across the complete multiverse (left axis) and with a fixed evaluation strategy (right axis). The multiverse with a fixed evaluation is scaled by a factor of four to be well visible. Fairness metric corresponds to $\Delta\text{Sensitivity}_{\text{max}}$ (A) and $\Delta\text{Precision}_{\text{max}}$ (B), both ranging from 0 to 1 with lower scores being preferable.

performance or fairness would come at a cost to the other. Interestingly, the distribution of models chosen by participants is clustered closer to these Pareto fronts. In particular, the most popular model is situated very closely to the technically "optimal" combination, indicating a competitive model. Repeating the analysis using Equalized Odds Difference yields similar results (Figure 18). Still, given the various trade-offs and nuanced implications of the different decision points, the "performance" of the most popular models cannot be fully reflected by the present metrics. Participants' preferences introduce a new dimension in its own right.

5 Discussion

In this work, we explore a novel workflow that uses participatory input to restrict the decision space during the design of an ML system. Using a case study of predicting public health care coverage on U.S. data, we illustrate this workflow, prototype a selection of decisions, collect citizen science data on which options people deem appropriate and evaluate data in the light of the resulting machine learning multiverse.

5.1 Summary of Findings

Our results indicate that the participatory multiverse approach shows great promise to be applicable in real-world scenarios, although care needs to be taken during design and evaluation. We were able to successfully collect diverse input from across the world, indicating a high willingness of participants to engage with the topic. We collected meaningful input on (complex) modeling and evaluation decisions, exhibiting differing levels of agreement across decisions. It should also be noted that while a large fraction of participants opted not to respond to the decisions, this behavior was to be expected as we solicited input for a hypothetical ML task on a citizen science platform rather than engaging affected individuals of an actual ADM system.

The most popular combinations of people's choices were very reasonable options across the different decisions. Especially in the context of evaluation strategies, the most popular evaluation strategy excluded "lazy" practices, such as evaluating fairness using a simplified setup of majority and minority data. This is a positive result, as these practices commonly occur in the literature, and it illustrates the opportunity for participatory input to combat them.

There were a few significant differences in participants' responses based on their membership to certain groups. However, many more comparisons could be made than the ones we presented here. We therefore argue for the importance of collecting data from a wide and diverse sample of people to represent all identities which will be impacted by a potential system, as anticipating the degree of diversity in views ahead of time will be difficult.

When comparing the complete multiverse of plausible ML models with one weighted by participatory input, we see that the participatory multiverse leans towards favorable models in both performance and fairness. Especially the most popular model, which would follow from a democratic vote, was situated closely to the Pareto front across different definitions of algorithmic fairness. This also aligns with findings from a study of participatory feature weighting, where participants' decisions tended to improve fairness evaluation [94].

As participatory input is a democratic process, one also has to anticipate and embrace disagreement. In the present data, we saw both high agreement and disagreement for different decisions. This should inform how different decisions are handled. For example, since agreement on the definition of a particular fairness metric was quite low, we chose to explore both options of the decision in more detail than other decisions which displayed significantly higher rates of agreement.

5.2 The Participatory Multiverse Workflow

We present the following steps as an outline for our workflow of collecting participatory input when designing an ML system.

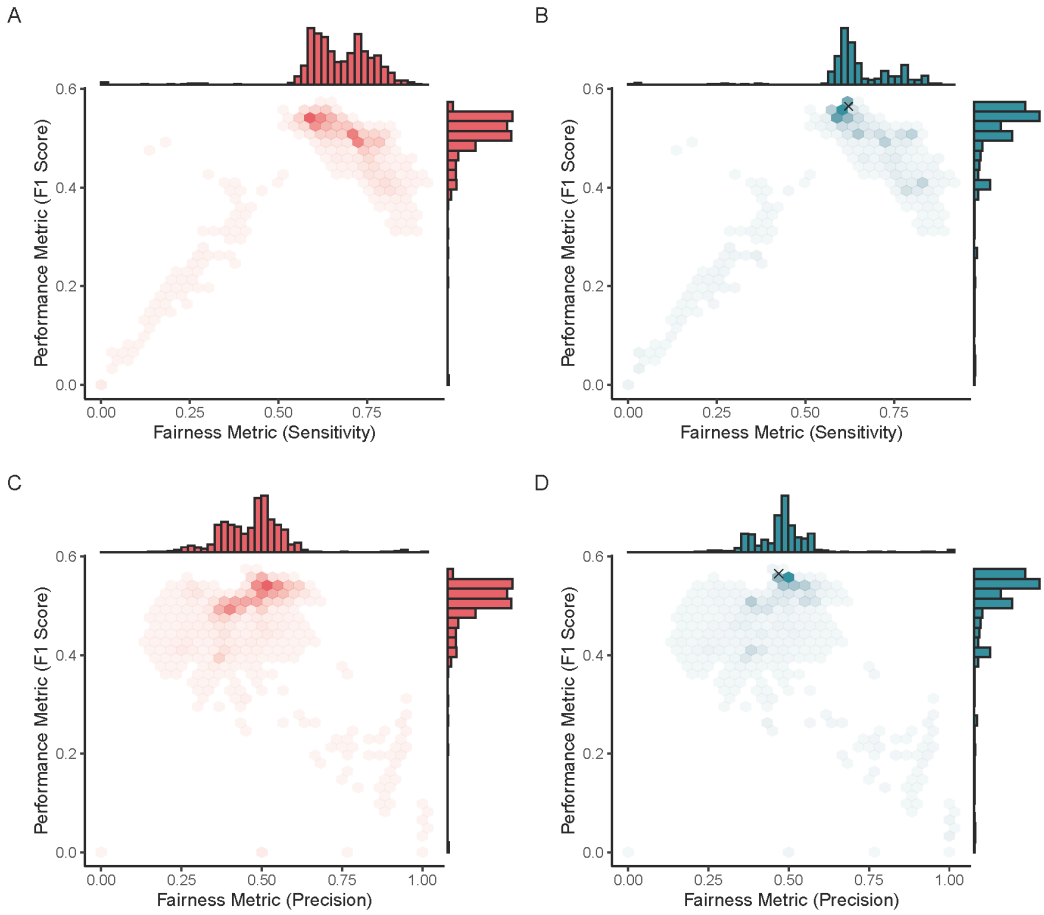


Figure 12: Models from the multiverse weighted by participants' votes are close to the Pareto front. Comparison between a complete multiverse of models (A, C) and one based on participants' votes (B, D) on performance (F1 Score, higher is better) and fairness metric (lower is better). Both multiverses are evaluated using a fixed strategy, split by definition of fairness metric ($\Delta\text{Sensitivity}_{\text{max}}$: A, B; $\Delta\text{Precision}_{\text{max}}$: C, D). Darker areas correspond to a higher clustering of models. Crosses indicate the most popular models among participants.

We also add recommendations for implementing the workflow in practice.

(1) **Decide on your target population and platform.** It is important to be clear on who the target population of both, the participatory input, but also the ML / ADM system is. While this would ideally be the same population in both cases, it may not always be feasible in real-world scenarios. Great care also needs to be taken in the actual sourcing of a chosen target sample. As demonstrated in this work, participants' characteristics may influence their responses to decisions, and it is therefore vital to include all relevant groups in the discussion. For both aspects, measurement and representation, participatory approaches can draw on the rich insights of survey science [11, 56]. Moreover, the platform in which participatory input is collected should fit and inform the intended scope and may range from a large-scale online platform to a town hall for in-depth discourse with stakeholders.

(2) **Source a list of decisions that affect your ML pipeline.** We recommend focusing on the complete pipeline here, including often-neglected steps such as the sourcing and (pre-)processing of data. It can be helpful to involve multiple people into this process to spot potentially overlooked

decision points. Co-designed fairness checklists [84] and documentation aids like model cards [92] can also be a starting point for identifying decisions.

(3) **Identify the different options for each decision.** One should keep an open mind when collecting the list of options, but also only include actually feasible options in each case.
(4) **Prepare the final list of decisions that you will gather participatory input on.** Depending on the application scenario, there may be practical limitations on which kinds of decisions you can present to participants and how many. In this case, it may be useful to prioritize value-related decisions, where stakeholders' attitudes are more likely to differ from those of ML engineers [71, 72, 134]. However, to the extent possible, we advocate for the inclusion of decisions that may seem to have an obvious correct choice and decisions that may be challenging for non-experts (cf. Section 3.1.1), as these may unearth implicit biases. In turn, this may foster reflection and counteract lazy practices.
(5) **Develop the actual wording for each decision and its options.** This should include introductory statements for each decision, outlining the trade-offs between different options. Care should be taken not to bias future participants towards any particular option in a decision [18]. We further recommend allowing participants to help highlight potential issues in the text surrounding a decision, e.g., by including other decision options as presented in this study (e.g., I don't understand the description; I prefer not to answer; suggest an alternative option [80]). The decision-design step should consist of multiple iterations where feedback is gathered from stakeholders in between each iteration, potentially including empirical data from a multiverse analysis. We recommend including additional, information-only sections to explain the context in which the system is used as well as more complicated foundations which may be required for certain decisions. We also recommend being as explicit and practical as possible when explaining decisions and options, opting for applied examples within the context of the ML system over more general statements.
(6) **Launch an initial small-scale pilot** to verify that participants understand the decisions and verify data is collected as expected. Participants should have an option here to provide suggestions for each decision as well as overall study feedback.
(7) **Gather participatory input.** Participants should be encouraged to freely choose which decisions to provide input on and which not. In online settings, we recommend only light input validation to allow participants who are not interested to skip parts of the study. The best course of action will depend on the particular context of participation (see Section 5.3).
(8) **Make decisions based on participatory input.** Participatory input will need to be turned into actual decisions to create the final ML model. There are many ways to achieve this, but committing on an approach ahead of time and explicitly communicating it to participants can serve to empower them. Care should be taken when choosing a strategy, as certain strategies, such as majority votes, risk reinforcing existing power imbalances [45]. Alternative approaches [35, 53] explicitly harnessing disagreement and participant characteristics can be worth exploring as well as approaches to quantify consensus, using it for thresholding and mapping out the opinion space [67, 119]. We suggest considering prohibitory approaches where participatory input has the power to rule out certain options and developers are left with options to choose in a limited space that is deemed acceptable by the public.

We highlight that implementing even parts of this workflow can be beneficial, as it allows for critical reflection of the ML pipeline and may be used to inform a follow-up multiverse analysis across decisions where participatory input may not have been feasible. Additional benefits for developers include surfacing new options they were not yet aware of while potentially saving costs by reducing the size of the multiverse left to explore.

5.3 Practical Implications

In this paper, we presented an illustrative application of the workflow focusing on the design of a real ML model, but for a "hypothetical" ADM system without real-world deployment. Below, we highlight key considerations for applying the workflow in ADM practice.

Special care should be taken when determining the *target population* for participatory input in light of the specific use case. When it is unclear which characteristics, expertise and lived experience are truly relevant, we recommend getting input from diverse sets of populations. Here, one should be open to recognize the different forms of expertise [39] different communities might possess, which can be hard to assess from the outside. In any case, all affected populations should always be considered for their input and thereby given a voice, especially so if they are part of marginalized populations, which can be more sensitive to biases [72].

The *mode, degree and setup of participation* will depend on the target population(s) as well as the decisions one plans to gather input on. While we specifically chose online citizen science to allow for broad participation and as a challenging form of participatory input to test the limits of the approach, we encourage considering the full breadth of participatory methods when applying the workflow in ADM practice. This also includes adequate forms of compensation to participants for the time and work they put into giving their input. This is especially important when working with potentially vulnerable populations.

Last, it is important to *avoid exploitative and extractive forms of participation* such as "participation washing" and manufactured consent [118]. At the same time, one should be aware that participatory design will not be able to solve all issues an AI system might face [13], including issues related to the data a system is built upon.

5.4 The Participatory Multiverse as a Participatory ML Method

With the Participatory Multiverse workflow, we contribute a novel method for gathering participatory input in the model design and evaluation steps of the ML pipeline. Thus, the workflow serves as a building block for a part of the pipeline that currently receives less

participatory input than earlier steps such as needs assessment [31]. It puts particular emphasis on fairness by reaching out to a wide audience of stakeholders and capturing their attitudes on questions that also address ethical and value-based decisions. Complementing related qualitative participatory approaches like Value-Sensitive Algorithm Design [134] and UI tools for fairness solicitation [26], we gather structured input that can inform decisions in a more quantitative manner. In practice, these methods can (and should) also be combined, e.g., identifying crucial decisions and choices with qualitative methods, inviting a smaller set of people but gathering in-depth insights, before following up with a participatory multiverse survey.

Mapping the Participatory Multiverse workflow to Delgado et al.'s [37] framework for evaluating participation, the *participation goal* of the workflow is to *include* participants to better align the model and its evaluation with stakeholders' preferences and values. The *scope of the participation* is *collaboration* with participants to query their preferences for *system design* aspects such as model features and *consultation* and *inclusion* for stakeholders' feedback and expertise, encouraging the suggestion of other options and providing space for additional comments. Finally, the *form of participation*, at the minimum, includes *consultation* via questionnaires, which makes it easy to apply the results to the multiverse analysis. In our case study, we ran a single iteration of the workflow, situated within an hypothetical ADM example. However, through repeated queries at different points in time, the approach could easily be extended to empower participants as *collaborators* involved in the ongoing decision-making processes. For example, participant input on design or evaluation decisions could be collected to adapt decision options for a second iteration, making these suggestions available to a broader audience. Repeated participatory input on model and evaluation decisions, even after initial deployment, could help adjust ML models in consideration of observed outcomes and potential biases. The participatory multiverse could also be used as a starting point for collaborating with specific groups of participants, e.g. those who voiced opinions that diverge from practices ML engineers would typically apply. For future work, we suggest further expansions towards higher-level participation (*collaborate* and *own*) that reduce power imbalances [13].

5.5 Limitations

There are several important limitations which apply to the results of the case study as well as the workflow itself.

It is important to note that, while geographically diverse, the sample used within this study is a convenience sample recruited from the internet. As such, it is not an accurate representation of the general public, and we do not claim that reported results represent effects, attitudes or convictions of the general public. Rather, results should be interpreted as a case study, illustrating that this particular workflow can work and produced highly useful results in this particular context. Nonetheless, the results emphasize that even a small case study can already provide a benefit by revealing controversial decisions.

Further, the decisions explored in this study represent a small subset of potential decisions encountered during the design of a real-world ML system. While they may serve as an inspiration for applications, they do not represent a holistic overview of decisions and are specific to this particular context. We also purposefully included potentially harmful decisions (related to lazy practices) to understand how these are addressed by participants. Any real-world usage of this workflow should most likely not include these decisions. Applying the workflow will require a careful evaluation of which decisions may be relevant in a particular context.

While able to provide useful information, the participatory workflow is by no means a replacement for expert input and one should not rely solely on participatory data to design an ML system. When implementing the workflow in practice, special care has to be taken when selecting which decisions to present to people, as despite introductory statements, certain concepts, especially in ML, may be too hard to convey within the context of a short survey or study. More elaborate settings, such as workshops with educational sections (e.g. [52]), may be used to address this limitation.

Due to practical limitations the present case study relies on informative participation, mostly *consulting* and *including* participants, rather than granting *ownership* [37] and *control* [31]. While practical limitations may restrict aspects of open online participatory design, one can borrow ideas from Delgado et al. [37] and Corbett et al. [31] to improve the mode of participation, such as including stakeholders in the design of the participatory system itself and enabling them to take part in the formulation of goals.

As monetary incentives may be unrealistic in many real-world scenarios of participatory input, where a system may be created by a government or NGO entity, this puts practical limitations on the scope, number and complexity of decisions one is able to implement. While we have demonstrated here that collection of participatory inputs without monetary incentives is quite feasible, it did inform our choice of decisions as well as their wording and framing. The lack of monetary incentives also meant that we were unable to include more detailed checks to confirm participants' understanding of the decisions asked to them. We did allow participants to explicitly check that they did not understand a decision, however, and we observed that popular choices largely reflected good decisions. Future research focusing solely on this issue will be required to better understand the degree to which participants may or may not be able to answer more technical ML decisions and how best to describe these.

Introductions and explanations for decisions also carry the risk of priming or influencing respondents to choose a particular option. While we had multiple rounds of iterative development in this study to minimize such influences, it is impossible to rule them out completely. In particular, the surprisingly high prevalence of choosing only "White alone" in the decision *Exclude Subgroups* could be influenced by the introductory text to the decision. Participatory input from the workflow should, therefore, always be evaluated with an eye on the specific wording of each decision. The framing of the study as secondary to the gamified section could have also influenced participants' responses. Although such an effect cannot be entirely ruled out, it is unlikely to have significantly impacted the results of this study. Participants who perceived the study as unimportant due to its framing had the option to refrain from responding. Consequently, it is possible that some of the non-responses can be attributed to this framing.

It is important to note that the workflow presented here represents just a small part of the bigger picture of creating and implementing an ML system. While participatory input can help reduce problematic practices, as demonstrated, it is by no means a replacement for critical reflection by the practitioners implementing the system. Care has to be taken to avoid a "tyranny of the majority". Rather, participatory input and careful reflection should be used jointly, each serving to constrain the garden of forking paths. This is especially the case when an ML system is implemented as part of an ADM pipeline.

6 Conclusion and Outlook

This work presents the *participatory multiverse* as a new workflow of incorporating participatory input into ML design and applies the workflow to a case study of predicting public health care coverage in the United States. Results from this case study demonstrate how participatory input can improve and inform the design of ML systems in an effective manner. In particular, results highlight how participatory input can be used to restrict both the multiverse of different models as well as different evaluation strategies, combating widespread lazy practices and aligning nuanced decisions and trade-offs with public preferences.

While we successfully apply the workflow in this work, future, theory-driven work will be necessary to better understand the interaction of participants' characteristics and their responses. The workflow is also currently limited in which decisions are suitable for participatory input by the complexity of certain concepts. Future work may help address this issue by focusing on particular concepts (e.g., metrics of algorithmic fairness) and how participatory input can be sourced on these.

We hope that adopting the participatory multiverse alongside standard ML workflows will lead to better overall systems, especially so in ADM, and give unheard voices a chance to be heard.

Acknowledgments

We extend our heartfelt thanks to all the participants who generously took part in our study. Their contributions were essential to the success of this research, and we deeply appreciate the time and effort they contributed to this project.

We would also like to express our gratitude to Malte Schierholz for his input and feedback during the writing process and Frauke Kreuter, Frederic Gerdon, Marcus Novotny and Clara Strasser Ceballos for their input and feedback during the review process of this work.

This work is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research and the Munich Center for Machine Learning.

Large language models (LLMs) were used to generate of a small number of helper functions during the technical implementation of the study and analyses. Code completions from GitHub Copilot were used when conducting the multiverse analysis. LLM-generated code presents less than 0.1% of the total code written for this project. LLMs were used to generate first drafts of figure descriptions for accessibility purposes, with all descriptions being reviewed, edited and often completely rewritten by the authors.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. arXiv:1803.02453 [cs]

[2] 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.

[3] R. Alba, R. G. Rumbaut, and K. Marotz. 2005. A Distorted Nation: Perceptions of Racial/Ethnic Group Sizes and Attitudes Toward Immigrants and Other Minorities. *Social Forces* 84, 2 (Dec. 2005), 901–919. https://doi.org/10.1353/sof.2006.0002

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

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *ProPublica* (May 2016), 254–264.

[6] Daniel Atherton. 2023. Incident Number 608. *AI Incident Database* (2023). https://incidentdatabase.ai/cite/608

[7] Kirk Bansak, Jeremy Ferwerda, Jens Hainmueller, Andrea Dillon, Dominik Hangartner, Duncan Lawrence, and Jeremy Weinstein. 2018. Improving refugee integration through data-driven algorithmic assignment. *Science* 359, 6373 (Jan. 2018), 325–329. https://doi.org/10.1126/science.aao4408

[8] 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.

[9] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

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

[11] Paul P Biemer. 2010. Total survey error: Design, implementation, and evaluation. *Public opinion quarterly* 74, 5 (2010), 817–848.

[12] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. *Fairlearn: A toolkit for assessing and improving fairness in AI.* Technical Report MSR-TR-2020-32. Microsoft. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[13] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, Arlington VA USA, 1–8. https://doi.org/10.1145/3551624.3555290

[14] 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 (March 2023). https://doi.org/10.1002/widm.1484

[15] Emily Black, Talia Gillis, and Zara Yasmine Hall. 2024. D-Hacking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 602–615. https://doi.org/10.1145/3630106.3658928

[16] Emily Black, Logan Koepke, Pauline Kim, Solon Barocas, and Mingwei Hsu. 2024. The Legal Duty to Search for Less Discriminatory Algorithms. arXiv:2406.06817 [cs]

[17] Emily Black, Manish Raghavan, and Solon Barocas. 2022. Model Multiplicity: Opportunities, Concerns, and Solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 850–863. https://doi.org/10.1145/3531146.3533149

[18] Kathrin Bogner and Uta Landrock. 2016. Response biases in standardised surveys. GESIS survey guidelines.

[19] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D^3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (Dec. 2011), 2301–2309. https://doi.org/10.1109/TVCG.2011.185

[20] Leo Breiman. 2001. Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author). *Statistical science* 16, 3 (2001), 199–231. https://doi.org/10.1214/ss/1009213726

[21] 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, Alejandro Ecker, Achim Edelmann, Maureen A. Eger, Simon Ellerbrock, Anna Forke, Andrea 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 (Nov. 2022), e2203150119. https://doi.org/10.1073/pnas.2203150119

[22] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-making in Child Welfare Services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–12. https://doi.org/10.1145/3290605.3300271

[23] US Census Bureau. 2021. Understanding and using the American Community Survey public use microdata sample files: What data users need to know.

[24] Simon Caton, Saiteja Malisetty, and Christian Haas. 2022. Impact of imputation strategies on fairness in machine learning. *Journal of Artificial Intelligence Research* 74 (2022), 1011–1035. https://doi.org/10.1613/jair.1.13197

[25] Chaofan Chen, Kangcheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang. 2018. An Interpretable Model with Globally Consistent Explanations for Credit Risk. https://doi.org/10.48550/arXiv.1811.12615 arXiv:1811.12615 [cs, stat]

[26] Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Steven Wu, and Haiyi Zhu. 2021. Soliciting Stakeholders' Fairness Notions in Child Maltreatment Predictive Systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–17. https://doi.org/10.1145/3411764.3445308

[27] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163. https://doi.org/10.1089/big.2016.0047

[28] 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.

[29] 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

[30] Ned Cooper and Alexandra Zafiroglu. 2024. From Fitting Participation to Forging Relationships: The Art of Participatory ML. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–9. https://doi.org/10.1145/3613904.3642775

[31] Eric Corbett, Emily Denton, and Sheena Erete. 2023. Power and Public Participation in AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, Boston MA USA, 1–13. https://doi.org/10.1145/3617694.3623228

[32] David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)* 20, 2 (1958), 215–232. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x Publisher: Wiley Online Library.

[33] Caroline Criado-Perez. 2019. *Invisible women: exposing data bias in a world designed for men*. Chatto & Windus, London. OCLC: 1084316434.

[34] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. 2022. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *Journal of Machine Learning Research* 23, 226 (2022), 1–61. https://www.jmlr.org/papers/v23/20-1335.html

[35] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics* 10 (Jan. 2022), 92–110. https://doi.org/10.1162/tacl_a_00449

[36] Joshua R. de Leeuw. 2015. jsPsych: A JavaScript Library for Creating Behavioral Experiments in a Web Browser. *Behavior Research Methods* 47, 1 (March 2015), 1–12. https://doi.org/10.3758/s13428-014-0458-y

[37] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, Boston MA USA, 1–23. https://doi.org/10.1145/3617694.3623261

[38] Kerstin Denecke, Elia Gabarron, Rebecca Grainger, Stathis Th. Konstantinidis, Annie Lau, Octavio Rivera-Romero, Talya Miron-Shatz, and Mark Merolli. 2019. Artificial Intelligence for Participatory Health: Applications, Impact, and Future Implications: Contribution of the IMIA Participatory Health and Social Media Working Group. *Yearbook of Medical Informatics* 28, 01 (Aug. 2019), 165–173. https://doi.org/10.1055/s-0039-1677902

[39] Mark Diaz and Angela D. R. Smith. 2024. What Makes An Expert? Reviewing How ML Researchers Define "Expert". *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7 (Oct. 2024), 358–370. https://doi.org/10.1609/aies.v7i1.31642

[40] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 6478–6490. https://proceedings.neurips.cc/paper_files/paper/2021/file/32e54441e6382a7fbacbbbaf3c450059-Paper.pdf

[41] Jiayun Dong and Cynthia Rudin. 2020. Variable Importance Clouds: A Way to Explore Variable Importance for the Set of Good Models. https://doi.org/10.48550/arXiv.1901.03209 arXiv:1901.03209 [cs, stat]

[42] 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).

[43] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery* 36, 6 (Sept. 2022), 2074–2152. https://doi.org/10.1007/s10618-022-00854-z

[44] Mike FC, Trevor L. Davis, and ggplot2 authors. 2024. *ggpattern: 'ggplot2' Pattern Geoms*. https://CRAN.R-project.org/package=ggpattern R package version 1.1.1.

[45] Michael Feffer, Hoda Heidari, and Zachary C. Lipton. 2023. Moral Machine or Tyranny of the Majority? *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 5 (June 2023), 5974–5982. https://doi.org/10.1609/aaai.v37i5.25739

[46] Michael Feffer, Michael Skirpan, Zachary Lipton, and Hoda Heidari. 2023. From Preference Elicitation to Participatory ML: A Critical Survey & Guidelines for Future Research. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Montréal QC Canada, 38–48. https://doi.org/10.1145/3600211.3604661

[47] Matthias Feurer and Frank Hutter. 2019. Hyperparameter Optimization. In *Automated Machine Learning: Methods, Systems, Challenges*, Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren (Eds.). Springer International Publishing, Cham, 3–33. https://doi.org/10.1007/978-3-030-05318-5_1

[48] Sam Firke. 2023. *janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor R package version 2.2.0.

[49] Ronald A Fisher. 1922. On the interpretation of χ 2 from contingency tables, and the calculation of P. *Journal of the royal statistical society* 85, 1 (1922), 87–94. https://doi.org/10.2307/2340521

[50] Andrew Gelman and Eric Loken. 2014. The statistical crisis in science. *American scientist* 102, 6 (2014), 460–465.

[51] Rayid Ghani and Malte Schierholz. 2021. Machine Learning. In *Big data and social science* (second edition ed.). CRC Press.

[52] Global AI Dialogue. 2024. A Workshop Series on the Impact of Artificial Intelligence (AI) on our Everyday Lives. https://perfectfuturedesign.com/global_ai_dialogue_info_english/.

[53] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association

for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3491102.3502004

[54] Travis Greene, Galit Shmueli, Jan Fell, Ching-Fu Lin, and Han-Wei Liu. 2022. Forks Over Knives: Predictive Inconsistency in Criminal Justice Algorithmic Risk Assessment Tools. *Journal of the Royal Statistical Society Series A: Statistics in Society* 185, Supplement_2 (Dec. 2022), S692–S723. https://doi.org/10.1111/rssa.12966

[55] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2016. The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making. In *NIPS Symposium on Machine Learning and the Law*, Vol. 1. Barcelona, Spain, 11.

[56] Robert M Groves, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau. 2011. *Survey methodology*. John Wiley & Sons.

[57] Aaron Halfaker and R. Stuart Geiger. 2020. ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–37. https://doi.org/10.1145/3415219

[58] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

[59] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature* 585, 7825 (Sept. 2020), 357–362. https://doi.org/10.1038/s41586-020-2649-2

[60] Galen Harrison, Kevin Bryson, Ahmad Emmanuel Balla Bamba, Luca Dovichi, Aleksander Herrmann Binion, Arthur Borem, and Blase Ur. 2024. JupyterLab in Retrograde: Contextual Notifications That Highlight Fairness and Bias Issues for Data Scientists. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–19. https://doi.org/10.1145/3613904.3642755

[61] Ewen Henderson. 2024. *ghibli: Studio Ghibli Colour Palettes*. https://CRAN.R-project.org/package=ghibli R package version 0.3.4.

[62] Courtney B. Hilton, Cody J. Moser, Mila Bertolo, Harry Lee-Rubin, Dorsa Amir, Constance M. Bainbridge, Jan Simson, Dean Knox, Luke Glowacki, Elias Alemu, Andrzej Galbarczyk, Grazyna Jasienska, Cody T. Ross, Mary Beth Neff, Alia Martin, Laura K. Cirelli, Sandra E. Trehub, Jinqi Song, Minju Kim, Adena Schachner, Tom A. Vardy, Quentin D. Atkinson, Amanda Salenius, Jannik Andelin, Jan Antfolk, Purnima Madhivanan, Anand Siddaiah, Caitlyn D. Placek, Gul Deniz Salali, Sarai Keestra, Manvir Singh, Scott A. Collins, John Q. Patton, Camila Scaff, Jonathan Stieglitz, Silvia Ccari Cutipa, Cristina Moya, Rohan R. Sagar, Mariamu Anyawire, Audax Mabulla, Brian M. Wood, Max M. Krasnow, and Samuel A. Mehr. 2022. Acoustic Regularities in Infant-Directed Speech and Song across Cultures. *Nature Human Behaviour* (July 2022), 1–12. https://doi.org/10.1038/s41562-022-01410-x

[63] Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Vol. 1. 278–282 vol.1. https://doi.org/10.1109/ICDAR.1995.598994

[64] Jake M. Hofman, Amit Sharma, and Duncan J. Watts. 2017. Prediction and Explanation in Social Systems. *Science* 355, 6324 (Feb. 2017), 486–488. https://doi.org/10.1126/science.aal3856

[65] 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. jstor:27594267

[66] Hsiang Hsu and Flavio Calmon. 2022. Rashomon Capacity: A Metric for Predictive Multiplicity in Classification. *Advances in Neural Information Processing Systems* 35 (Dec. 2022), 28988–29000.

[67] Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. 2024. Collective Constitutional AI: Aligning a Language Model with Public Input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1395–1417. https://doi.org/10.1145/3630106.3658979

[68] Frank Hutter, Holger Hoos, and Kevin Leyton-Brown. 2014. An Efficient Approach for Assessing Hyperparameter Importance. In *Proceedings of the 31st International Conference on Machine Learning*. PMLR, 754–762. https://proceedings.mlr.press/v32/hutter14.html

[69] Rashidul Islam, Shimei Pan, and James R. Foulds. 2021. Can We Obtain Fairness For Free?. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Virtual Event USA, 586–596. https://doi.org/10.1145/3461702.3462614

[70] Sofia Jaime and Christoph Kern. 2024. Ethnic Classifications in Algorithmic Fairness: Concepts, Measures and Implications in Practice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 237–253. https://doi.org/10.1145/3630106.3658902

[71] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 310–323. https://doi.org/10.1145/3531146.3533097

[72] Sara Kingsley, Jiayin Zhi, Wesley Hanwen Deng, Jaimie Lee, Sizhe Zhang, Motahhare Eslami, Kenneth Holstein, Jason I. Hong, Tianshi Li, and Hong Shen. 2024. Investigating What Factors Influence Users' Rating of Harmful Algorithmic Bias and Discrimination. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 12 (Oct. 2024), 75–85. https://doi.org/10.1609/hcomp.v12i1.31602

[73] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, Carol Willing, and Jupyter development team. 2016. Jupyter Notebooks - a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, Fernando Loizides and Birgit Scmidt (Eds.). IOS Press, Netherlands, 87–90. https://eprints.soton.ac.uk/403913/

[74] John Körtner and Giuliano Bonoli. 2023. Predictive algorithms in the delivery of public employment services. In *Handbook of Labour Market Policy in Advanced Democracies*. Edward Elgar Publishing, 387–398. https://doi.org/10.4337/9781800880887.00037

[75] Jon A. Krosnick. 1991. Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology* 5, 3 (1991), 213–236. https://doi.org/10.1002/acp.2350050305

[76] Bogdan Kulynych, David Madras, Smitha Milli, Inioluwa Deborah Raji, Angela Zhou, and Richard Zemel. 2020. Participatory Approaches to Machine Learning. International Conference on Machine Learning Workshop.

[77] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.

[78] Edward E. Leamer. 1985. Sensitivity Analyses Would Help. *The American Economic Review* 75, 3 (1985), 308–313. jstor:1814801

[79] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–35. https://doi.org/10.1145/3359283

[80] Timo Lenzner and Natalja Menold. 2016. Question Wording. GESIS survey guidelines.

[81] Bria Long, Jan Simson, Andrés Buxó-Lugo, Duane G. Watson, and Samuel A. Mehr. 2023. How Games Can Make Behavioural Science Better. *Nature* 613, 7944 (Jan. 2023), 433–436. https://doi.org/10.1038/d41586-023-00065-6

[82] Carol Long, Hsiang Hsu, Wael Alghamdi, and Flavio Calmon. 2023. Individual Arbitrariness and Group Fairness. *Advances in Neural Information Processing Systems* 36 (Dec. 2023), 68602–68624.

[83] Kristian Lum, Yunfeng Zhang, and Amanda Bower. 2022. De-Biasing "Bias" Measurement. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 379–389. https://doi.org/10.1145/3531146.3533105

[84] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–14. https://doi.org/10.1145/3313831.3376445

[85] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2021. On the Applicability of Machine Learning Fairness Notions. *SIGKDD Explor. Newsl.* 23, 1 (may 2021), 14–23. https://doi.org/10.1145/3468507.3468511

[86] Michele Mauri, Tommaso Elli, Giorgio Caviglia, Giorgio Uboldi, and Matteo Azzi. 2017. RAWGraphs: A Visualisation Platform to Create Open Outputs. In *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter* (Cagliari, Italy) *(CHItaly '17)*. ACM, New York, NY, USA, Article 28, 5 pages. https://doi.org/10.1145/3125571.3125585

[87] 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

[88] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6 (2021). https://doi.org/10.1145/3457607

[89] Michelle M. Mello and Sherri Rose. 2024. Denial—Artificial Intelligence Tools and Health Insurance Coverage Decisions. *JAMA Health Forum* 5, 3 (March 2024), e240622. https://doi.org/10.1001/jamahealthforum.2024.0622

[90] Lisa Messeri and M. J. Crockett. 2024. Artificial Intelligence and Illusions of Understanding in Scientific Research. *Nature* 627, 8002 (March 2024), 49–58. https://doi.org/10.1038/s41586-024-07146-0

[91] 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

[92] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, Atlanta GA USA, 220–229. https://doi.org/10.1145/3287560.3287596

[93] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8, 1 (2021), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

[94] Yuri Nakao, Simone Stumpf, Subeida Ahmed, Aisha Naseer, and Lorenzo Strappelli. 2022. Toward Involving End-users in Interactive Human-in-the-loop AI Fairness. *ACM Transactions on Interactive Intelligent Systems* 12, 3 (Sept. 2022), 1–30. https://doi.org/10.1145/3514258

[95] Nteract Contributors. 2017. Papermill: Parameterize and Run Jupyter and Nteract Notebooks.

[96] 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 (Oct. 2019), 447–453. https://doi.org/10.1126/science.aax2342

[97] Observable Team. [n. d.]. Observable: Build Expressive Charts and Dashboards with Code. https://observablehq.com/.

[98] Jeroen Ooms. 2014. The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. *arXiv:1403.2805 [stat.CO]* (2014). https://arxiv.org/abs/1403.2805

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

[100] Thomas Lin Pedersen. 2024. *patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork R package version 1.2.0.

[101] 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.

[102] Emma Pierson. 2018. Demographics and Discussion Influence Views on Algorithmic Fairness. https://doi.org/10.48550/arXiv.1712.09124 arXiv:1712.09124

[103] Marc Pinski and Alexander Benlian. 2023. AI Literacy - Towards Measuring Human Competency in Artificial Intelligence. In *Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2023.021

[104] Pipenv Maintainer Team. 2017. Pipenv: Python Development Workflow for Humans.

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

[106] Cathy Roche, Dave Lewis, and PJ Wall. 2021. Artificial Intelligence Ethics: An Inclusive Global Discourse? *arXiv preprint arXiv:2108.09959* (2021).

[107] Kit T. Rodolfa, Erika Salomon, Lauren Haynes, Iván Higuera Mendieta, Jamie Larson, and Rayid Ghani. 2020. Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 142–153. https://doi.org/10.1145/3351095.3372863

[108] Cynthia Rudin, Chudi Zhong, Lesia Semenova, Margo Seltzer, Ronald Parr, Jiachang Liu, Srikar Katta, Jon Donnelly, Harry Chen, and Zachery Boner. 2024. Amazing Things Come from Having Many Good Models. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[109] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. 2019. Aequitas: A Bias and Fairness Audit Toolkit. https://doi.org/10.48550/arXiv.1811.05577 arXiv:1811.05577 [cs]

[110] Lesia Semenova, Cynthia Rudin, and Ronald Parr. 2022. On the Existence of Simpler Machine Learning Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM, 1827–1858. https://doi.org/10.1145/3531146.3533232

[111] Hong Shen, Haojian Jin, Ángel Alexander Cabrera, Adam Perer, Haiyi Zhu, and Jason I. Hong. 2020. Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–22. https://doi.org/10.1145/3415224

[112] Hong Shen, Leijie Wang, Wesley H. Deng, Ciell Brusse, Ronald Velgersdijk, and Haiyi Zhu. 2022. The Model Card Authoring Toolkit: Toward Community-centered, Deliberation-driven AI Design. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 440–451. https://doi.org/10.1145/3531146.3533110

[113] 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 (Nov. 2011), 1359–1366. https://doi.org/10.1177/0956797611417632

[114] 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

[115] Jan Simson. 2024. Multiversum: A Helper Package to Conduct Multiverse Analyses in Python. https://pypi.org/project/multiversum/

[116] Jan Simson, Alessandro Fabris, and Christoph Kern. 2024. Lazy Data Practices Harm Fairness Research. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 642–659. https://doi.org/10.1145/3630106.3658931

[117] Jan Simson, Florian Pfisterer, and Christoph Kern. 2024. One Model Many Scores: Using Multiverse Analysis to Prevent Fairness Hacking and Evaluate the Influence of Model Design Decisions. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1305–1320. https://doi.org/10.1145/3630106.3658974

[118] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2022. Participation Is not a Design Fix for Machine Learning. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, Arlington VA USA, 1–6. https://doi.org/10.1145/3551624.3555285

[119] Christopher Small. 2021. Polis: Scaling Deliberation by Mapping High Dimensional Opinion Spaces. *RECERCA. Revista de Pensament i Anàlisi* (July 2021). https://doi.org/10.6035/recerca.5516

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

[121] Andy South. 2011. rworldmap: A New R package for Mapping Global Data. *The R Journal* 3, 1 (2011), 35–43. https://doi.org/10.32614/RJ-2011-006

[122] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2022. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society* 9, 2 (July 2022), 205395172211151. https://doi.org/10.1177/20539517221115189

[123] Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science* 11, 5 (2016), 702–712. https://doi.org/10.1177/1745691616658637 arXiv:https://doi.org/10.1177/1745691616658637 PMID: 27694465.

[124] Jan-Philipp Stein, Tanja Messingschlager, Timo Gnambs, Fabian Hutmacher, and Markus Appel. 2024. Attitudes towards AI: Measurement and Associations with Personality. *Scientific Reports* 14, 1 (Feb. 2024), 2909. https://doi.org/10.1038/s41598-024-53335-2

[125] Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3465416.3483305

[126] United States District Court, District of Minnesota. 2023. Lokken v UnitedHealth Group Inc. CASE: 0:23-cv-03514.

[127] Guido Van Rossum and Fred L. Drake Jr. 1995. *Python Tutorial*. Vol. 620. Centrum voor Wiskunde en Informatica Amsterdam.

[128] 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

[129] 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. *Journal of Open Source Software* 4, 43 (Nov. 2019), 1686. https://doi.org/10.21105/joss.01686

[130] World-Wide-Lab Developers. 2024. World-Wide-Lab. https://worldwidelab.org

[131] Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. 2022. Exploring the Whole Rashomon Set of Sparse Decision Trees. https://doi.org/10.48550/arXiv.2209.08040 arXiv:2209.08040 [cs]

[132] Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos. 2018. Grounding Interactive Machine Learning Tool Design in How Non-Experts Actually Build Models. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, Hong Kong China, 573–584. https://doi.org/10.1145/3196709.3196729

[133] Achim Zeileis, Jason C. Fisher, Kurt Hornik, Ross Ihaka, Claire D. McWhite, Paul Murrell, Reto Stauffer, and Claus O. Wilke. 2020. colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes. *Journal of Statistical Software* 96, 1 (2020), 1–49. https://doi.org/10.18637/jss.v096.i01

[134] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–23. https://doi.org/10.

Table 1: Distribution of participants across different countries. Due to the high number of countries, different countries with the same sample size are grouped into one row. Sample size and percentages apply to each individual country.

N	Percentage	Countries
199	37.27%	United States
37	6.93%	United Kingdom
31	5.81%	Canada
30	5.62%	Germany
27	5.06%	Australia
21	3.93%	France
10	1.87%	Aotearoa New Zealand; Brazil
9	1.69%	Russia
8	1.50%	Denmark
7	1.31%	Austria; Poland; Turkey
6	1.12%	India; South Korea
5	0.94%	Argentina; Belgium; Italy; Norway; Saudi Arabia; Sweden
4	0.75%	Hong Kong; Romania; Spain; Switzerland
3	0.56%	China; Iran; Ireland; Japan; Portugal; Serbia; Taiwan; The Netherlands
2	0.37%	Afghanistan; Albania; Andorra; Azerbaijan; Egypt; Finland; Greece; Hungary; Indonesia; Mexico; Philippines; Singapore; Vietnam
1	0.19%	Algeria; Angola; Armenia; Artsakh; Barbados; Bulgaria; Chad; Chile; Dominican Republic; Estonia; Georgia; Israel; Kenya; Kuwait; Latvia; Lithuania; Mongolia; Myanmar; Slovenia; South Africa; Sri Lanka; Thailand; Ukraine

1145/3274463

A Participants: Sample Composition

Participants self-reported a total of 69 different countries as their country of origin. The three most frequent countries are the United States (n = 199, 37.26%), the United Kingdom (n = 37, 6.92%) and Canada (n = 31, 5,80%). Participation rates by country are shown in Figure 3, with detailed numbers per country available in Table 1.

The majority of participants identified as male (n = 329, 61.61%), with 31.65% identifying as female (n = 169) and 6.74% identifying with another gender (n = 36). The average reported age is 29.11 years (*SD* = 15.56).

Participants from the United States were asked about their race and identified as predominantly White (n = 109, 54.77%), Asian (n = 38, 19.1%) or with more than one race (n = 18, 9.05%). Only a small number of people identified with other races (Black or African American n = 9, 4.52%; American Indian/Alaska Native n = 4, 2.01%; Native Hawaiian or other Pacific Islander n = 1, 0.5%) and about 10% of participants preferred not to answer the item (n = 20).

As racial identities are a highly complex topic and vary greatly within different geographical and social contexts [70], non-U.S. participants were not asked about their race but rather whether they identify as a minority in their respective country. Wording of the question was adapted from the European Social Survey. The majority of people did not identify as members of a minority (n = 260, 77.61%), with 14.03% of participants identifying with one (n = 47). For analyses using minority status, we also coded U.S. racial data into minority membership with the biggest group (White) coded as majority and all other racial groups as minorities.

B Research Materials

B.1 Introductory Text

In this last section, we have some questions for you about artificial intelligence (AI). We are building an AI system — and we need your help!

Our system is going to try to predict something important in the USA: **whether or not someone has public health insurance**. By "public health insurance", we mean free healthcare provided by the government, like Medicare or Veterans Health Administration coverage.

Here's how it will work. We'll give the AI some information (like a person's age and income) and it will learn how this information can reliably predict — or not — whether or not someone has health insurance. Then we'll see where its predictions went wrong, and help the AI to figure out how to make better predictions.

This could be really useful for helping to figure out who needs more support in getting health insurance. It could help local governments to look out for people in need.

Here's the problem: Sometimes, the people who design AI systems make bad decisions about how to train these models, and those decisions can lead to unfair, biased AI. **We want to know which sorts of decisions you think are good and which sorts you think are bad in designing an AI system.**

B.2 Introductory Text (Evaluation)

How should we evaluate whether an AI system is fair for different groups of people? The fairness of an AI system can be assessed with fairness metrics. These metrics allow quantifying the degree to which an AI system treats different groups of people equally based on sensitive characteristics like race, sex, or age. These

metrics evaluate how well a system's predictions or decisions align across these groups to identify and reduce biases.

We want to make our AI system as fair as we can across different **races** (i.e., the AI system's predictions should work equally across different groups of the characteristic "race").

B.3 Decisions

B.3.1 Exclude Subgroups. **Should the AI system analyze all groups of people, or only some groups of people?** When we work with data from different groups, especially when some groups are very small or uncommon, it can be challenging to decide how best to handle their data.

Sometimes, small groups are left out to protect people's privacy, because the data might not be reliable, or excluding them might make the data easier to analyze. But doing so means they are not represented in the data anymore.

Here are a set of race and ethnicity subgroups of the US population. **Which groups do you think should be included in the AI system?** (you can choose as many as you like)

Answering options were not randomized for this question. Order and options based on the ACS PUMS [23].

Variant 1: No Percentages

- ☐ **White alone:** Use data from everyone identifying mainly as White.
- ☐ **Black or African American alone:** Use data from everyone identifying mainly as Black or African American.
- ☐ **American Indian alone:** Use data from everyone identifying mainly as American Indian.
- ☐ **Alaska Native alone:** Use data from everyone identifying mainly as Alaska Native.
- ☐ **Anyone who indicated that they are American Indian and/or Alaska Native and specified a tribe:** Use data from everyone identifying as American Indian or Alaska Native and who specified information about their tribe.
- ☐ **Asian alone:** Use data from everyone identifying as Asian.
- ☐ **Native Hawaiian and Other Pacific Islander alone:** Use data from everyone identifying mainly as Native Hawaiian or Pacific Islander.
- ☐ **Some Other Race alone:** Use data from anyone identifying mainly with another race than the ones mentioned here.
- ☐ **Two or More Races:** Use data from anyone identifying with two or more races (biracial).
- ☐ I don't understand the description
- ☐ I prefer not to answer
- ☐ Suggest an alternative option: _____

Variant 2: With Percentages

The percentages correspond to the size of the race/ethnicity group in the dataset.[8]

- ☐ **White alone (56.7%):** Use data from everyone identifying mainly as White.
- ☐ **Black or African American alone (5.8%):** Use data from everyone identifying mainly as Black or African American.
- ☐ **American Indian alone (0.9%):** Use data from everyone identifying mainly as American Indian.
- ☐ **Alaska Native alone (<0.1%):** Use data from everyone identifying mainly as Alaska Native.
- ☐ **Anyone who indicated that they are American Indian and/or Alaska Native and specified a tribe (0.3%):** Use data from everyone identifying as American Indian or Alaska Native and who specified information about their tribe.
- ☐ **Asian alone (15.0%):** Use data from everyone identifying as Asian.
- ☐ **Native Hawaiian and Other Pacific Islander alone (0.4%):** Use data from everyone identifying mainly as Native Hawaiian or Pacific Islander.
- ☐ **Some Other Race alone (15.8%):** Use data from anyone identifying mainly with another race than the ones mentioned here.
- ☐ **Two or More Races (5.0%):** Use data from anyone identifying with two or more races (biracial).
- ☐ I don't understand the description
- ☐ I prefer not to answer
- ☐ Suggest an alternative option: _____

B.3.2 Exclude Features. **What kind of information should an AI system use?** Sometimes AI systems take into account potentially sensitive characteristics (like a person's sex or race) and sometimes these characteristics are excluded due to legal or privacy reasons. But excluding this information doesn't always make AI systems fairer, as these characteristics can be related to other information.

Which of these options do you think are acceptable? (you can choose more than one)

Answering options were not randomized for this question.

- ☐ **Do not exclude any sensitive characteristics:** This means the AI is trained with all available information, including sensitive characteristics like race and sex.
- ☐ **Exclude race from the system:** This means the AI uses all information available but **not** race.
- ☐ **Exclude sex from the system:** This means the AI uses all information available but **not** sex.
- ☐ **Exclude both race and sex from the system:** This means the AI uses all information available but **not** race and sex.
- ☐ I don't understand the description
- ☐ I prefer not to answer
- ☐ Suggest an alternative option: _____

B.3.3 Preprocess Income. **How should the AI system handle numbers?** When working with data that are numerical (like household income, or people's ages), it is often useful to *bin* these numbers into categories. This can make the data easier to understand and compare, but it also means the system is using less detailed information.

Which of these options do you think are acceptable for income data? (you can choose more than one)

- ☐ **No binning:** Keep the income data as it is.
- ☐ **Binning into bins of size $10,000:** Put each income into a group that covers ten thousand dollars, like $0-$9,999; $10,000-$19,999; and so on.

[8] *This text was erroneously displayed with brackets for a brief part of data collection (5.4% of sessions), before being updated. Previous version:* (the percentages correspond to the size of the race/ethnicity group in the dataset).

- □ **Binning into three evenly sized groups:** Divide all incomes into three equal groups: lower income, middle income and higher income.
- □ **Binning into four evenly sized groups:** Divide all incomes into four equal groups: lower income, middle income, upper middle income and higher income.
- □ I don't understand the description
- □ I prefer not to answer
- □ Suggest an alternative option: _____

B.3.4 Model. **How complicated should the AI system be?** One of the key decisions in designing an AI system is choosing the type of model to use. Below are some possibilities. (it's okay if you don't know how these work!)

Which of these options do you think are acceptable? (you can choose more than one)

- □ **Simple, more understandable model (Logistic Regression):** This type of model is easier to understand and interpret than many alternatives. It may not be able to learn as many relations as other, more powerful models.
- □ **Complex, more flexible model (Random Forest):** This type of model is able to learn intricate relations in the data, but is harder to interpret and understand.
- □ I don't understand the description
- □ I prefer not to answer
- □ Suggest an alternative option: _____

B.3.5 Eval Fairness Definition. **How should we evaluate our AI system?** Two important metrics are available to evaluate AI systems. Since we can often not be perfect on both of them, we may have to focus on one in particular.

For an AI system that predicts whether a person has public health insurance, which is most important?

This decision allowed selecting only a single answering option.

- □ **Sensitivity:** It is more important that out of all people without public health insurance, the AI system correctly identifies as many as possible. This would minimize the number of people who really do not have public insurance who are incorrectly identified as having public insurance.
- □ **Precision:** It is more important that out of all people where the AI system thinks that they do not have health insurance, many individuals indeed have no insurance. This would minimize the number of people who really have public insurance, but are incorrectly identified as not having public insurance.
- □ I don't understand the description
- □ I prefer not to answer
- □ Suggest an alternative option: _____

B.3.6 Eval Fairness Grouping. **How should our data be grouped?** When checking whether our AI system is fair or not, we need to choose how to group the data. This could influence how we calculate whether the AI system is fair or not.

Which of these options do you think are acceptable? (you can choose more than one)

- □ **Two Groups:** Create just two groups: the largest group and all other groups combined. We then check whether the AI system works equally well for these two groups. The difference between these two groups is used as the fairness metric.
- □ **All Group Comparisons:** Consider all the different groups separately without combining them. We then check whether the AI system works equally well for each possible pair of groups. The largest difference between any two groups is used as the fairness metric.
- □ I don't understand the description
- □ I prefer not to answer
- □ Suggest an alternative option: _____

B.3.7 Eval on Subset. **Which people should an AI system be tested on?** It's not practical to test out an AI system on everybody in a whole country, but it's also hard to choose who it should be tested on.

For an AI system that predicts whether a person has public health insurance, Which of these options do you think are acceptable? (you can choose more than one)

- □ **Collecting data from the most populous area:** This means testing the AI system using data from the area with the most people living in it, like a big city.
- □ **Collecting data from the area where the most people have public health insurance:** This means testing the AI system using data from an area where lots of people have public health insurance, but a few people don't have public health insurance.
- □ **Collecting data from the closest major city:** This means evaluating the AI system using data only from a city close-by to the people building the AI system.
- □ **Collecting data from as many people as possible, but excluding military veterans:** Being a veteran can impact healthcare needs, so it might change the AI's predictions to test the model on veterans. This option means testing the AI system only using data from non-veterans living in the area where the AI is being tested.
- □ **Collecting data of only U.S. citizens:** This means testing the AI system using data from U.S. citizens and not other people living in the area where the AI is being tested.
- □ **Collecting data from the overall population:** This means testing the AI in a similar way to how political polls are conducted, by studying a representative sample of people in the US.
- □ I don't understand the description
- □ I prefer not to answer
- □ Suggest an alternative option: _____

B.3.8 Eval Exclude Subgroups. When creating an AI system, one might exclude certain smaller groups from the data to simplify the process or with the intention to protect their privacy. There already was an earlier question about this regarding the exclusion of data from **certain** groups when creating the AI system.

This decision now is about including or excluding the same small groups when **evaluating** how good the AI system works.

Which of these options do you think are acceptable? (you can choose more than one)

This decision was only displayed if (1) an answer for Exclude Subgroups was provided, (2) that answer was one of the valid options and (3) not all options of Exclude Subgroups were selected.

- □ **Keep all groups for evaluation:** This means evaluating the AI system with data from all groups, also ones that were excluded earlier.
- □ **Exclude the same groups during evaluation:** This means using only data from the groups that were also included when creating the AI, excluding data from the same groups that were excluded earlier.
- □ I don't understand the description
- □ I prefer not to answer
- □ Suggest an alternative option: _____

C Software used in Analyses

Analyses were conducted with R version 4.2.2 [105] using packages from the tidyverse [129] with support of multiple other packages [4, 44, 48, 61, 98, 100, 121, 133].

The complete multiverse of decisions was simulated and explored with Python version 3.8 [127], using pandas [99] for data manipulation and scikit-learn [101] for modeling alongside multiple other packages [12, 40, 59, 73, 95, 104]. Diagrams of the multiverse were generated using RawGraphs [86], d3 [19] and Observable [97].

D Supplementary Tables and Figures

This section contains supplementary tables, figures and information on statistical analyses described in the main body of this work.

Tables with statistical details for group comparisons contain test details for each comparison which was calculated. Odds ratios are provided for comparisons between two groups. The column *Sig. Threshold* refers to Bonferroni corrected significance thresholds for a p-value of $\alpha = 0.05$.

Table 2: Overview of the two decision blocks, the actual decisions examined in the case study and their respective options. For the decision *Exclude Subgroups* the combination of options was used. The decision *Eval Fairness Definition* allowed choosing only one option. For the participatory input, each decision includes three additional *other* options: "I don't understand the description," "I prefer not to answer" and "Suggest an alternative option".

Block	Decision	Options
Model Design Decisions (Section 3.1.1)		
Data Selection	*Exclude Subgroups*	(1) white-alone; (2) black-or-african-american-alone; (3) american-indian-alone; (4) alaska-native-alone; (5) american-indian-and-or-alaska-native-and-tribe; (6) asian-alone; (7) native-hawaiian-and-other-pacific-islander-alone; (8) some-other-race-alone; (9) two-or-more-races
	Exclude Features	(1) none; (2) race; (3) sex; (4) race-sex
Preprocessing	*Preprocess Income*	(1) none; (2) bins-10000; (3) quantiles-3; (4) quantiles-4
Modeling	*Model*	(1) simple; (2) complex
Evaluation Decisions (Section 3.1.2)		
Metric	*Eval Fairness Definition*	(1) sensitivity; (2) precision
Evaluation	*Eval Fairness Grouping*	(1) majority-minority; (2) race-all
	Eval On Subset	(1) locality-largest-only; (2) locality-most-privileged; (3) locality-city; (4) exclude-military; (5) exclude-non-citizens; (6) full
	Eval Exclude Subgroups	(1) keep-in-eval; (2) exclude-in-eval

Table 3: Statistical details of group comparisons for the decision *Exclude Subgroups*, comparing different groups by *Country Of Residence*.

Option	Odds Ratio	p-value	Sig. Threshold
White (56.7%)	0.55	0.0018	0.0038
Some Other Race (15.8%)	0.64	0.0224	0.0038
Asian (15.0%)	0.62	0.0123	0.0038
Black/African American (5.8%)	0.61	0.0119	0.0038
Two Or More Races (5.0%)	0.59	0.0064	0.0038
American Indian (0.9%)	0.60	0.0104	0.0038
Nat. Hawaiian & Pac. Islander (0.4%)	0.62	0.0143	0.0038
Am. Indian & Alaska Native Incl. Tribe (0.3%)	0.82	0.3244	0.0038
Alaska Native (<0.1%)	0.84	0.4247	0.0038
Dont Understand	1.81	0.2278	0.0038
Prefer No Answer	1.51	0.1003	0.0038
Suggest Alternative	1.40	0.2973	0.0038
None Checked	1.06	0.8392	0.0038

Table 4: Statistical details of group comparisons for the decision *Exclude Subgroups*, based on whether percentages of the relative size of each subgroup were visible.

Option	Odds Ratio	p-value	Sig. Threshold
White (56.7%)	1.12	0.5836	0.0038
Some Other Race (15.8%)	1.27	0.2491	0.0038
Asian (15.0%)	1.09	0.7100	0.0038
Black/African American (5.8%)	1.07	0.7794	0.0038
Two Or More Races (5.0%)	1.21	0.3468	0.0038
American Indian (0.9%)	1.09	0.7030	0.0038
Nat. Hawaiian & Pac. Islander (0.4%)	0.98	1.0000	0.0038
Am. Indian & Alaska Native Incl. Tribe (0.3%)	0.95	0.8488	0.0038
Alaska Native (<0.1%)	0.95	0.8472	0.0038
Dont Understand	2.20	0.0786	0.0038
Prefer No Answer	1.03	1.0000	0.0038
Suggest Alternative	0.96	1.0000	0.0038
None Checked	0.79	0.2803	0.0038

Table 5: Statistical details of group comparisons for the decision *Exclude Features*, comparing different groups of the attribute *gender*.

Option	p-value	Sig. Threshold
None	0.0112	0.0063
Race	0.7743	0.0063
Sex	0.2613	0.0063
Race Sex	0.2362	0.0063
Dont Understand	0.7586	0.0063
Prefer No Answer	0.0685	0.0063
Suggest Alternative	0.0152	0.0063
None Checked	0.9346	0.0063

Table 6: Statistical details of group comparisons for the decision *Exclude Features*, comparing different groups of the attribute minority status. Minority status was self-reported for participants outside of the U.S. and computed based on majority-group membership based on self-reported race for U.S. participants.

Option	p-value	Sig. Threshold
None	0.0132	0.0063
Race	0.0929	0.0063
Sex	0.4921	0.0063
Race Sex	0.7785	0.0063
Dont Understand	0.4599	0.0063
Prefer No Answer	0.0680	0.0063
Suggest Alternative	0.7344	0.0063
None Checked	0.6405	0.0063

Table 7: Statistical details of group comparisons for the decision *Eval Fairness Definition*, comparing three equally-sized groups based on self-reported *AI Literacy*.

Option	p-value	Sig. Threshold
Sensitivity	0.2756	0.0083
Precision	0.4475	0.0083
Dont Understand	0.3651	0.0083
Prefer No Answer	0.5722	0.0083
Suggest Alternative	0.0009	0.0083
None Checked	0.0002	0.0083

Table 8: Statistical details of group comparisons for the option *prefer not to answer*, comparing three equally-sized groups based on self-reported *AI Attitudes*.

Decision	p-value	Sig. Threshold
Exclude Subgroups	0.1427	0.0063
Exclude Features	0.0290	0.0063
Model	0.0065	0.0063
Preprocess Income	0.0619	0.0063
Eval Fairness Definition	0.5947	0.0063
Eval Fairness Grouping	0.9266	0.0063
Eval On Subset	0.4232	0.0063
Eval Exclude Subgroups	0.2421	0.0063

Table 9: Statistical details of group comparisons for the option *none checked*, comparing three equally-sized groups based on self-reported *AI Attitudes*.

Decision	p-value	Sig. Threshold
Exclude Subgroups	0.0128	0.0063
Exclude Features	0.0051	0.0063
Model	0.0053	0.0063
Preprocess Income	0.0064	0.0063
Eval Fairness Definition	0.0141	0.0063
Eval Fairness Grouping	0.0514	0.0063
Eval On Subset	0.0476	0.0063
Eval Exclude Subgroups	0.3343	0.0063

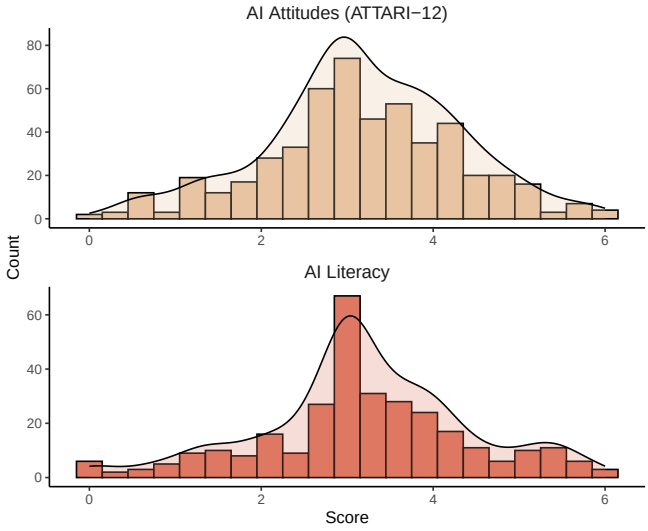


Figure 13: Histograms showing the overall distribution of AI attitudes [124] (above) and AI literacy [103] (below) scores across participants. Both scales emit a high degree of variation in the present sample, with a slight tendency towards more positive AI attitudes and higher AI literacy.

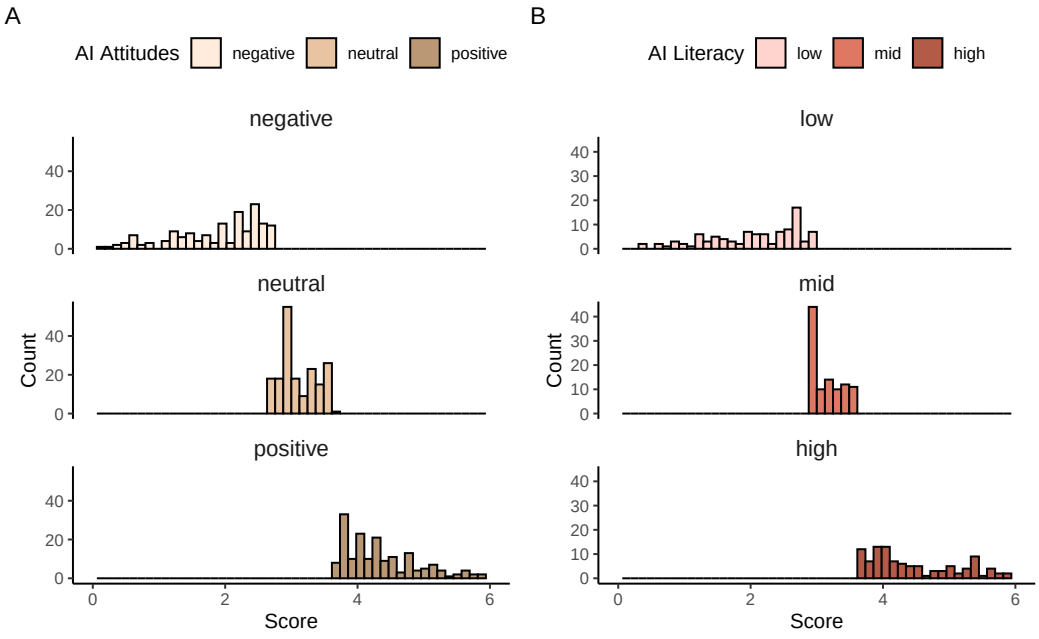


Figure 14: Histograms showing the distribution of AI attitudes (A) and AI literacy (B) scores across the three equally sized groups per metric, which were used for later group comparisons. The overall distribution of both scales is shown in Figure 13.

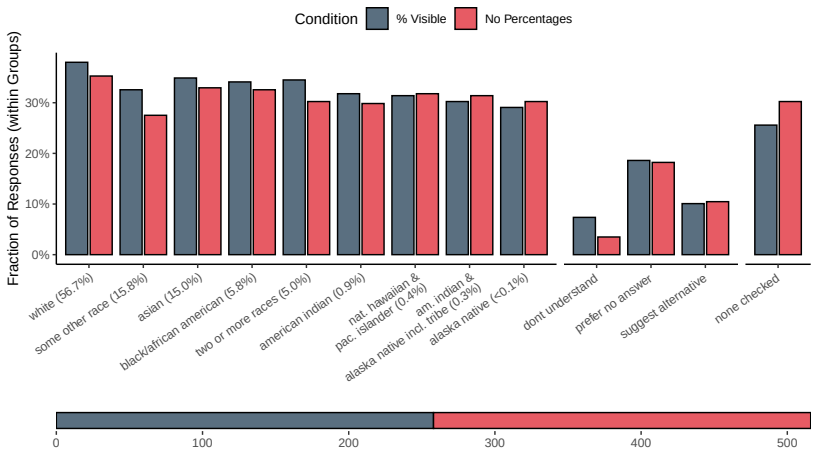


Figure 15: Inclusion of subgroups split by whether or not percentages were displayed next to groups for the decision *Exclude Subgroups*. The bar below the plot indicates the raw group distribution and number of votes.

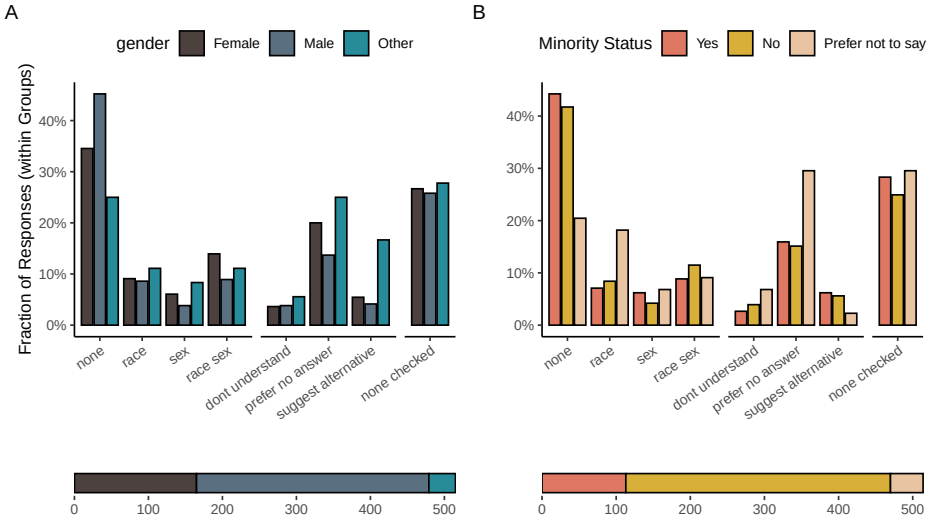


Figure 16: Exclusion of sensitive features split by gender (A) and minority status (B) for the decision *Exclude Features*. Bars below plot indicate the raw group distribution and number of votes.

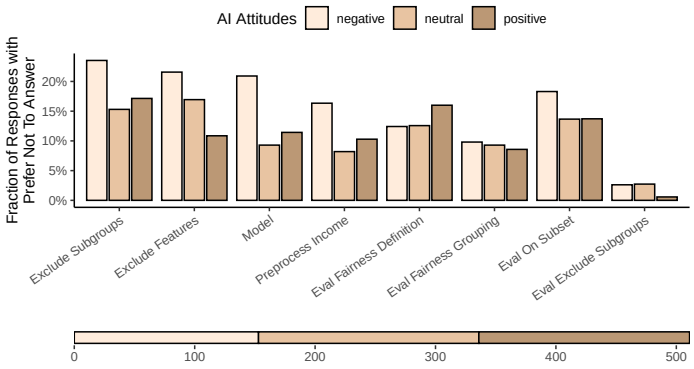


Figure 17: Fraction of participants choosing *"I Prefer Not To Answer"* across decisions split by self-reported AI attitudes. The bar below the plot indicates the raw group distribution and number of votes.

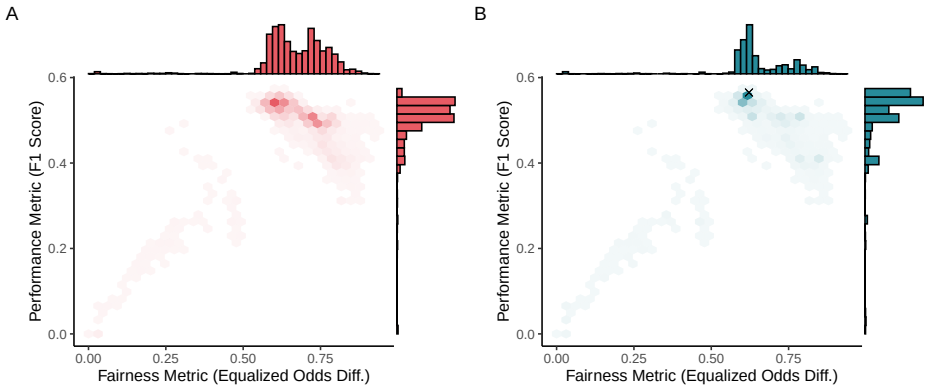


Figure 18: Comparison between complete multiverse of models (A) and one based on participants' votes (B), both evaluated using a fixed strategy with equalized odds difference as fairness metric and F1 score as performance metric. Darker areas correspond to a higher clustering of models. Cross indicates the most popular model among participants.

Taming the Machine Learning Multiverse Using Participatory Design (#185870)

Author(s)
Jan Simson (LMU Munich, Munich Center for Machine Learning (MCML)) - jan.simson@lmu.de
Fiona Draxler (University of Mannheim) - fiona.draxler@ifi.lmu.de
Samuel Mehr (University of Auckland, Yale University) - mehr@hey.com
Christoph Kern (LMU Munich, Munich Center for Machine Learning (MCML)) - christoph.kern@stat.uni-muenchen.de

Pre-registered on: 08/08/2024 08:12 AM (PT)

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
The study hypothesizes that participatory input from the general public can be a helpful tool in constraining the garden of forking paths encountered during the design of machine learning (ML) pipelines.

Specifically, it introduces a workflow of translating ML design decisions into a language that empowers laypeople to give input on the design of an ML system. The main question of this study is whether and to which degree this workflow is usable and provides high quality results.

3) Describe the key dependent variable(s) specifying how they will be measured.
The key dependent variable will be respondents' answers to multiple ML pipeline design decisions, including decisions in regards to pre-processing and evaluation.

The wording and options presented in design decisions may be changed throughout the duration of the study, based on respondents' input (see Section 5).

4) How many and which conditions will participants be assigned to?
Respondents will be presented with design decisions in blocks following the order these decisions would be made during the design of an ML system. Decisions in the same block are presented in randomized order.

In the unlikely scenario that there is a surprisingly high degree of dropout during the study, respondents may randomly be shown only a subset of the complete survey to reduce response burden.

For one of the decisions (ExcludeSubgroups) respondents will be randomly assigned to one of two conditions, varying whether or not they are shown percentages of group size next to the options.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
The main focus of analyses will be the degree of (dis)agreement of respondents on their responses to particular decisions. We will also examine whether respondents indicate that they do not understand a decision's description, decide not to respond or suggest an alternative option. We will examine these in particular after weeks one and two of data collection to update decisions if necessary. These analyses will be descriptive.

We will further investigate whether respondents' answers differ significantly based on their attitudes towards and literacy of artificial intelligence and socio-demographic characteristics. Depending on the observed distribution of scores on the AI scales we will separate them either into two (high/low) or three (high/medium/low) groups.

We present respondents with a case study based in the U.S. and will therefore examine whether data from U.S. respondents differs from those of other countries; the nuance of these analyses will be dependent on the final distribution of countries and responses per country.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We will exclude trials from any respondents where the response time was less than 2s for the decision items, as each of these items includes sufficient text that reading within 2s would be implausible.

Depending on the amount of data available, we may restrict analyses to only use data from respondents who have responded for all decisions (i.e. who did not drop out of the study before the end of this part).

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
The sample size will be determined by the amount of organic traffic our study receives. We aim to collect data for a duration of up to one month.

If we do not recruit enough data (N < 50), we will use a paid data provider.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
Nothing else to pre-register.

7. Bias Begins with Data: The FairGround Corpus for Robust and Reproducible Research on Algorithmic Fairness

Contributing article

Simson, J., Fabris, A., Fröhner, C., Kreuter, F. & Kern, C. (2024). Bias Begins with Data: The FairGround Corpus for Robust and Reproducible Research on Algorithmic Fairness. *Preprint*. doi: 10.48550/arXiv.2510 URL https://doi.org/10.48550/arXiv.2510.22363

Website

https://reliable-ai.github.io/fairground/

Package on PyPI

https://pypi.org/project/fairml-datasets/

Code repositories

Package and Website: https://github.com/reliable-ai/fairground/

Analyses: https://github.com/reliable-ai/fairground-analysis/

Author contributions

J. Simson provided the initial idea for the collaboration and project. J. Simson designed the annotation process and C. Fröhner performed most annotations in the work. A. Fabris reviewed a random subset of annotations and provided detailed comments. J. Simson implemented the Python package, online resources and empirical experiments in the work; analyzed the data and created all figures and tables. J. Simson lead the writing, submission and revision process of the work. C. Kern particularly contributed to the introduction (Section 1), A. Fabris particularly contributed to related work (Section 2) and with detailed edits across the manuscript. C. Fröhner particularly contributed to the section on computed metadata (C.6). All authors contributed through fruitful comments, proofreading and revisions of the manuscript.

A. Szimmat and F. Weber, mentioned in the acknowledgements of the paper provided help with the annotation process.

Bias Begins with Data: The *FairGround* Corpus for Robust and Reproducible Research on Algorithmic Fairness

Jan Simson JAN.SIMSON@LMU.DE
Department of Statistics, LMU Munich
Munich Center for Machine Learning (MCML)
Munich, 80539, Germany

Alessandro Fabris ALESSANDRO.FABRIS@UNITS.IT
University of Trieste
Trieste, 33170, Italy

Cosima Fröhner C.FROEHNER@CAMPUS.LMU.DE
Department of Statistics, LMU Munich
Munich, 80539, Germany

Frauke Kreuter FRAUKE.KREUTER@LMU.DE
Department of Statistics, LMU Munich
Munich Center for Machine Learning (MCML)
Munich, 80539, Germany

Christoph Kern CHRISTOPH.KERN@LMU.DE
Department of Statistics, LMU Munich
Munich Center for Machine Learning (MCML)
Munich, 80539, Germany

Abstract

As machine learning (ML) systems are increasingly adopted in high-stakes decision-making domains, ensuring fairness in their outputs has become a central challenge. At the core of fair ML research are the datasets used to investigate bias and develop mitigation strategies. Yet, much of the existing work relies on a narrow selection of datasets–often arbitrarily chosen, inconsistently processed, and lacking in diversity–undermining the generalizability and reproducibility of results.

To address these limitations, we present *FairGround*: a unified framework, data corpus, and Python package aimed at advancing reproducible research and critical data studies in fair ML classification. FairGround currently comprises 44 tabular datasets, each annotated with rich fairness-relevant metadata. Our accompanying Python package standardizes dataset loading, preprocessing, transformation, and splitting, streamlining experimental workflows. By providing a diverse and well-documented dataset corpus along with robust tooling, FairGround enables the development of fairer, more reliable, and more reproducible ML models. All resources are publicly available to support open and collaborative research.

Keywords: algorithmic fairness, dataset collections, dataset usage

1 Introduction

The field of algorithmic fairness has grown rapidly, reflecting the increasing recognition of fairness as a core concern in machine learning (Mehrabi et al., 2021; Pessach and Shmueli, 2023). Progress in this field is inevitably tied to data as the central ingredient to developing, testing and benchmarking more equitable algorithms. Given that these algorithms and fairness-enhancing techniques are often deployed in high-risk contexts [e.g., healthcare (Obermeyer et al., 2019; Barda et al., 2020), criminal justice (Angwin et al., 2016; Carton et al., 2016), jobseeker profiling (Kern et al., 2024; Achterhold et al., 2025)], systematic and transparent evaluations based on principled rather than ad-hoc selections of datasets are critical to understand which method works reliably under which conditions and which might not yet be ready for deployment.

Progress in Fair ML is challenged by (1) opacity in data practices and (2) critical limitations of the most prominent datasets currently used. A number of studies have shown that seemingly minor data processing and algorithmic design choices can significantly impact fairness outcomes, raising important questions about the robustness and generalizability of existing fairness interventions (Simson et al., 2024b; Friedler et al., 2019; Caton et al., 2022). Compounding these concerns, recent work has also highlighted reproducibility challenges that hinder consistent evaluation across settings (Cooper et al., 2024; Simson et al., 2024a). Furthermore, large-scale comparisons of fairness algorithms not only show strong sensitivity to data processing decisions, but also considerable performance differences between datasets, underlining the importance of the exact collection of data used for benchmarking and evaluation (Agrawal et al., 2021). Unfortunately, current studies commonly still focus on a narrow set of benchmark datasets–such as *Adult* (Kohavi, 1996), *COMPAS* (Angwin et al., 2016) and *German Credit* (Hofmann, 1994)–which suffer from known limitations, including contrived prediction tasks, noisy data, and severe coding mistakes (Ding et al., 2021; Bao et al., 2022; Grömping, 2019a). Taken together, these practices can lead to evaluations of fairness algorithms that are driven by methodological artifacts rather than representing reliable performance tests that justify the (non-)deployment of a given method in practice.

Addressing these limitations, this work introduces *FairGround*: a framework that emphasizes reproducible data processing pipelines, standardized evaluation protocols, and diverse collections of datasets tailored to specific needs (Figure 1). Our corpus contains 136 scenarios, i.e. combinations of 44 tabular datasets with available sensitive attributes. Each dataset comes with rich metadata (35 annotated and 27 computed meta-features), which allows for a principled selection of benchmarking collections and for failure testing of algorithms to identify data scenarios under which a proposed method struggles to perform. We further provide a data selection algorithm and associated collections of datasets that are small but diverse, i.e., present challenging scenarios with data that capture the variability present in the larger corpus. Our Python package facilitates transparent data practices in fair ML through reproducible and standardized, but customizable, processing pipelines. With FairGround, we contribute infrastructure that supports more robust and generalizable evaluation of fairness-aware machine learning methods.

2 Related work

2.1 Comparing fairness-enhancing algorithms

A number of prior studies have carried out systematic comparisons of fairness-enhancing algorithms across different settings (Friedler et al., 2019; Agrawal et al., 2021; Biswas and Rajan, 2020; Cruz and Hardt, 2024; Defrance et al., 2024; Han et al., 2023; Hort et al., 2021; Islam et al., 2021; L. Cardoso et al., 2019). While these comparative efforts have contributed valuable insights, they are often constrained by a narrow and inconsistently chosen set of benchmark datasets. In many cases, dataset selection is neither well-documented nor critically examined, resulting in evaluations that are difficult to reproduce and limited in scope.

The broader field continues to face fundamental challenges related to reproducibility and transparency in experimental design (Simson et al., 2024a; Cooper et al., 2024). One prominent issue is the lack of principled approaches to dataset processing and selection. Many existing works make ad hoc or arbitrary choices when selecting datasets (Ding et al., 2021; Bao et al., 2022; Grömping, 2019a), often relying on convenience or popularity rather than representativeness or relevance. These decisions can unintentionally bias results and restrict the generalizability of conclusions. A core concern here is that the datasets typically used in fairness evaluations do not adequately reflect the diversity and complexity of real-world deployment scenarios. The dominance of a small set of benchmark datasets has led to evaluations that cover only a limited subset of the problem space fairness algorithms are meant to address (Fabris et al., 2022).

Compounding this, there remains little clarity around the specific data conditions under which fairness methods are expected to succeed or fail. Without a systematic understanding of these contexts, practitioners are left with limited guidance on which algorithms to apply in practice (Richardson et al., 2021; Holstein et al., 2019), reducing the effectiveness and reliability of fairness interventions in real-world systems.

2.2 Fairness toolkits and data studies

Fairness datasets have been examined from both granular and comparative perspectives. Some works offer deep, dataset-specific critiques (Bandy and Vincent, 2021; Ding et al., 2021; Bao et al., 2022; Birhane et al., 2023), while others survey broader patterns across multiple datasets (Crawford and Paglen, 2021; Fabbrizzi et al., 2022; Fabris et al., 2022; Zhao et al., 2024). In parallel, fairness-focused toolkits such as AIF360 (Bellamy et al., 2018), Fairlearn (Weerts et al., 2023), and Aequitas (Jesus et al., 2024) implement popular algorithmic interventions and metrics, providing an accessible entry point for numerical comparisons—while including only a few illustrative datasets (Table A3). Despite overlapping goals, these two strands have remained largely disconnected. Toolkits often treat datasets as ancillary components and dataset-focused studies fail to produce machine-readable resources designed for seamless integration with software frameworks. Bridging critical data studies and fairness toolkits is essential for advancing the field, as meaningful integration can enable more rigorous, interpretable, and reproducible fairness evaluations–particularly by linking dataset properties to the behavior and impact of fairness interventions (Li et al., 2022; Favier et al., 2023).

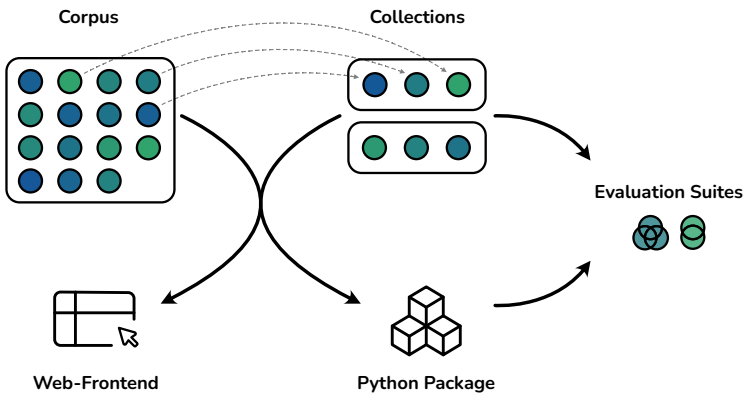


Figure 1: **The different components in the FairGround corpus.** We provide a comprehensive corpus of datasets and extract diverse collections of datasets via a selection algorithm. Both the corpus, collections and individual datasets are made accessible via a Python package and web-frontend. Collections paired with reproducible dataset loading and preparation allows for novel evaluation suites.

A recent article, most closely related to ours, lists several fairness resources and provides a tool for data fetching, but does not address integrated processing or annotation pipelines (Hirzel and Feffer, 2023). In this paper, we address this gap by introducing a benchmark suite that combines (1) a curated corpus of datasets accompanied by rich quantitative and qualitative annotations, (2) reproducible data fetching and processing pipelines, and (3) standardized collections and evaluation protocols. Our annotations provide a foundation for aligning datasets with fairness-aware methods in a consistent, reproducible, and extensible manner.

3 Framework

We introduce a unified framework of resources designed to support reproducible research and critical data studies in fair ML. While our current implementation focuses on tabular classification, which is prominent in fair ML research Mehrabi et al. (2021); Caton and Haas (2024), the underlying design is broadly applicable to other contexts.

3.1 Corpus

Building on and extending beyond prior surveys of datasets in fair ML research (Fabris et al., 2022; Le Quy et al., 2022), we compile a curated corpus of $N = 44$ tabular datasets. Each dataset is annotated with extensive fairness-relevant metadata, both quantitative and qualitative. While an additional 11 datasets were partially annotated, we excluded them from the final release due to issues such as dubious provenance or access restrictions (details in Section C.7).

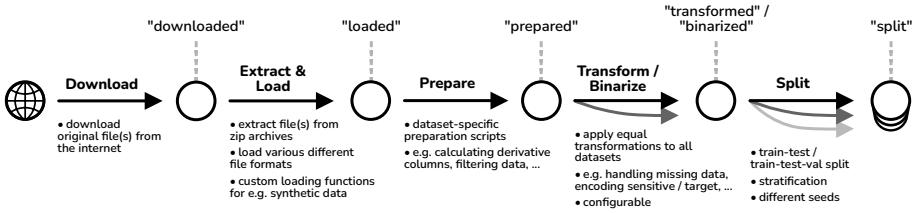


Figure 2: **The pipeline of steps involved when loading and processing a dataset in the package.** Datasets can be accessed / exported after each of the steps in the pipeline and most steps allow for configuration.

The corpus spans a wide range of dataset sizes, from 118 to over 3.2 million records, and 4 to 1,941 features. Most datasets originate from domains such as economics and law (each 23.4%), followed by finance (12.7%) and education (10.7%). Geographical representation is notably skewed: nearly 60% of datasets originate from the United States, with limited coverage from other regions (see Tables A and A2 for details). The dataset metadata can be explored interactively at: `https://reliable-ai.github.io/fairground/`.

Following prior work (Fabris et al., 2022; Le Quy et al., 2022), we annotate each dataset with contextual information (e.g., dataset name, domain), data-specific attributes (e.g., geography, time period), and technical metadata required for loading and preparing the data. Where multiple variants of a dataset exist, each version is treated as a distinct entry (cf. Section C.4). We also provide annotations relevant to fair ML tasks, including sensitive attribute selection, target variable definitions, and required preprocessing. While we do not claim our annotations are definitive, they serve as principled defaults that make implicit modeling decisions explicit, encouraging transparency in fair ML research (Simson et al., 2024a). Full details on our annotation procedure are provided in Sections C.4 and C.5.

In addition to manual annotations, we compute a range of metadata to support dataset selection, benchmarking, and critical analysis. This includes structural properties (e.g., missing values, feature types), statistical characteristics (e.g., bivariate correlations, sensitive AUC), and fairness-related properties (e.g., protected group prevalence, base rates, Gini-Simpson index) (Brzezinski et al., 2024; Mecati et al., 2022; Holland et al., 2020). These computed metadata features are detailed in Appendix C.6 and integrated into our Python tooling for streamlined access.

3.2 Infrastructure

To enable reproducible and scalable use of the corpus, we provide a Python package that operationalizes our framework. This package automates dataset acquisition, preprocessing, transformation, and splitting, applying the annotations to prepare datasets for downstream fair ML tasks (Figure 2). The package supports diverse data formats and includes default transformations, such as standard feature selection, handling of missing values and encoding

of sensitive attributes. We re-emphasize that defaults are not intended as universally correct, but rather as transparent baselines that can be fully customized. By surfacing and standardizing preprocessing decisions, the package encourages methodological rigor and reduces hidden variability in experimental pipelines (Simson et al., 2024b).

In particular, FairGround supports the following transformations to export data in a readily usable format. Users can choose to retain either the complete set of columns in a dataset or only the essential subset, which includes frequently-used features, sensitive attributes, and the target variable (default). To handle missing values, the framework supports three options: dropping the entire column, removing only rows with missing values, or imputing missing values using the median (default for numerical) or a placeholder value (default for categorical). The target variable can be binarized in several ways: based on an annotated preferable label, redefined to reflect a majority/minority split, or automatically selected between these options depending on metadata availability (default is based on the preferable label if provided). When multiple sensitive attributes exist, users can keep them separate (default) or combine them into a single binary attribute that captures their intersection (default in the binarized setting). Sensitive attribute values can be left unchanged or grouped into majority and minority categories (again, grouping is the default in binarized datasets). For categorical features, FairGround supports either leaving them as-is or converting them into binary indicators via dummy encoding (default). To control for high cardinality in categorical or text fields, the package applies an optional limit–by default, restricting each categorical or text column to a maximum of 200 unique values, with less frequent categories grouped together once this limit is exceeded.

The package also supports automatic metadata extraction (see Section 3.1). Importantly, we avoid redistributing raw data directly to respect licensing constraints and datasets are instead downloaded from their original sources and optionally cached locally.
The package is open source and available at: `https://github.com/reliable-ai/fairground`
Releases are archived on Zenodo: `https://doi.org/10.5281/zenodo.17288596`
Package installation: `pip install fairml-datasets`
Package documentation: `https://reliable-ai.github.io/fairground/docs/`
Code examples are provided in Appendix C.3.

In parallel, we release an interactive website that allows browsing the dataset corpus, metadata, and example usage. The site also offers sample code for specific datasets and is available at `https://reliable-ai.github.io/fairground/`.

3.3 Collections

To further support reproducible benchmarking and targeted experimentation, we define several curated dataset collections derived from the full corpus with an extensible algorithm. These include: two collections (small and large) optimized for diversity in algorithmic performance; three collections with permissive licenses; and three collections emphasizing geographic diversity (Tables A5–A7).

Combined with standardized data splits from our package, which are critical to fair ML reproducibility (Friedler et al., 2019), these collections provide ready-to-use evaluation suites for fair ML development.

4 Experiments

Leveraging the full FairGround dataset corpus, we conduct a series of experiments to systematically investigate the extent to which the choice of dataset influences the evaluation and observed performance of fairness-aware machine learning methods.

To reflect common practice in fairness research and enable broad coverage of methodological approaches, we evaluate a representative set of fairness-aware debiasing techniques spanning the three main intervention stages in the ML pipeline: *pre-processing*, *in-processing*, and *post-processing*. Specifically, we compare the following seven algorithms: *Learning Fair Representations* (pre) (Zemel et al., 2013), *Disparate Impact Remover* (pre) (Feldman et al., 2015), *Adversarial Debiasing* (in) (Zhang et al., 2018), *Meta-Algorithm* (in) (Celis et al., 2019), *Rich Subgroup Fairness / GerryFair* (in) (Kearns et al., 2018), *Grid Search Reduction* (in) (Agarwal et al., 2018), and *Group-Specific Thresholds* (post) (Hardt et al., 2016). We use logistic regression as a standard model for pre- and post-processing.

To satisfy the input constraints of all methods, datasets were converted to binarized numerical representations using the default transformation settings provided by our accompanying Python package (see Section 3.2). This ensures compatibility while preserving consistency across experiments.

Given that most fairness techniques are designed to optimize fairness with respect to a single sensitive attribute, we adopt a principled approach to define sensitive attribute configurations. For datasets containing fewer than four sensitive attributes, we evaluate all individual attributes and their pairwise intersections. For datasets with four or more sensitive attributes, we restrict evaluation to individual attributes to avoid combinatorial complexity. We refer to each combination of a dataset and its corresponding sensitive attribute selection as a unique ***scenario***.

We apply each of the seven processing methods and a baseline to each of the $n = 136$ datasets and sensitive attribute combinations (scenarios) across five separate seeds and train-test splits. This results in a total of $N = 5440$ different models that are trained and compared. For each model, we compute two commonly used measures of performance (*Balanced Accuracy*, Eq. 1; *F1 Score*, Eq. 2) and two measures of algorithmic fairness (*Equalized Odds Difference*, Eq. 3; *Demographic Parity Difference*, Eq. 4). The computational infrastructure (Section C.8) and software (Section C.9) used for experiments are described in the technical appendix.

4.1 Results

The experiments reveal substantial variation in both fairness and performance metrics across datasets and methods. F1 score, equalized odds difference, and demographic parity difference span the full $[0, 1]$ range, while balanced accuracy varies from approximately 0.2 to 1.0. To facilitate comparisons, we compute delta scores—metric differences relative to a logistic regression baseline without fairness interventions (Eq. 5). Figure 3 illustrates this calculation for one dataset, scenario, seed, and metric, with dashed lines indicating differences from the baseline.

The overall distribution of delta scores across all four metrics is shown in Figure B1. Importantly, fairness interventions often lead to minor deviations in scores, as highlighted by the large gray bar indicating an absolute change of ≤ 0.01, which correspond to scenarios

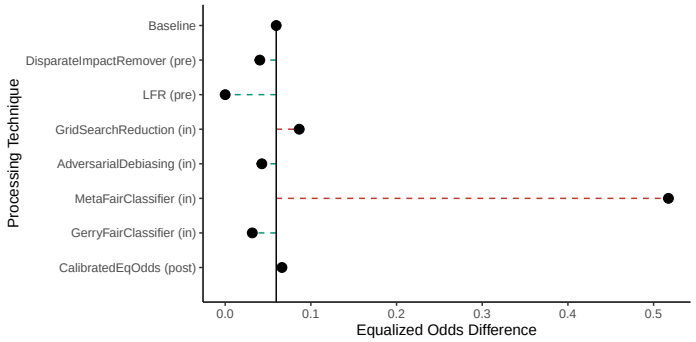


Figure 3: Scores from a single dataset (*Bank*), scenario (sensitive attribute: *Age*), seed (*80539*), and metric (*Equalized Odds Difference*) illustrating how delta scores with respect to baseline logistic regression are calculated. Delta scores correspond to dashed lines.

where popular fair ML methods are ineffective. A sizable portion produces meaningful differences, typically reflecting the well-known tradeoff between fairness and performance (Menon and Williamson, 2018; Islam et al., 2021): improvements in fairness often coincide with declines in predictive accuracy.

4.2 Rankings of Debiasing Techniques are not Stable

To reflect how practitioners might compare processing techniques in practice, we analyze the relative rankings of different methods. While some methods–such as *LFR*, *Grid Search Reduction*, and *Adversarial Debiasing*–tend to rank favorably, their positions vary considerably across scenarios, and no single method consistently outperforms the rest (Figure 4). High-performing methods often come with caveats. For instance, *LFR* occasionally fails due to convergence issues or label collapse during rebalancing, rendering it unusable in some cases. *Adversarial Debiasing* often presents sharp tradeoffs between fairness and predictive performance. These variations are influenced by the dataset and scenario characteristics.

4.3 Identifying Important Dataset Characteristics

To uncover which dataset properties affect method performance, we train simple machine learning models (random forests (Ho, 1995)) for each debiasing technique. These models use only computed metadata (Sections 3.1, C.6) to predict method effectiveness across individual scenarios. As shown in Figure B3, they capture substantial variance in observed outcomes. We analyze feature importance scores from these models to assess which dataset characteristics matter most. Figure 5 displays importances for predicting *Equalized Odds Difference*. A key trend is that the predictability of sensitive features from non-sensitive ones (`meta_sens_predictability_roc_auc`, top row) is influential across all methods. Base rate differences are critical for some techniques but negligible for others. These metadata-derived

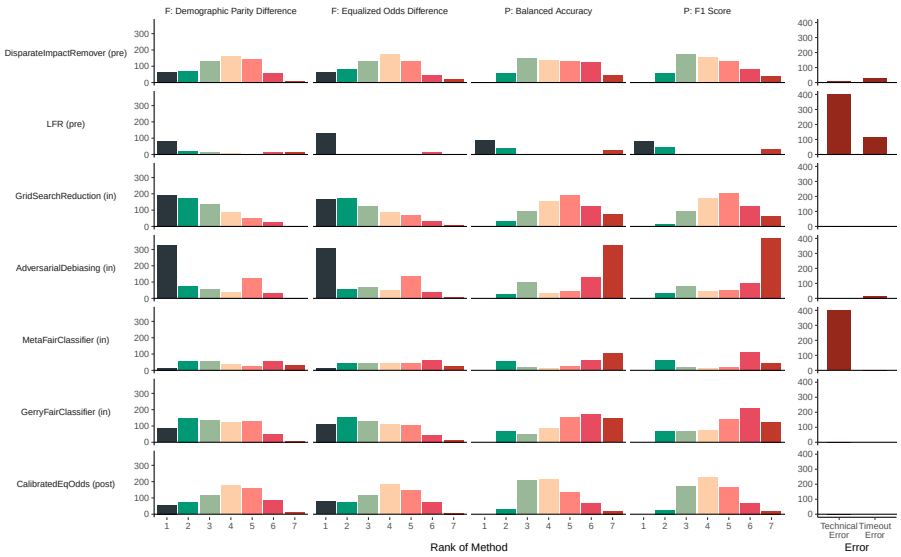


Figure 4: **Relative performance and efficacy of different fairness interventions is highly variable.** Relative ranking of different processing techniques across datasets and seeds (A), as well as prevalence of practical and timeout errors (B).

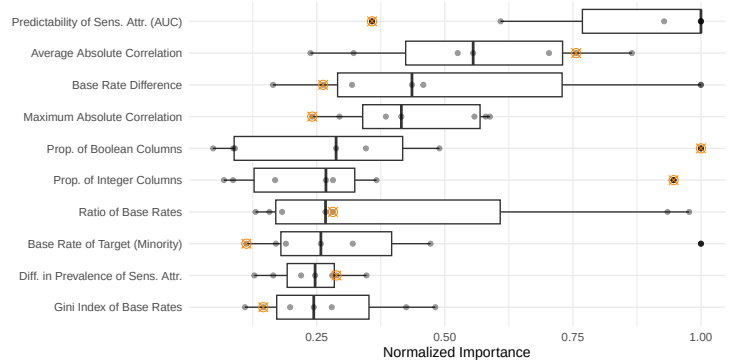


Figure 5: **The importance of different dataset characteristics can be highly variable between debiasing algorithms.** Normalized feature importance of the 10 most important computed metadata features to predict the difference in *Equalized Odds Difference* across all processing methods, ordered by average importance. Feature importance for *Adversarial Debiasing (in)* is highlighted in orange.

features help characterize the conditions under which fairness interventions are likely to succeed. Notably, *Adversarial Debiasing*, highlighted in orange, relies less on sensitive attribute predictability and more on structural features such as the proportion of boolean and integer columns. Relative importances for other metrics appear in Figure B2.

4.4 Developing Diverse Collections of Datasets

Evaluating fairness interventions across all possible datasets and scenarios is ideal but rarely feasible due to practical constraints like limited compute. To address this, we construct eight curated dataset collections, each optimized for a specific purpose. We use a principled algorithm to construct subsets of scenarios that exhibit diverse properties. We explicitly target predictive accuracy and fairness properties by building a collection of scenarios whose pairwise spearman correlations of delta scores (Eq. 5), across *Balanced Accuracy* (Eq. 1), *F1 Score* (Eq. 2), *Equalized Odds Difference* (Eq. 3) and *Demographic Parity Difference* (Eq. 4) are as low as possible. The underlying assumption is that datasets where debiasing techniques yield divergent fairness-performance tradeoffs make for more informative and challenging benchmarks. The algorithm greedily builds collections by adding the least correlated scenario while fulfilling optional secondary constraints, including selecting only a single scenario per dataset. The algorithm supports two different cutoff values, providing either a fixed number of k scenarios or a fixed upper bound for dataset correlation ($\bar{r}_{j\mathcal{C}} < \tau$) when added to the collection. The algorithm is described in detail in Section C.2. We use this selection process both to construct benchmark collections and to define default scenarios per dataset. We demonstrate how the *FairGround* corpus as well as its collections exhibit higher diversity in algorithm performance compared to other dataset collections (Table A4).

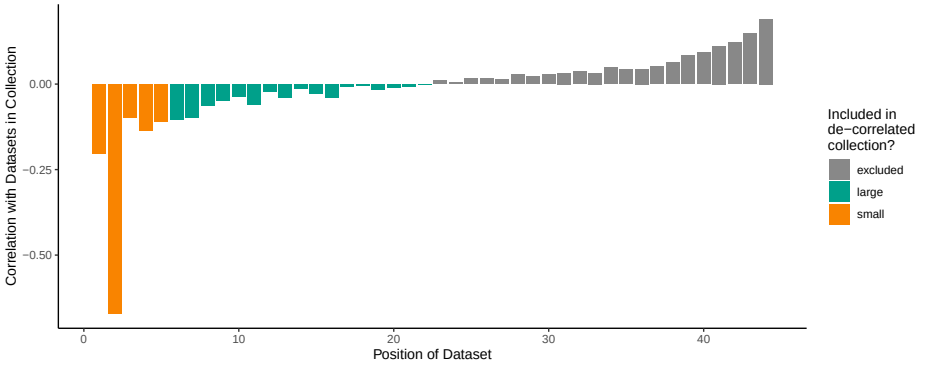


Figure 6: **A large number of negatively intercorrelated datasets is available for collection creation.** Average Spearman correlation of delta scores between the scenarios already in the collection and candidate scenarios at the time they are added to the collection. The very first scenario minimizes the average correlation with all other scenarios.

De-Correlated Datasets We construct two benchmark collections using the correlation-based algorithm with cutoffs $k = 5$ and $\tau < 0$, yielding sets of $n = 5$ and $n = 22$ scenarios, respectively (Table A5). A UMAP projection (McInnes et al., 2018) from the high-dimensional space of computed metadata (Figure 7) confirms that selected datasets span a wide range of characteristics.

Permissively Licensed Datasets To facilitate open sharing and reuse, we build three collections containing only datasets with permissive licenses. We construct these collections by only allowing datasets to be added to the collection which (1) have licensing information available and (2) are permissively licensed (e.g. Creative Commons, Apache, GNU licenses). One collection uses a fixed $k = 5$ cutoff, one uses a $\tau < 0$ threshold ($n = 16$), and one includes all permissively licensed datasets without filtering ($n = 32$). All three are listed in Table A6. We release these datasets in both prepared and binarized formats.

Geographically Diverse Datasets To address regional bias, we create three collections ensuring that no two datasets originate from the same country. We apply the selection algorithm with constraints and cutoffs of $k = 5$ and $\tau < 0$ ($n = 6$), as well as an unfiltered collection ($n = 10$) (Table A7). While this offers greater geographic diversity than is typical in ML fairness benchmarks, it remains insufficient. As prior work has emphasized (Septiandri et al., 2023; Mihalcea et al., 2025), future data efforts must expand beyond WEIRD contexts while carefully balancing this goal with ethical data practices.

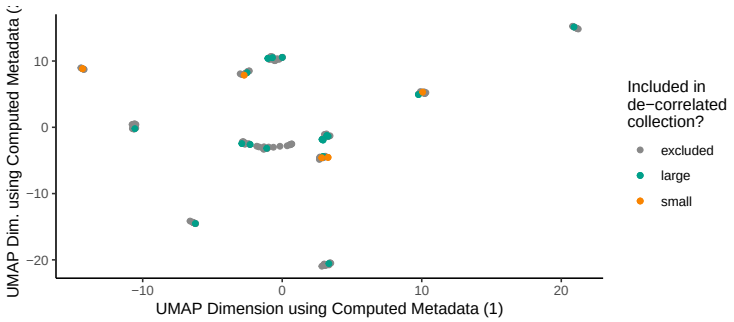


Figure 7: **Datasets in the de-correlated collections capture variability in the computed metadata features well.** Two dimensional mapping of datasets using UMAP on computed metadata features. Scenarios in the de-correlated collections are highlighted in different colors.

5 Limitations

While this work takes a substantial step toward improving reproducibility and empirical rigor in fair ML, it also operates within known constraints. Benchmarking, particularly in fairness research, can risk oversimplifying complex sociotechnical issues. Fairness cannot be fully captured by metrics or solved solely through optimization, and responsible development and evaluation of fair ML requires critical engagement with the broader context.

Our preprocessing and annotation decisions are not intended as universally optimal; their suitability depends on the specific dataset and use case. The experimental results presented here are illustrative rather than prescriptive—they demonstrate the kinds of analyses our corpus enables but are not meant to be definitive benchmarks.

Importantly, our dataset corpus is designed to be dynamic. Gaps in representation, especially with respect to geographic and demographic diversity, remain. We explicitly encourage community contributions of new datasets to help close these gaps (cf. Section 3.1). To support this, we provide a modular, versioned Python package that ensures transparency and reproducibility as the corpus evolves.

While our current focus is on tabular classification–a core setting in fair ML research–our framework is general. LLM evaluations, for example, may also benefit from the current tabular corpus through approaches such as folktexts (Cruz et al., 2025). In future work, we aim to extend our methodology and infrastructure to other data modalities, including text and image domains.

6 Discussion

We introduce *FairGround*, a comprehensive framework, dataset corpus, and Python package developed to address long-standing challenges in fair ML research. By curating a diverse collection of 44 tabular datasets, encompassing 136 scenarios, and providing fairness-relevant

metadata and reproducible preprocessing tools, FairGround enables transparent, rigorous, and extensible experimentation. The accompanying Python package supports reproducibility by exposing (and providing defaults for) key data processing decisions. We demonstrate its utility through a large-scale case study, illustrating how the framework facilitates robust comparative evaluations of debiasing techniques. Specifically, we show how our provided data collections better reflect the diverse performance of debiasing algorithms in comparison to collections currently used in fair ML research, while enabling new fairness analyses by connecting algorithm performance to dataset characteristics.

The significance of this work extends beyond its immediate technical contributions. By foregrounding the role of data infrastructure, FairGround highlights how dataset design, composition, and documentation fundamentally shape research trajectories and outcomes. These elements influence algorithmic behavior, reproducibility, and downstream system impact–making them critical to both scientific rigor and ethical responsibility.

Our framework is designed not only to support method development but also to position datasets as first-class research objects. It prompts researchers to interrogate representational biases, data provenance, and the implications of dataset selection–core concerns for equitable and socially responsible AI. In doing so, FairGround fosters deeper engagement with the sociotechnical dimensions of ML, encouraging reflection on how benchmarks reflect and reinforce power structures.

Additionally, FairGround lays essential groundwork for linking dataset characteristics to model fairness outcomes. This connection has important implications for anti-discrimination policy and regulation. For instance, under the EU AI Act, high-risk AI systems are subject to strict data governance requirements, including the obligation to assess datasets for bias and representational gaps (European Parliament, 2024). The metadata and fairness-relevant characteristics computed within FairGround can serve as a foundation for quantitative dataset documentation aligned with these legal mandates.

Broader Impact Statement

Our work aims to improve data practices in the field of algorithmic fairness, which in turn can lead to more robust and reproducible research, better and more ethical handling of datasets and increased transparency around dataset usage. By highlighting and quantifying the lack of geographic representation in popular datasets, we hope our work inspires the collection of novel and geographically diverse datasets. These positive changes have the possibility of affecting practices beyond research, ideally leading to the deployment of better and fairer algorithmic decision making and ML systems in production settings. Beyond the field of algorithmic fairness the *FairGround* framework provides a template for other fields to start developing dataset corpora and collections.

While this work encourages better data practices, there is a risk of it contributing to a benchmarking culture overly focused on quantitative and superficial notions of fairness, which we explicitly want to warn against. While it is important to use a diverse collection of datasets for evaluation, it is equally important, especially in applied contexts, to be aware of the sociotechnical context (ML) systems are developed and deployed in.

Acknowledgments and Disclosure of Funding

We would like to thank F. Weber and A. Szimmat for their help in the annotation process. The authors gratefully acknowledge the computational and data resources provided by the Leibniz Supercomputing Centre (www.lrz.de).

This work is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research and the Munich Center for Machine Learning (MCML).

This work is supported by BERD@NFDI and the Simons Institute for the Theory of Computing at the University of California, Berkeley.

References

Eva Achterhold, Monika Mühlböck, Nadia Steiber, and Christoph Kern. Fairness in algorithmic profiling: The AMAS case. *Minds and Machines*, 35(1):9, 2025.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR, 2018.

Agency for Healthcare Research and Quality. Medical expenditure panel survey (meps), 8 2018. URL `https://www.ahrq.gov/data/meps.html`.

Ashrya Agrawal, Florian Pfisterer, Bernd Bischl, Francois Buet-Golfouse, Srijan Sood, Jiahao Chen, Sameena Shah, and Sebastian Vollmer. Debiasing classifiers: is reality at variance with expectation?, 2021. URL `https://arxiv.org/abs/2011.02407`.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. *ProPublica*, 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, Published on May 23, 2016.

Pranjal Awasthi, Alex Beutel, Matthäus Kleindessner, Jamie Morgenstern, and Xuezhi Wang. Evaluating fairness of machine learning models under uncertain and incomplete information. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 206–214, 2021.

Jack Bandy and Nicholas Vincent. Addressing "documentation debt" in machine learning: A retrospective datasheet for bookcorpus. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. It's COMPASlicated: The messy relationship between RAI datasets and algorithmic fairness benchmarks, April 2022.

Noam Barda, Dan Riesel, Amichay Akriv, Joseph Levy, Uriah Finkel, Gal Yona, Daniel Greenfeld, Shimon Sheiba, Jonathan Somer, Eitan Bachmat, et al. Developing a COVID-19 mortality risk prediction model when individual-level data are not available. *Nature communications*, 11(1):4439, 2020.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018. URL https://arxiv.org/abs/1810.01943.

Abeba Birhane, Sanghyun Han, Vishnu Boddeti, Sasha Luccioni, et al. Into the laion's den: Investigating hate in multimodal datasets. *Advances in neural information processing systems*, 36:21268–21284, 2023.

Sumon Biswas and Hridesh Rajan. Do the machine learning models on a crowd sourced platform exhibit bias? An empirical study on model fairness. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2020, pages 642–653, New York, NY, USA, November 2020. Association for Computing Machinery. ISBN 978-1-4503-7043-1. doi: 10.1145/3368089.3409704.

Dariusz Brzezinski, Julia Stachowiak, Jerzy Stefanowski, Izabela Szczech, Robert Susmaga, Sofya Aksenyuk, Uladzimir Ivashka, and Oleksandr Yasinskyi. Properties of fairness measures in the context of varying class imbalance and protected group ratios. *ACM Transactions on Knowledge Discovery from Data*, 2024.

Samuel Carton, Jennifer Helsby, Kenneth Joseph, Ayesha Mahmud, Youngsoo Park, Joe Walsh, Crystal Cody, CPT Estella Patterson, Lauren Haynes, and Rayid Ghani. Identifying police officers at risk of adverse events. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 67–76, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939698. URL https://doi.org/10.1145/2939672.2939698.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):166:1–166:38, 2024. doi: 10.1145/3616865. URL https://doi.org/10.1145/3616865.

Simon Caton, Saiteja Malisetty, and Christian Haas. Impact of Imputation Strategies on Fairness in Machine Learning. *Journal of Artificial Intelligence Research*, 74, September 2022. ISSN 1076-9757. doi: 10.1613/jair.1.13197.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 319–328, 2019.

Chicago Data Portal. Strategic subject list - historical, 2020. URL https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List-Historical/4aki-r3np.

Consumer Financial Protection Bureau. Home mortgage disclosure act (hmda) data, September 2022. URL https://www.consumerfinance.gov/data-research/hmda/.

A. Feder Cooper, Katherine Lee, Madiha Zahrah Choksi, Solon Barocas, Christopher De Sa, James Grimmelmann, Jon Kleinberg, Siddhartha Sen, and Baobao Zhang. Arbitrariness and social prediction: The confounding role of variance in fair classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22004–22012, March 2024. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v38i20.30203.

Paulo Cortez and Alice Maria Gonçalves Silva. Using data mining to predict secondary school student performance. 2008.

Kate Crawford and Trevor Paglen. Excavating AI: the politics of images in machine learning training sets, 2021. URL `https://excavating.ai/`.

André F. Cruz and Moritz Hardt. Unprocessing Seven Years of Algorithmic Fairness, March 2024.

André F. Cruz, Moritz Hardt, and Celestine Mendler-Dünner. Evaluating language models as risk scores. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2025. Curran Associates Inc. ISBN 9798331314385.

MaryBeth Defrance, Maarten Buyl, and Tijl De Bie. ABCFair: an adaptable benchmark approach for comparing fairness methods. *Advances in Neural Information Processing Systems*, 37:40145–40163, December 2024.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring Adult: New Datasets for Fair Machine Learning. page 13, 2021. doi: 10.48550/ARXIV.2108.04884.

Michele Donini, Luca Oneto, Shai Ben-David, John S. Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2791–2801, 2018.

Martin Durant. fastparquet: A python interface to the parquet file format. https://pypi.org/project/fastparquet/.

European Parliament. Artificial intelligence act. `https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf`, 2024.

Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. A survey on bias in visual datasets. *Computer Vision and Image Understanding*, 223:103552, 2022.

Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, September 2022. ISSN 1573-756X. doi: 10.1007/s10618-022-00854-z.

Marco Favier, Toon Calders, Sam Pinxteren, and Jonathan Meyer. How to be fair? A study of label and selection bias. *Machine Learning*, 112(12):5081–5104, 2023. doi: 10.1007/S10994-023-06401-1. URL `https://doi.org/10.1007/s10994-023-06401-1`.

Jake Fawkes, Nic Fishman, Mel Andrews, and Zachary Lipton. The fragility of fairness: Causal sensitivity analysis for fair machine learning. *Advances in Neural Information Processing Systems*, 37:137105–137134, 2024.

Elaine Fehrman, Vincent Egan, and Evgeny Mirkes. Drug consumption (quantified). UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C5TC7S.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 329–338. Association for Computing Machinery, January 2019. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287589.

Ulrike Grömping. South german credit data: Correcting a widely used data set. *Rep. Math., Phys. Chem., Berlin, Germany, Tech. Rep*, 4:2019, 2019a.

Ulrike Grömping. South german credit. UCI Machine Learning Repository, 2019b. DOI: https://doi.org/10.24432/C5X89F.

H. Guvenir, Burak Acar, Haldun Muderrisoglu, and R. Quinlan. Arrhythmia. UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C5BS32.

Xiaotian Han, Jianfeng Chi, Yu Chen, Qifan Wang, Han Zhao, Na Zou, and Xia Hu. FFB: A fair fairness benchmark for in-processing group fairness methods. In *The Twelfth International Conference on Learning Representations*, October 2023.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016. URL `https://arxiv.org/abs/1610.02413`.

Martin Hirzel and Michael Feffer. A suite of fairness datasets for tabular classification. *arXiv preprint arXiv:2308.00133*, 2023.

Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282. IEEE, 1995.

Hans Hofmann. Statlog (german credit data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label. *Data Protection and Privacy*, 12(12):1, 2020.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík, and Hanna M. Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and

Vassilis Kostakos, editors, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 600. ACM, 2019. doi: 10.1145/3290605.3300830. URL https://doi.org/10.1145/3290605.3300830.

Max Hort, Jie M. Zhang, Federica Sarro, and Mark Harman. Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2021, pages 994–1006, New York, NY, USA, August 2021. Association for Computing Machinery. ISBN 978-1-4503-8562-6. doi: 10.1145/3468264.3468565.

Rashidul Islam, Shimei Pan, and James R. Foulds. Can we obtain fairness for free? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 586–596, Virtual Event USA, July 2021. ACM. ISBN 978-1-4503-8473-5. doi: 10.1145/3461702.3462614.

Shahin Jabbari, Han-Ching Ou, Himabindu Lakkaraju, and Milind Tambe. An empirical study of the trade-offs between interpretability and fairness. In *ICML Workshop on Human Interpretability in Machine Learning, International Conference on Machine Learning (ICML)*, 2020.

Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano. Heart disease data set. UCI Machine Learning Repository, 1988. URL https://archive.ics.uci.edu/ml/datasets/heart+Disease.

Sérgio Jesus, Pedro Saleiro, Beatriz M Jorge, Rita P Ribeiro, João Gama, Pedro Bizarro, Rayid Ghani, et al. Aequitas Flow: Streamlining fair ML experimentation. *arXiv preprint arXiv:2405.05809*, 2024.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pages 2564–2572. PMLR, 2018.

Christoph Kern, Ruben Bach, Hannah Mautner, and Frauke Kreuter. When small decisions have big impact: Fairness implications of algorithmic profiling schemes. *ACM Journal on Responsible Computing*, 1(4), November 2024. doi: 10.1145/3689485. URL https://doi.org/10.1145/3689485.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207, 1996.

Max Kuhn and Hadley Wickham. *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles*, 2020. URL https://www.tidymodels.org.

Rodrigo L. Cardoso, Wagner Meira Jr., Virgilio Almeida, and Mohammed J. Zaki. A Framework for Benchmarking Discrimination-Aware Models in Machine Learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pages 437–444,

New York, NY, USA, January 2019. Association for Computing Machinery. ISBN 978-1-4503-6324-2. doi: 10.1145/3306618.3314262.

Preethi Lahoti, Krishna P. Gummadi, and Gerhard Weikum. Operationalizing individual fairness with pairwise fair representations. arXiv preprint arXiv:1907.01439, 2019.

Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3):e1452, 2022.

Nianyun Li, Naman Goel, and Elliott Ash. Data-centric factors in algorithmic fairness. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 396–410, 2022.

Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating ML's impact disparity require treatment disparity? *Advances in Neural Information Processing Systems*, 31, 2018.

Debmalya Mandal, Samuel Deng, Suman Jana, Jeannette Wing, and Daniel J Hsu. Ensuring fairness beyond the training data. *Advances in neural information processing systems*, 33: 18445–18456, 2020.

Charlie Marsh. uv: An extremely fast python package and project manager, 2024. URL `https://github.com/astral-sh/uv`.

Fernando Martínez-Plumed, Cèsar Ferri, David Nieves, and José Hernández-Orallo. Fairness and missing values. *arXiv preprint arXiv:1905.12728*, 2019.

Norman Matloff and Wenxi Zhang. A novel regularization approach to fair ml. *arXiv preprint arXiv:2208.06557*, 2022.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Wes McKinney. Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, 445:51–56, 2010.

Mariachiara Mecati, Antonio Vetrò, and Marco Torchiano. Detecting risk of biased output with balance measures. *ACM J. Data Inf. Qual.*, 14(4):25:1–25:7, 2022. doi: 10.1145/3530787. URL `https://doi.org/10.1145/3530787`.

Mariachiara Mecati, Marco Torchiano, Antonio Vetrò, and Juan Carlos De Martin. Measuring imbalance on intersectional protected attributes and on target variable to forecast unfair classifications. *IEEE Access*, 11:26996–27011, 2023.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54 (6):1–35, 2021.

Aditya Krishna Menon and Robert C. Williamson. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 107–118. PMLR, January 2018.

Weiwen Miao. Did the results of promotion exams have a disparate impact on minorities? Using statistical evidence in Ricci v. DeStefano. *Journal of Statistics Education*, 18(3), 2010.

Rada Mihalcea, Oana Ignat, Longju Bai, Angana Borah, Luis Chiruzzo, Zhijing Jin, Claude Kwizera, Joan Nwatu, Soujanya Poria, and Thamar Solorio. Why AI is WEIRD and shouldn't be this way: Towards AI for everyone, with everyone, by everyone. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28657–28670, 2025.

Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

New York City Police Department. The nypd stop, question, and frisk database, 2012. URL `https://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page`.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453, 2019. doi: 10.1126/science.aax2342. URL `https://www.science.org/doi/abs/10.1126/science.aax2342`.

Thomas Lin Pedersen. *patchwork: The Composer of Plots*, 2024. URL `https://CRAN.R-project.org/package=patchwork`. R package version 1.3.0.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct): 2825–2830, 2011.

Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Comput. Surv.*, 55(3), February 2022. ISSN 0360-0300. doi: 10.1145/3494672. URL `https://doi.org/10.1145/3494672`.

Dana Pessach and Erez Shmueli. Algorithmic fairness. In *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*, pages 867–886. Springer, 2023.

Python Core Team. *Python: A dynamic, open source programming language*. Python Software Foundation, 2019. URL `https://www.python.org/`. Python version 3.10.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL `https://www.R-project.org/`.

Vladislav Rajkovic. Nursery. UCI Machine Learning Repository, 1989. DOI: https://doi.org/10.24432/C5P88W.

Karthik Ram and Hadley Wickham. *wesanderson: A Wes Anderson Palette Generator*, 2023. URL `https://CRAN.R-project.org/package=wesanderson`. R package version 0.3.7.

Michael Redmond. Communities and Crime. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C53W3X.

Brianna Richardson, Jean Garcia-Gathright, Samuel F. Way, Jennifer Thom, and Henriette Cramer. Towards fairness in practice: A practitioner-oriented rubric for evaluating fair ML toolkits. In Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Mark Drucker, editors, *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, pages 236:1–236:13. ACM, 2021. doi: 10.1145/3411764.3445604. URL `https://doi.org/10.1145/3411764.3445604`.

Sivan Sabato and Elad Yom-Tov. Bounding the fairness and accuracy of classifiers from population statistics. In *International conference on machine learning*, pages 8316–8325. PMLR, 2020.

Ali Akbar Septiandri, Marios Constantinides, Mohammad Tahaei, and Daniele Quercia. WEIRD FAccTs: How western, educated, industrialized, rich, and democratic is FAccT? In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 160–171, 2023.

Jan Simson. Multiversum: A helper package to conduct multiverse analyses in python, 2024. URL `https://github.com/jansim/multiversum`.

Jan Simson, Alessandro Fabris, and Christoph Kern. Lazy data practices harm fairness research. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 642–659, New York, NY, USA, June 2024a. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658931.

Jan Simson, Florian Pfisterer, and Christoph Kern. One model many scores: Using multiverse analysis to prevent fairness hacking and evaluate the influence of model design decisions. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 1305–1320, New York, NY, USA, June 2024b. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658974.

Antonio Vetrò, Marco Torchiano, and Mariachiara Mecati. A data quality approach to the identification of discrimination risk in automated decision making systems. *Government Information Quarterly*, 38(4):101619, 2021.

Hao Wang, Berk Ustun, and Flavio Calmon. Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In *International Conference on Machine Learning*, pages 6618–6627. PMLR, 2019.

Yanchen Wang and Lisa Singh. Analyzing the impact of missing values and selection bias on fairness. *International Journal of Data Science and Analytics*, 12(2):101–119, 2021.

Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and improving fairness of ai systems. *J. Mach. Learn. Res.*, 24(1), January 2023. ISSN 1532-4435.

Austin Wehrwein. *awtools: misc tools and themes for austinwehrwein.com*, 2025. URL `https://github.com/awhstin/awtools`. R package version 0.2.1.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

Linda F. Wightman. Lsac national longitudinal bar passage study. 1998.

I-Cheng Yeh. Default of Credit Card Clients. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C55S3H.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR, 2017.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

Dora Zhao, Morgan Klaus Scheuerman, Pooja Chitre, Jerone Theodore Alexander Andrews, Georgia Panagiotidou, Shawn Walker, Kathleen H. Pine, and Alice Xiang. A taxonomy of challenges to curating fair datasets. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL `http://papers.nips.cc/paper_files/paper/2024/hash/b142e78db191e19b17e60c1425a28b52-Abstract-Datasets_and_Benchmarks_Track.html`.

Appendix A. Supplementary Tables

This section includes supplementary tables that provide additional information supporting the results presented in the main text.

Table A1: Overview of datasets in the corpus. Row and column counts apply to the prepared data prior to further transformations.

	Name and Citation	Rows	Columns	License
1	Adult (Kohavi, 1996)	32,560	15	CC BY 4.0
2	Arrhythmia (Guvenir et al., 1998)	451	280	CC BY 4.0
3	Bank (additional + full) (Moro et al., 2014)	41,188	21	CC BY 4.0
4	Bank (additional) (Moro et al., 2014)	4,119	21	CC BY 4.0
5	Bank (full) (Moro et al., 2014)	45,211	17	CC BY 4.0
6	Bank (Moro et al., 2014)	4,521	17	CC BY 4.0
7	Communities (Redmond, 2009)	1,993	128	CC BY 4.0
8	Communities (unnormalized) (Lahoti et al., 2019)	2,214	147	CC BY 4.0
9	COMPAS (2 years) (Angwin et al., 2016)	6,172	53	?
10	COMPAS (2 years, violent) (Angwin et al., 2016)	4,743	54	?
11	COMPAS (Angwin et al., 2016)	11,757	47	?
12	CreditCard (Yeh, 2009)	30,000	25	CC BY 4.0
13	Drug (Fehrman et al., 2015)	1,885	32	CC BY 4.0
14	Dutch (Le Quy et al., 2022)	60,420	12	[a]
15	German Credit (Hofmann, 1994)	1,000	21	CC BY 4.0
16	German Credit (numeric) (Hofmann, 1994)	1,000	25	CC BY 4.0
17	South German Credit (Grömping, 2019b)	1,000	21	CC BY 4.0
18	German Credit (onehot) (Hofmann, 1994)	1,000	65	Apache License
19	Heart Disease (Janosi et al., 1988)	303	14	CC BY 4.0
20	HMDA (Consumer Financial Protection Bureau, 2022)	2,000,000	19	?
21	Law School (tensorflow) (Wightman, 1998)	22,407	39	CC BY-SA 4.0
22	Law School (LeQuy) (Wightman, 1998; Le Quy et al., 2022)	18,692	12	CC BY-SA 4.0
23	MEPS (Panel 19, FY2015) (Agency for Healthcare Research and Quality, 2018)	15,830	1,831	[b]
24	MEPS (Panel 20, FY2015) (Agency for Healthcare Research and Quality, 2018)	17,570	1,831	[b]
25	MEPS (Panel 21, FY2016) (Agency for Healthcare Research and Quality, 2018)	15,675	1,941	[b]

	Name and Citation	Rows	Columns	License
26	Nursery (Rajkovic, 1989)	12,960	9	CC BY 4.0
27	ricci (Miao, 2010)	118	5	?
28	Stop, Question and Frisk Data (New York City Police Department, 2012)	8,947	83	[c]
29	Chicago Strategic Subject List (Chicago Data Portal, 2020)	398,684	48	NA
30	Student (Cortez and Silva, 2008)	395	33	CC BY 4.0
31	Student (Language) (Cortez and Silva, 2008)	649	33	CC BY 4.0
32	generate_synthetic_data (Zafar et al., 2017)	2,000	4	GPL-3.0
33	Lipton synthetic hiring dataset (Lipton et al., 2018)	2,000	4	CC 0
34	synth (Donini et al., 2018)	6,400	4	?
35	Folktables ACSIncome (Ding et al., 2021)	1,664,500	11	CC 0
36	Folktables ACSPublicCoverage (Ding et al., 2021)	1,138,289	20	CC 0
37	Folktables ACSMobility (Ding et al., 2021)	620,937	22	CC 0
38	Folktables ACSEmployment (Ding et al., 2021)	3,236,107	17	CC 0
39	Folktables ACSTravelTime (Ding et al., 2021)	1,466,648	17	CC 0
40	Folktables ACSIncome (small) (Ding et al., 2021)	245,673	11	CC 0
41	Folktables ACSPublicCoverage (small) (Ding et al., 2021)	174,178	20	CC 0
42	Folktables ACSMobility (small) (Ding et al., 2021)	98,081	22	CC 0
43	Folktables ACSEmployment (small) (Ding et al., 2021)	478,236	17	CC 0
44	Folktables ACSTravelTime (small) (Ding et al., 2021)	216,385	17	CC 0

[a] Copyright 2001, Centraal Bureau voor de Statistiek (CBS) (Statistics Netherlands) and Minnesota Population Center.

[b] See `https://meps.ahrq.gov/data_stats/data_use.jsp`.

[c] "All rights reserved", see `https://www.nyc.gov/home/terms-of-use.page`.

Appendix B. Supplementary Figures

This section contains supplementary figures that complement the primary results and provide further context for the analyses discussed in the main manuscript.

Table A2: Countries represented in fair ML data. Each count represents a dataset that includes data from the specified country. There is one dataset representing data from across the world and one representing data from Hungary, Switzerland and the United States.

Country	Count	Percentage (%)
United States	28	59.57
Portugal	6	12.77
Germany	4	8.51
N/A	3	6.38
Hungary, Switzerland & United States	1	2.13
Netherlands	1	2.13
Slovenia	1	2.13
Taiwan	1	2.13
Turkey	1	2.13
World	1	2.13

Table A3: Quantitative comparison of datasets available in different fairness libraries. *FairGround allows for the input of any custom fairness methods by users.

		Number of		Meta-	Collec-
Library	**Main Focus**	**Datasets**	**Methods**	**Features**	**tions**
ABCFair	methods, metrics	7 (5)	10	✗	✗
Aequitas Flow	methods, metrics, guides	11 (11)	10	✗	✗
AIF360	methods, metrics	8 (8)	15	✗	✗
Fairlearn	methods, metrics, guides	6 (4)	6	✗	✗
FairGround (ours)	data	44	7*	✓	✓

Table A4: Comparison of dataset collections in FairGround and other work, showing whether a debiasing method is ever the best performing method for any of the datasets for Equalized Odds Difference (left) and Demographic Parity Difference (right). For outside collections the closest matching scenarios within FairGround are selected.

	FairGround							
	All	**Open (all)**	**Open (lg.)**	**Open (sm.)**	**ABC Fair[a]**	**AIF 360[b]**	**Fried-ler[c]**	**Typ. 3[d]**
DisparateImpactRemover (pre)	✓/ ✗	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✗/ ✓	✓/ ✓	✗/ ✗
LFR (pre)	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✗/ ✗	✓/ ✓	✓/ ✓	✓/ ✓
GridSearchReduction (in)	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓
AdversarialDebiasing (in)	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓
MetaFairClassifier (in)	✓/ ✓	✓/ ✓	✓/ ✓	✗/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓
GerryFairClassifier (in)	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✗/ ✗	✗/ ✗
CalibratedEqOdds (post)	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✗/ ✗	✗/ ✗	✗/ ✗
No. of Datasets	44	32	16	5	5	7	5	3

[a] Five out of seven datasets in ABCFair (Defrance et al., 2024) are used.
[b] Seven out of eight datasets in AIF360 (Bellamy et al., 2018) are used, the skipped dataset is available in FairGround, but used as a regression dataset in AIF360.
[c] Friedler et al. (2019)
[d] "Typical 3" refers to Adult, Compas and German Credit, the three most commonly used datasets in fairness research (Fabris et al., 2022).

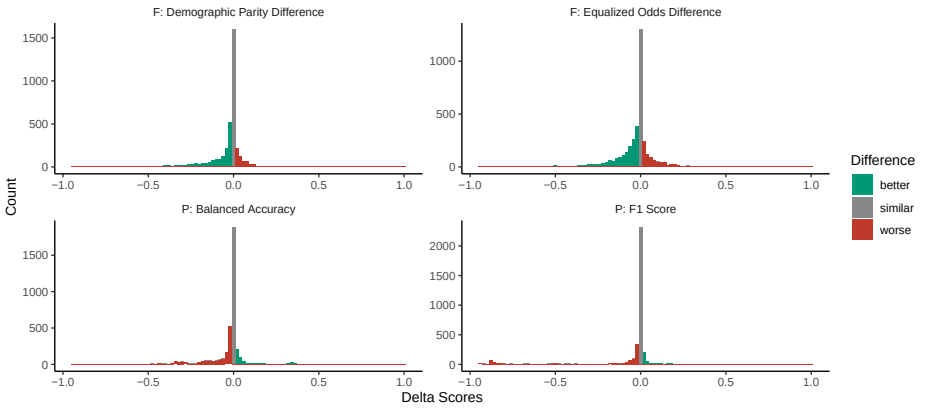


Figure B1: Delta scores across all four metrics are highly variable. Distribution of delta values for metrics of performance and fairness across different processing algorithms. Color-coding indicates whether the change is sizable (above an absolute threshold of 0.01) and corresponds to better (green) or worse (red) scores. For algorithmic fairness metrics lower scores are more desirable, whereas for metrics of performance higher scores are more desirable.

Table A5: Scenarios in the *De-Correlated Datasets* collection. Column C denotes collection membership: k corresponds to the small collection with a cutoff value of $k = 5$; τ corresponds to the bigger collection with a cutoff value of $\tau = 0$. The larger collection encompasses the smaller one. Scenarios are listed based on insertion order.

	C	Dataset	Sens. Attributes	Domain
1	k	folktables_acspubliccoverage	RAC1P	economics
2	k	heart_disease	sex	cardiology
3	k	hmda	applicant_sex_name; applicant_race_name_1	finance
4	k	stop_question_and_frisk_data	SUSPECT_SEX; SUSPECT_RACE_DESCRIPTION; SUSPECT_REPORTED_AGE	law
5	k	folktables_acsemployment_small	RAC1P	economics
6	τ	folktables_acstraveltime	RAC1P	economics
7	τ	compas	sex; age	law
8	τ	folktables_acsincome_small	RAC1P	economics
9	τ	compas_2_years	age	law
10	τ	communities_unnormalized	pct12-21	law
11	τ	arrhythmia	sex	cardiology
12	τ	folktables_acspubliccoverage_small	RAC1P	economics
13	τ	compas_2_years_violent	age	law
14	τ	south_german_credit	age; foreign_worker	finance
15	τ	dutch	age	demography
16	τ	folktables_acsmobility_small	RAC1P	economics
17	τ	law_school_tensorflow	gender	education
18	τ	german_credit_onehot	<= 25 years	finance
19	τ	communities	racePctAsian	law
20	τ	nursery	finance	education
21	τ	german_credit_numeric	age	finance
22	τ	chicago_strategic_subject_list	RACE CODE CD	law

Table A6: Scenarios in the *Permissively Licensed Datasets* collection. Column C denotes collection membership: k corresponds to the small collection with a cutoff value of $k = 5$; τ corresponds to the bigger collection with a cutoff value of $\tau = 0$; an empty value corresponds to the full collection. The larger collections encompass the smaller ones. Scenarios are ordered based on when they were added to the collection.

	C	Dataset	Sens. Attributes	license
1	k	folktables_acspubliccoverage	RAC1P	CC 0
2	k	heart_disease	sex	CC BY 4.0
3	k	communities_unnormalized	pct12-21	CC BY 4.0
4	k	lipton_synthetic_hiring_dataset	sex	CC 0
5	k	bank	age; marital	CC BY 4.0
6	τ	german_credit_onehot	$>$ 25 years	Apache License
7	τ	folktables_acsincome	RAC1P	CC 0
8	τ	south_german_credit	age	CC BY 4.0
9	τ	folktables_acsemployment_small	RAC1P	CC 0
10	τ	german_credit_numeric	age	CC BY 4.0
11	τ	student	sex; age	CC BY 4.0
12	τ	folktables_acstraveltime_small	RAC1P	CC 0
13	τ	folktables_acspubliccoverage_small	RAC1P	CC 0
14	τ	communities	agePct16t24	CC BY 4.0
15	τ	folktables_acsmobility	RAC1P	CC 0
16	τ	law_school_tensorflow	gender	CC BY-SA 4.0
17		arrhythmia	sex	CC BY 4.0
18		adult	race	CC BY 4.0
19		nursery	finance; parents	CC BY 4.0
20		folktables_acsincome_small	RAC1P	CC 0
21		creditcard	SEX	CC BY 4.0
22		folktables_acsmobility_small	RAC1P	CC 0
23		student_language	age	CC BY 4.0
24		drug	ethnicity	CC BY 4.0
25		law_school_lequy	racetxt; male	CC BY-SA 4.0
26		folktables_acstraveltime	RAC1P	CC 0
27		bank_additional_full	age; marital	CC BY 4.0
28		german_credit	foreign_worker	CC BY 4.0
29		generate_synthetic_data	s1	GPL-3.0
30		bank_additional	age	CC BY 4.0
31		folktables_acsemployment	RAC1P	CC 0
32		bank_full	age	CC BY 4.0

Table A7: Scenarios in the *Geographically Diverse Datasets* collection. Column C denotes collection membership: k corresponds to the small collection with a cutoff value of $k = 5$; τ corresponds to the bigger collection with a cutoff value of $\tau = 0$; an empty value corresponds to the full collection. The larger collections encompass the smaller ones. Scenarios are ordered based on when they were added to the collection.

	C	Dataset	Sens. Attributes	country
1	k	folktables_acspubliccoverage	RAC1P	USA
2	k	heart_disease	sex	HUN;CHE;USA
3	k	dutch	age; citizenship	NLD
4	k	creditcard	SEX	TWN
5	k	german_credit_onehot	$>$ 25 years	DEU
6	τ	student	sex	PRT
7		arrhythmia	sex	TUR
8		nursery	finance; parents	SVN
9		synth	sensible_feature	NA
10		drug	ethnicity	WORLD

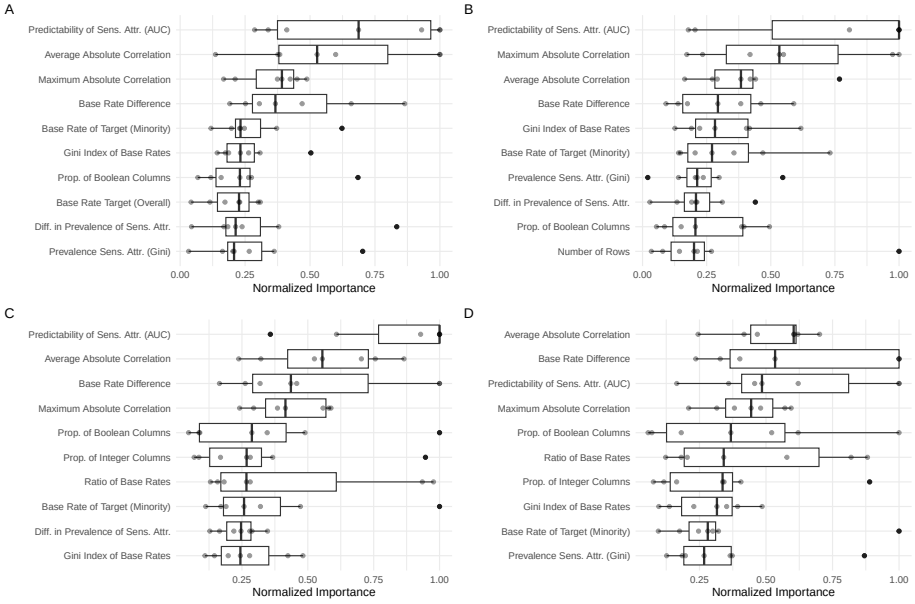


Figure B2: Normalized feature importance of the 10 most important computed metadata features to predict the difference in *Balanced Accuracy* (A), *F1 Score* (B), *Equalized Odds Difference* (C) and *Demographic Parity Difference* (D) across different processing methods.

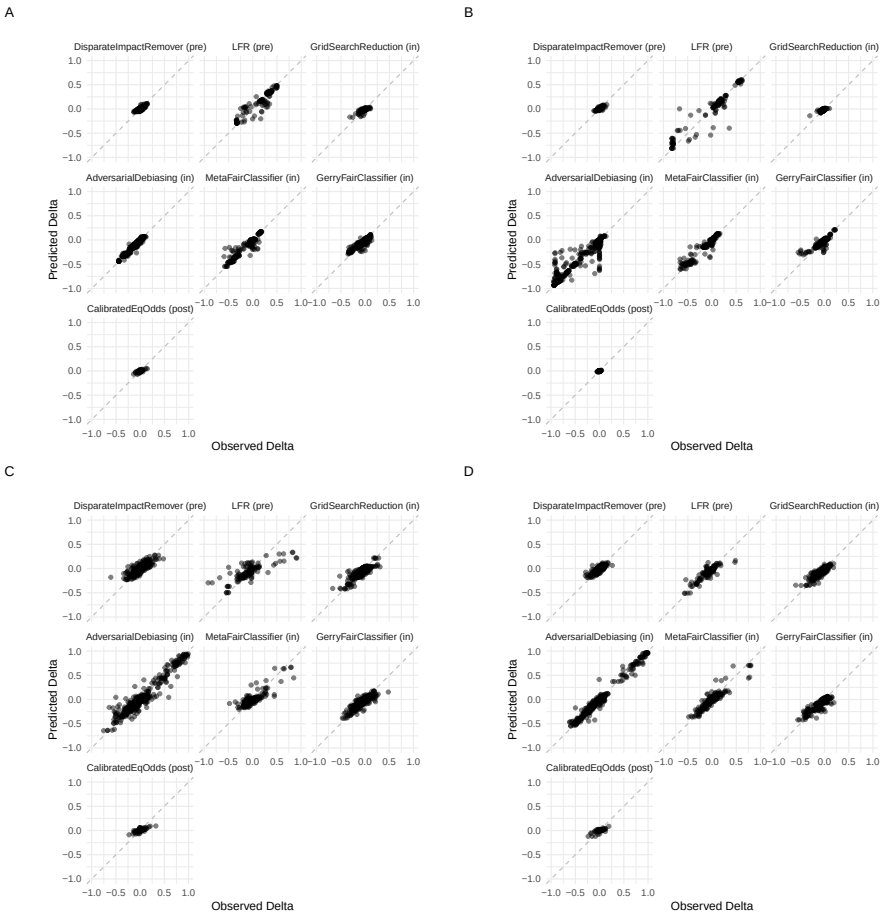


Figure B3: Comparison between observed and model-predicted values for *Balanced Accuracy* (A), *F1 Score* (B), *Equalized Odds Difference* (C) and *Demographic Parity Difference* (D) across different processing methods.

Appendix C. Technical Appendix

C.1 Metrics

$$\begin{aligned}\text{Precision} &= \Pr(y = 1|\hat{y} = 1)\\ \text{Recall} &= \Pr(\hat{y} = 1|y = 1)\\ \text{Specificity} &= \Pr(\hat{y} = 0|y = 0)\end{aligned}$$

We use Balanced Accuracy (bAcc; Eq. 1) and F1 Score (Eq. 2) as measures of performance. The two performance metrics are defined as follows:

$$\text{bACC} = \frac{\text{Specificity} + \text{Recall}}{2} \tag{1}$$

$$\text{F1 Score} = \frac{2}{\text{Precision}^{-1} + \text{Recall}^{-1}} \tag{2}$$

We use Equalized Odds Difference (EOD; Eq. 3) and Demographic Parity Difference (DPD; Eq. 4) as measures of algorithmic fairness. The two fairness metrics are defined as follows:

$$\text{EOD} = \max_g \Pr(\hat{y} = 1|y = 1, S = g) - \min_g \Pr(\hat{y} = 1|y = 1, S = g) \tag{3}$$

$$\text{DPD} = \max_g \Pr(\hat{y} = 1|S = g) - \min_g \Pr(\hat{y} = 1|S = g) \tag{4}$$

When comparing different fairness aware methods, we use delta scores ($\Delta_{a,b}$) for their comparison. These scores are computed for each performance and fairness metric and are defined as follows:

$$\Delta_{a,b} = score_{a,b} - score_{a,\text{baseline}} \tag{5}$$

C.2 Selection Algorithm

Given the corpus of datasets and their associated scenarios $\mathcal{D} = \{D_1, D_2, \ldots, D_N\}$, where each dataset D_i consists of a set of scenarios $D_i = \{s_{i1}, s_{i2}, \ldots\}$, the goal is to construct a collection of scenarios $\mathcal{C}$ such that the pairwise spearman correlations of delta scores (Eq. 5), across *Balanced Accuracy* (Eq. 1), *F1 Score* (Eq. 2), *Equalized Odds Difference* (Eq. 3) and *Demographic Parity Difference* (Eq. 4) between members of $\mathcal{C}$ are as low as possible across different families of fair ML algorithms (*Learning Fair Representations* (Zemel et al., 2013), *Disparate Impact Remover* (Feldman et al., 2015), *Adversarial Debiasing* (Zhang et al., 2018), *Meta-Algorithm* (Celis et al., 2019), *Rich Subgroup Fairness / GerryFair* (Kearns et al., 2018), *Grid Search Reduction* (Agarwal et al., 2018), *Group-Specific Thresholds* (Hardt et al., 2016)). To control the number of scenarios in $\mathcal{C}$, we use either a fixed number k or a correlation threshold τ. While this work uses only one of these constraints at a time, they can be combined if desired. The algorithm proceeds as follows:

1. Let r_{ab} denote the *Spearman rank correlation* between scenarios s_a and s_b, where $s_a, s_b \in \bigcup_{i=1}^{N} D_i$.

2. For each scenario s_a, compute the average Spearman correlation to all other scenarios:

$$\bar{r}_a = \frac{1}{M-1} \sum_{b \neq a} r_{ab}$$

   where M is the total number of scenarios in the corpus. Select the scenario s_m with the lowest average correlation:

$$m = \arg\min_a \bar{r}_a$$

   Initialize the selected set $\mathcal{C} = \{s_m\}$, and the remaining pool $\mathcal{R} = \left(\bigcup_{i=1}^{N} D_i\right) \setminus D_{i(m)}$, where $D_{i(m)}$ is the dataset containing scenario s_m.

3. Repeat the following until $|\mathcal{C}| = k$ or no candidate in $\mathcal{R}$ has an average Spearman correlation strictly less than τ with all members of $\mathcal{C}$:

   (a) For each scenario $s_j \in \mathcal{R}$, compute the average correlation with the current set $\mathcal{C}$:

$$\bar{r}_{j\mathcal{C}} = \frac{1}{|\mathcal{C}|} \sum_{s_i \in \mathcal{C}} r_{ij}$$

   (b) Identify the scenario s_{j^*} with the lowest such average:

$$j^* = \arg\min_{j \in \mathcal{R}} \bar{r}_{j\mathcal{C}}$$

   (c) If $\bar{r}_{j^*\mathcal{C}} < \tau$, add s_{j^*} to $\mathcal{C}$, and remove all scenarios in the same dataset $D_{i(j^*)}$ from $\mathcal{R}$.

4. The algorithm terminates when $|\mathcal{C}| = k$ or no remaining scenario has an average Spearman correlation below τ with the current set $\mathcal{C}$. The resulting subset $\mathcal{C}$ is returned as the final collection of minimally correlated scenarios.

C.3 Example Code using the Package

The following subsection contains exemplary code illustrating the usage of the Python package. We recommend readers to review the online package documentation at `https://brave-ocean-078c2100f.6.azurestaticapps.net/` for a more in-depth description of the package's functions.

C.3.1 Using a Dataset

```
from fairml_datasets import Dataset

# Get the dataset
dataset = Dataset.from_id("folktables_acsemployment")
```

```
# Load as pandas DataFrame
df = dataset.load()  # or df = dataset.to_pandas()
print(f"Dataset shape: {df.shape}")

# Get the target column
target_column = dataset.get_target_column()
print(f"Target column: {target_column}")

# Get sensitive attributes (before transformation)
sensitive_columns_org = dataset.sensitive_columns

# Transform to e.g. impute missing data
df_transformed, transformation_info = dataset.transform(df)
# Sensitive columns may change due to transformation
sensitive_columns = transformation_info.sensitive_columns

# Split into train and test sets
train_df, test_df = dataset.train_test_split(df, test_size=0.3)

# Run analyses on the data
```

C.3.2 Using a Collection of Datasets / Scenarios

```
from fairml_datasets.collections import DeCorrelatedSmall

collection = DeCorrelatedSmall()

# The collection consists of scenarios
for scenario in collection:
    # Each scenario behaves just like a dataset

    # Load as pandas DataFrame
    df = scenario.load()  # or df = scenario.to_pandas()
    print(f"Dataset shape: {df.shape}")

    # Get the target column
    target_column = scenario.get_target_column()
    print(f"Target column: {target_column}")

    # Get sensitive attributes (before transformation)
    sensitive_columns_org = scenario.sensitive_columns

    # Transform to  e.g. impute missing data
```

```
df_transformed, transformation_info = scenario.transform(df)
# Sensitive columns may change due to transformation
sensitive_columns = transformation_info['sensitive_columns']

# Split into train and test sets
train_df, test_df = scenario.train_test_split(df, test_size=0.3)

# Run analyses on the data
```

C.4 Annotation Procedure

We started the annotation process by collecting all tabular datasets used for fair classification tasks in a large survey of fair ML datasets (Fabris et al., 2022). This provided a list of $n = 37$ unique datasets. Additionally, we added the folktables (Ding et al., 2021) collection of datasets, due to its recent popularity and as the datasets specifically try to address issues in the most popular dataset in the survey: *Adult* (Kohavi, 1996).

For each dataset, we annotated the information required to practically use the dataset in a fair classification task, as well as key qualitative and quantitative data regarding the information represented in each dataset. During this process, a critical issue quickly became apparent: While datasets are commonly referenced by name as if they were uniquely identified, this is often not the case in practice. A striking example is the widely used Bank dataset, one of the most frequently cited datasets in Fair ML (Fabris et al., 2022). Although typically referred to as Bank or Bank Marketing, the primary source[1] actually comprises four distinct datasets, each differing in their respective number of instances and attributes. Recognizing this ambiguity, we adapted our annotation methodology to explicitly capture dataset variants, significantly increasing the number of distinct datasets in the corpus. In our framework, we treat these variants as separate datasets while preserving their connection to maintain clarity and traceability.

When collecting the information required to download and load datasets, we were forced to exclude $n = 11$ datasets due to data not being publicly available or with restricted access. We excluded a further $n = 18$ datasets, if there were issues with recreating how a dataset was generated or the dataset's usage did not fit into schema of a "classic" fairML classification task including features, a target column and sensitive attribute(s). A detailed breakdown of excluded datasets and the reasons for their exclusion is available in Section C.7.

After exclusion of non-eligible datasets and inclusion of different variants, we arrive at a list of $N = 44$ datasets.

Datasets were annotated by two of the authors with help from research assistants. A random subset of annotations was reviewed by a third author.

C.5 Annotated Columns

The following section provides descriptions of columns which were manually annotated for each dataset in the corpus.

1. https://archive.ics.uci.edu/dataset/222/bank+marketing

new_dataset_id A unique identifier for each dataset. Usually derived from the dataset name.

dataset_name An official, common, or known name of the dataset that is unique across datasets.

base_dataset_name In case there are different variants of the same dataset, this field holds a common name to group all these variants together.

description_public This is a free-text field reporting (1) the aim/purpose of a data artifact (i.e., why it was developed/collected), as stated by curators or inferred from context; (2) a high-level description of the available features; (3) the labeling procedure for annotated attributes, with special attention to sensitive ones, if any; (4) the envisioned ML task, if any.

notes_public Any notes or comments regarding this dataset / task combination.

dataset_aliases Any names that this dataset is called by. While 'dataset_name' only contains the single most common name, this field holds possible aliases used to reference this dataset.

affiliation Affiliation of the creators of the dataset. Based on reports, articles, or official web pages presenting the dataset.

domain_class_main The main field where the data is used (e.g., computer vision for ImageNet) or the field studying the processes and phenomena that produced the dataset (e.g., radiology for CheXpert).

domain_class_multi The primary fields where the data is used (e.g., computer vision for ImageNet) or the fields studying the processes and phenomena that produced the dataset (e.g., radiology for CheXpert). Multiple domains are possible in this feature.

domain_freetext Fine-grained domain of the prediction task. Summarized with 1 - 2 words.

sample_size Dataset cardinality. Rough estimate of the size of the dataset.

year_last_updated The last known update to the dataset. For resources whose collection and curation are ongoing (e.g., HMDA), we write “present”.

years_data The timespan covered in the data. This refers to the "social realities" captured in the data i.e., data from which year(s) is present in the data.

citation The main / official source to cite this dataset in BibTeX format. For synthetic datasets, this refers to the original paper where the dataset was first introduced.

main_url The main landing page or website related to the dataset. This is a website with information on the dataset and not the dataset itself, which is referenced via 'download_url'.

related_urls List of related links and resources to the dataset.

license Under which license is the dataset made available? A "?" indicates that no license was found.

continent Continent(s) where the dataset is sourced. In two-letter format. If "n/a", this concept is not applicable for a dataset (e.g., a synthetic one).

country Countrie(s) where the dataset is sourced. In ISO3 format. If "n/a", this concept is not applicable for a dataset (e.g., a synthetic one).

dataset_variant_id This ID is used to identify different datasets belonging to the same original dataset e.g., COMPAS has 3 unique smaller datasets belonging to this one bigger one. In cases like this, each smaller dataset gets its own dataset_variant_id.

dataset_variant_description Description outlining how this "sub-dataset" is different from the others. Only filled out if there are multiple "dataset_variant_ids".

is_accessible Is the dataset publicly accessible? "Manual download" indicates that an automated download is not possible.

download_url URL to the dataset file itself, if it is publicly accessible.

custom_download Are there some extra steps needed to download the dataset itself, e.g., unpacking a ZIP archive?

filename_raw Filename of the dataset for downloading it or finding it in a ZIP archive.

format Format of the dataset. Corresponds to the format the data is in, not the extension of the dataset e.g., CSV for comma-separated-values, TSV for tab-separated-values, FIXED-WIDTH for fixed-width formats etc.

colnames Column names to use if the dataset file does not include them.

processing Does the dataset need some special pre-processing to be in the correct format?

sensitive_attributes Sensitive attributes that are *available* in the dataset. Supports multiple entries, separated with a semicolon and a space: '; '.

typical_col_sensitive All columns containing available sensitive attributes and the information they contain in a categorical fashion. Covering the attributes listed in 'sensitive_attributes'. Formatted as a JSON dictionary.

typical_col_features All columns typically used as features / predictors. Either a list of column names indicating a positive selection or a list of column names prefixed with a - indicating a negative selection i.e. all columns except the listed ones. A - indicates using all available columns (except the target).

typical_col_target Column(s) which are being predicted. If more than one, separated by semicolons.

target_lvl_good Which value of the target variable is considered desirable? Desirable here means good for any person impacted by a system built using this data.

target_lvl_bad Which value of the target variable is considered undesirable? Undesirable here means bad for any person impacted by a system built using this data.

dataset_size Whether a dataset is exceptionally large.

C.6 Computed Metadata

The following section provides descriptions of the computed metadata features which are implemented in the Python package and computed for each of the datasets in the corpus. The technical implementation can be reviewed in the publicly available source code of the package.

Size As Ding et al. (2021) note, increasing dataset size does not necessarily reduce observed disparities due to persistent structural inequalities. We try to cover a broad range of dataset sizes in our corpus and compute dataset sizes by rows (samples) and columns (attributes) of both prepared (`meta_pretrans_n_rows, meta_pretrans_n_cols`) and transformed datasets (`meta_n_rows, meta_n_cols`).

Missing values To address potential bias from missing data (e.g. see Pessach and Shmueli, 2022; Wang and Singh, 2021; Martínez-Plumed et al., 2019), we

calculate the fraction of missing data per dataset. Metadata was computed prior to processing to assess the proportion of rows (`meta_pretrans_prop_NA_rows`), columns (`meta_pretrans_prop_NA_cols`) and cells (`meta_pretrans_prop_NA_cells`) that contain missing values. We further calculate missingness within each group of the protected attribute (only when binarizing; `meta_prop_NA_sens_minority, meta_prop_NA_sens_majority`).

Attribute types We calculate the proportions of different numeric (`meta_prop_cols_float, meta_prop_cols_int`) and logical (`meta_prop_cols_bool`) data types in the data to assess their potential influence.

Sensitive AUC Non-sensitive attributes can act as proxies for sensitive ones (e.g. see Pessach and Shmueli, 2022; Mehrabi et al., 2021; Fawkes et al., 2024). Identifying and addressing such proxies can help mitigate unfairness (Pessach and Shmueli, 2022; Matloff and Zhang, 2022). To assess this, we define *Sensitive AUC* as the ROC-AUC of a random forest model (Ho, 1995) trained to predict the sensitive attribute using only non-sensitive features (`meta_sens_predictability_roc_auc`). A higher Sensitive AUC suggests that non-sensitive attributes may encode sensitive information.

Bivariate correlations Serving as an additional indicator of potential proxy variables, we computed the correlation between each non-sensitive feature and the sensitive attribute, using the average and maximum correlation values (`meta_average_absolute_correlation, meta_maximum_absolute_correlation`).

Number of protected groups Some fairness methods require binary representations of protected attributes, leading to the binarization of categorical or numerical sensitive attributes during preprocessing. Documenting the original number of protected groups before processing (`meta_pretrans_unique_group_counts_pre_agg`) helps track this process and may provide insight into how such simplifications affect the performance and suitability of fairness methods.

Prevalence We computed the proportions of minority and majority groups within the dataset (only when binarizing; `meta_prev_sens_minority, meta_prev_sens_majority`), along with the absolute difference between them (`meta_prev_sens_difference`) and the imbalance ratio (`meta_prev_sens_ratio`). A smaller absolute difference and an imbalance ratio closer to 1 indicate a more balanced distribution of the sensitive attribute.

Base Rate Similar to prevalence, we computed the probability of the favorable outcome overall (`meta_base_rate_target`) and for each group (only when binarizing; `meta_base_rate_target_sens_minority`, `meta_base_rate_target_sens_majority`) along with the absolute difference (`meta_base_rate_difference`) and ratio (`meta_base_rate_ratio`) between them.

Gini-Simpson Index The Gini-Simpson Index measures the probability that two randomly selected individuals belong to different groups. Similar indices have been previously used by Mecati et al. (2023) and Vetrò et al. (2021) to assess balance and detect potential unfairness in datasets. We compute the Gini-Simpson Index for both group prevalence and base rates

$$GS = 1 - \sum_i p_i^2,$$

where p_i is the proportion of instances in group $i \in \{1, 2\}$ (protected or non-protected). For prevalence, this is is the proportion of individuals per group relative to the entire dataset (`meta_prev_sens_gini`). For base rates, p_i denotes the proportion of favorable outcomes within each group (`meta_base_rate_sens_gini`).

C.7 Excluded Datasets

This subsection contains explanations for additional datasets that were excluded from the corpus. The annotation procedure is described in detail in Section C.4.

2016 Presidential Elections (2 datasets) This dataset from the FiveThirtyEight 2016 Election Forecast was developed with the goal of providing an aggregated estimate of the probability that Trump/Clinton wins the 2016 election based on multiple polls, weighting each input according to sample size, recency, and historical accuracy of the polling organization. For each poll, the dataset provides the period of data collection, its sample size, the pollster conducting it, their rating, and a url linking to the source data. The dataset does not contain any sensitive attributes and was therefore excluded. One annotated but excluded dataset came from ABC News, and another, potentially deviating, from (Sabato and Yom-Tov, 2020).

Cancer Cases and Deaths (3 datasets) The main dataset reports state-level cancer prevalence for 18 cancer types, based on data from the CDC's NPCR and the NCI's SEER program. Mortality data come from the CDC's National Vital Statistics System. As it contains only aggregated data on state-level, it was excluded from our analysis. Two additional datasets provided the source data on new cases and deaths. As neither was used in isolation in our annotations, both were excluded with the main dataset.

Clinical Annotations / Warfarin Dosage / PharmaGKB (4 datasets) The data, collected by the International Warfarin Pharmacogenetics Consortium and co-curated by PharmGKB, was used to study algorithmic estimation of optimal warfarin dosage. The original data includes thousands of patient demographics, comorbidities, medications, genetics, and effective warfarin doses. However, the available datasets do not contain demographic details and only a specialty group column indicates few pediatric cases. Due to the absence of sensitive attributes, these datasets were excluded. The excluded datasets comprised: 1) meta-data for each clinical annotation; 2) genotype/allele-based annotation text with CPIC-assigned function, if available; 3) supporting annotation details (variant, guideline, label); and 4) clinical annotation history with creation and update dates.

COMPAS (4 datasets) We retain the original COMPAS data published by ProPublica Angwin et al. (2016). Specific versions of the COMPAS dataset were excluded, including an unofficial version published on Kaggle, used in one reviewed study (Jabbari et al., 2020), and two others, each appearing in a single paper (Wang et al., 2019; Mandal et al., 2020), due to a lack of clarity in the differences and processing from the original ProPublica release. The COMPAS repository[2] also includes a file with "raw" scores, named `compas-scores-raw.csv`, which we decided not to include, as it is not further utilized in the analysis.

FICO Credit Score, Credit Score Performance (2 datasets) The dataset originates from a 2007 Federal Reserve report to the US Congress on credit scoring and its

2. `https://github.com/propublica/compas-analysis`

effects on the availability and affordability of credit. The collection, creation, processing, and aggregation was carried out by the working group and is based on a sample of 301,536 TransUnion TransRisk scores from 2003. The dataset contains only aggregated statistics per FICO score and race/ethnicity group and was therefore excluded. A second version with unclear differences was also excluded.

Fifa 20 Complete Player This dataset was scraped by Stefano Leone and shared on Kaggle. It contains player data from FIFA Career Mode (FIFA 15-20). We excluded this dataset, because relevant sensitive attributes and target variables were unclear. A paper by Awasthi et al. (2021) created a sensitive attribute by predicting nationality from player names using LSTM, an approach that could introduce unnecessary uncertainty and therefore may have reduced comparability.

Pima Diabetes This dataset was derived from a medical study of Native Americans from the Gila River Community, often called Pima. Conducted by the National Institute of Diabetes and Digestive and Kidney Diseases since the 1960s, the study found a large prevalence of *diabetes mellitus* in this population. The dataset includes a subset of the original study, focusing on women of age 21 or older. It reports diabetes test results and eight key risk factors, such as number of pregnancies, skin thickness, and BMI. Relevant sensitive attributes were not clear based on the papers we reviewed, so we decided to exclude the dataset.

US Census (1990) This dataset is derived from the 1990 US Census. In the reviewed literature, the classification task was often unclear or unsuitable for our analysis goals (e.g., Sabato and Yom-Tov, 2020). Another meta-analysis referenced 25 selected numeric attributes without specifying them.

C.8 Computational Infrastructure

Experiments were run on a shared Linux compute cluster with partitions and compute infrastructure chosen based on availability of resources. Experiments were run as four consecutive jobs, the first running experiments at high concurrency and the later re-running errored out experiments at lower concurrency.

The first job was run on a node with access to 76 CPU cores and 512 GB of memory over a duration of 11 hours. Later jobs were run on a node with 96 CPUs and 1 TB of memory, using 5-fold parallelism and a maximum execution time of 2.5 hours for the second and third run and 5 hours for the last run. Experiments were conducted using only CPU compute.

C.9 Software

Simulation experiments were conducted using Python (Python Core Team, 2019) version 3.10 and the Python package `multiversum` (Simson, 2024) version 0.7.0. We used the implementations of fairness-aware processing methods from the package `AIF360` (Bellamy et al., 2018) and used `scikit-learn` (Pedregosa et al., 2011) to fit logistic regressions. Data were processed using the newly developed `fairml_datasets` package, utilizing `pandas` (McKinney, 2010), `fastparquet` (Durant) and `scikit-learn`. Multiple other packages were utilized as (peer) dependencies of the named packages. We use `uv` (Marsh, 2024) for virtual environment management.

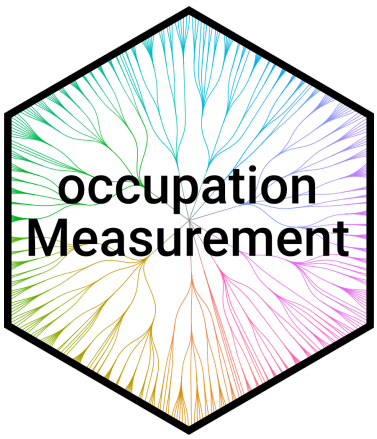
Results from the experiments were analysed using R version 4.4.1 (R Core Team, 2024) with packages from the `tidyverse` (Wickham et al., 2019), `patchwork` (Pedersen, 2024) and `tidymodels` (Kuhn and Wickham, 2020). Color schemes are used from the R packages `awtools` (Wehrwein, 2025) and `wesanderson` (Ram and Wickham, 2023). We use `renv` for virtual environment management.

Lockfiles for both Python and R packages are provided with the codebase.

Experiments were executed using a docker container converted to the `enroot` format[3].

3. https://github.com/NVIDIA/enroot

8. occupationMeasurement: A Comprehensive Toolbox for Interactive Occupation Coding in Surveys



Contributing article

Simson, J., Kononykhina, O., & Schierholz, M. (2023). occupationMeasurement: A Comprehensive Toolbox for Interactive Occupation Coding in Surveys. In *Journal of Open Source Software*, 8(88), 5505. doi: 10.21105/joss.05505 URL https://doi.org/10.21105/joss.05505

Code repository

https://github.com/occupationMeasurement/occupationMeasurement

Package on CRAN

https://cran.r-project.org/web/packages/occupationMeasurement/index.html

Package Documentation

https://occupationMeasurement.github.io/occupationMeasurement/

Author contributions

M. Schierholz contributed the idea and funding for the project through a DFG work package. M. Schierholz developed a first prototype of the software in prior work. J. Simson developed the final package based on the initial prototype, re-organizing and re-writing most of the codebase. J. Simson lead the submission of the package to CRAN, the writing of the paper and the revision process. All authors contributed through fruitful comments, proofreading and revisions of the manuscript.

The Journal of Open Source Software

occupationMeasurement: A Comprehensive Toolbox for Interactive Occupation Coding in Surveys

Jan Simson [1¶], **Olga Kononykhina**[1], **and Malte Schierholz** [1]

1 Department of Statistics, Ludwig-Maximilians-Universität München, Germany ¶ Corresponding author

DOI: 10.21105/joss.05505

Software

- Review
- Repository
- Archive

Editor: Chris Vernon

Reviewers:

- @welch16
- @danielruss

Submitted: 30 March 2023
Published: 24 August 2023

Summary

People earn a living a multitude of ways which is why the occupations they pursue are almost as diverse as people themselves. This makes quantitative analyses of free-text occupational responses from surveys hard to impossible, especially since people may refer to the same occupations with different terms. To address this problem, a variety of different classifications have been developed, such as the International Standard Classification of Occupations 2008 (ISCO) (ILO, 2012) and the German Klassifikation der Berufe 2010 (KldB) (Bundesagentur für Arbeit, 2011), narrowing down the amount of occupation categories into more manageable numbers in the mid hundreds to low thousands and introducing a hierarchical ordering of categories. This leads to a different problem, however: Coding occupations into these standardized categories is usually expensive, time-intensive and plagued by issues of reliability.

Here we present a new instrument that implements a faster, more convenient and interactive occupation coding workflow where respondents are included in the coding process. Based on the respondent's answer, a novel machine learning algorithm generates a list of suggested occupational categories from the Auxiliary Classification of Occupations (Schierholz, 2018), from which one is chosen by the respondent (see Figure 1). Issues of ambiguity within occupational categories are addressed through clarifying follow-up questions. We provide a comprehensive toolbox including anonymized German training data and pre-trained models without raising privacy issues, something not possible yet with other algorithms due to the difficulties of anonymizing free-text data.

Statement of Need

Assigning occupations to standardized codes is a critical task frequently encountered in research, public administration and beyond: They are used in government censuses (e.g. USA, UK, Germany) and administrative data to better understand economic activity, in epidemiology to estimate exposure to health hazards, and in sociology to obtain a person's socio-economic status, to name a few examples.

To date, the standard approach to coding occupations is to collect one or two free-text responses and later hand-coding these descriptions by trained personnel with a classification manual, possibly assisted by computer software. Since coding typically occurs after data collection, based on the responses only and without the ability to request clarifying information from the respondent, the assignment of categories is often inaccurate. This approach to occupational coding is costly due to the experts' time needed and often suffers from low inter-coder reliability[1].

[1]An international review found rates of agreement between 44% and 89% when different coders code the same answers across different studies (Mannetje & Kromhout, 2003). For a more in-depth discussion of the factors affecting the reliability of occupation coding, see Conrad et al. (2016) and Massing et al. (2019).

The Journal of Open Source Software

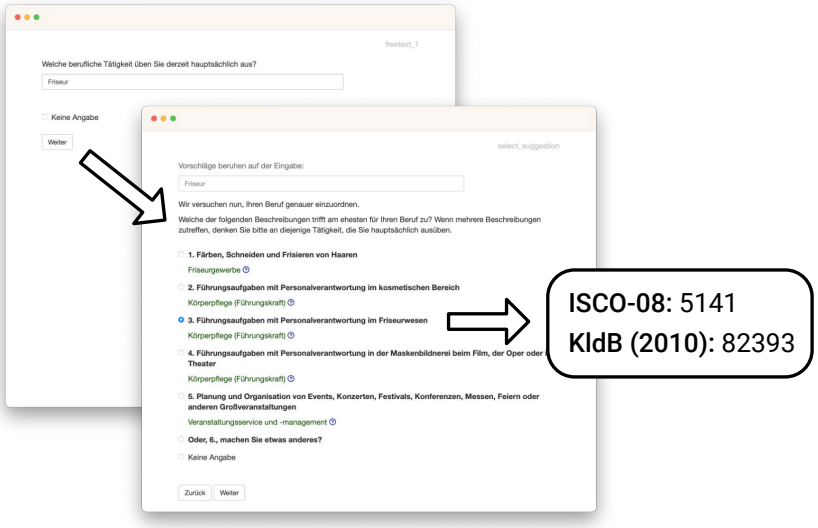


Figure 1: Typical flow of the interactive application. 1. A respondent provides a free text response describing her occupation. 2. The machine learning model then generates a list of suggested categories, from which the respondent will select one. 3. As a result, the associated occupational category codes from both the German KldB-2010 and the international ISCO-08 are returned.

Given the limitations of manual coding, a technical solution that can generate a suggested code fast enough to elicit immediate feedback from respondents would be a boon. However, implementations of this idea are few, and none are openly available. Technological solutions using machine learning have been proposed (Creecy et al., 1992; Gweon et al., 2017; Russ et al., 2014, 2016; Schierholz & Schonlau, 2021) but face problems obtaining training data and sharing trained models due to privacy issues, as training data often contains sensitive free-text responses that may personally identify respondents.

Our toolbox addresses these issues by implementing an occupation coding workflow during the interview as discussed in Peycheva et al. (2021) and Schierholz et al. (2018). Difficulties of data and model sharing are resolved by using a novel machine learning algorithm specifically crafted to work with anonymized occupational data.

Functionalities

We provide an open-source implementation of a machine learning algorithm for occupation coding with immediate feedback and verification, available as an R (R Core Team, 2019) package on CRAN[2]. An introductory "Getting Started" guide is available for anyone looking to use the package. To make it widely useful, our toolkit can be readily integrated in both self-administered web surveys as well as interviewer-administered (telephone) surveys using the included questionnaires. Programmers can adapt these questionnaires to fit a wide array of requirements. The toolbox includes custom survey software built on top of the shiny (Chang et al., 2023) framework to integrate machine learning predictions into surveys. On the off-chance that further flexibility is needed, we offer direct API access for completely custom

[2]Package "occupationMeasurement" on CRAN: https://cran.r-project.org/web/packages/occupationMeasurement/index.html.

data collection and integration into existing survey software. As we built the toolbox on top of the shiny (Chang et al., 2023) and plumber (Schloerke et al., 2022) frameworks, deployments on the Web are easy. We further provide pre-built container images for even easier deployment in production environments.

The toolbox uses a specifically developed list of occupational task descriptions, the Auxiliary Classification of Occupations, designed to be easier to understand and less ambiguous than existing lists of job titles (Schierholz, 2018). Alongside this list, it provides matching follow-up questions to enable a fine-grained assignment into existing classification systems.

The machine-learning algorithm used in our instrument is able to work with anonymized training data while retaining predictive performance on-par with other machine learning and non-machine-learning algorithms (Schierholz, 2019; Schierholz & Schonlau, 2021). This enables sharing training data as well as trained models, allowing on the one hand out-of-the-box usage of our instrument without the need for labeled data or pre-training, but also the further sharing of newly trained models by users of the toolbox. Anonymized training data and pre-trained models in German are included with the package.

The toolbox is released under the MIT license alongside extensive online documentation. Quality of the software is ensured using automated testing via continuous integration. The toolbox has successfully been piloted with various modes of data collection in collaboration with different German survey institutes.

Related Work

This project is not the first to apply technology to assist in the coding of occupations, although it is the first tool to be released as open-source software and to offer this level of convenience and flexibility. Notable examples of prior work include the WISCO[3] database (Tijdens, 2010), providing a search tree with levels of nested categories of occupations for use in Web surveys. Another prominent tool is CASCOT (Elias et al., 2014), short for Computer Assisted Structured Coding Tool. CASCOT uses a mixture of a coding index, which requires manual editing, text distances and manually specified rules to code responses into occupational categories. A promising tool has also been developed for the US Standard Occupational Classification System (SOC) (U. S. Bureau of Labor Statistics, n.d.), called SOCcer (Russ et al., 2014, 2016). Similar to the software presented here, SOCcer relies on using a machine learning algorithm. Unfortunately, neither SOCcer nor CASCOT are open-sourced, with the former offering coding via a free online version[4] and the latter requiring payment for use.

Acknowledgements

This project is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project numbers 290773872 and 460037581.

References

Bundesagentur für Arbeit. (2011). *Klassifikation der Berufe 2010: Vols. 1 & 2*. Bundesagentur für Arbeit.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2023). *Shiny: Web application framework for r*. https://shiny.rstudio.com/

[3] The WISCO database used different names over time, but kept the same acronym. The latest description is: "World database of occupations, coded ISCO". The database is available at https://surveycodings.org.

[4] Online version available at: https://soccer.nci.nih.gov/soccer/.

Conrad, F. G., Couper, M. P., & Sakshaug, J. W. (2016). Classifying Open-Ended Reports: Factors Affecting the Reliability of Occupation Codes. *Journal of Official Statistics*, *32*(1), 75–92. https://doi.org/10.1515/jos-2016-0003

Creecy, R. H., Masand, B. M., Smith, S. J., & Waltz, D. L. (1992). Trading MIPS and memory for knowledge engineering. *Communications of the ACM*, *35*(8), 48–64. https://doi.org/10.1145/135226.135228

Elias, P., Birch, M., & Ellison, R. (2014). CASCOT international version 5 user guide. *Institute for Employment Research, University of Warwick, Coventry*. https://warwick.ac.uk/fac/soc/ier/software/cascot/internat/cascot_international_user_guide.pptx

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., & Steiner, S. (2017). Three Methods for Occupation Coding Based on Statistical Learning. *Journal of Official Statistics*, *33*(1), 101–122. https://doi.org/10.1515/jos-2017-0006

ILO. (2012). *International Standard Classification of Occupations 2008 (ISCO-08): Structure, group definitions and correspondence tables*. http://www.ilo.org/global/publications/ilo-bookstore/order-online/books/WCMS_172572/lang--en/index.htm

Mannetje, A. 't., & Kromhout, H. (2003). The use of occupation and industry classifications in general population studies. *International Journal of Epidemiology*, *32*(3), 419–428. https://doi.org/10.1093/ije/dyg080

Massing, N., Wasmer, M., Wolf, C., & Zuell, C. (2019). How Standardized is Occupational Coding? A Comparison of Results from Different Coding Agencies in Germany. *Journal of Official Statistics*, *35*(1), 167–187. https://doi.org/10.2478/jos-2019-0008

Peycheva, D. N., Sakshaug, J. W., & Calderwood, L. (2021). Occupation Coding During the Interview in a Web-First Sequential Mixed-Mode Survey. *Journal of Official Statistics*, *37*(4), 981–1007. https://doi.org/10.2478/jos-2021-0042

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Russ, D. E., Ho, K.-Y., Colt, J. S., Armenti, K. R., Baris, D., Chow, W.-H., Davis, F., Johnson, A., Purdue, M. P., Karagas, M. R., Schwartz, K., Schwenn, M., Silverman, D. T., Johnson, C. A., & Friesen, M. C. (2016). Computer-based coding of free-text job descriptions to efficiently identify occupations in epidemiological studies. *Occupational and Environmental Medicine*, *73*(6), 417–424. https://doi.org/10.1136/oemed-2015-103152

Russ, D. E., Ho, K.-Y., Johnson, C. A., & Friesen, M. C. (2014). *2014 IEEE 27th International Symposium on Computer-Based Medical Systems (CBMS)*. 347–350. https://doi.org/10.1109/CBMS.2014.79

Schierholz, M. (2018). Eine Hilfsklassifikation mit Tätigkeitsbeschreibungen für Zwecke der Berufskodierung. *AStA Wirtschafts- und Sozialstatistisches Archiv*, *12*(3-4), 285–298. https://doi.org/10.1007/s11943-018-0231-2

Schierholz, M. (2019). *New Methods for Job and Occupation Classification* [PhD thesis].

Schierholz, M., Gensicke, M., Tschersich, N., & Kreuter, F. (2018). Occupation coding during the interview. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*(2), 379–407. https://doi.org/10.1111/rssa.12297

Schierholz, M., & Schonlau, M. (2021). Machine learning for occupation coding—a comparison study. *Journal of Survey Statistics and Methodology*, *9*(5), 1013–1034. https://doi.org/10.1093/jssam/smaa023

Schloerke, B., Allen, J., Tremblay, B., Dunné, F. van, Vandewoude, S., & RStudio. (2022). *Plumber: An API generator for r*. https://CRAN.R-project.org/package=plumber

Tijdens, K. (2010). *Measuring occupations in web-surveys: the WISCO database of occupations*. https://dare.uva.nl/search?identifier=2a72aad5-24dc-4891-9bc4-05deddad520b

U. S. Bureau of Labor Statistics. (n.d.). *Standard Occupational Classification System (SOC) System*. https://www.bls.gov/soc/

Part III.

Software

9. Multiversum

Description

`multiversum` is a package designed to make it easy to conduct multiverse analyses in Python. The package is intended to seamlessly integrate into a normal analysis or ML workflow and can also be added to an existing machine learning or analysis pipeline. It supports parallel execution at scale, the analysis of specific universes, as well as pausing and restarting an analysis.

The package has been used to run experiments in Simson et al. (2025a) (Chapter 6) and Simson et al. (2025b) (Chapter 7). Early variants of the package have been used to run experiments in Simson et al. (2024b) (Chapter 5) and Julia Sophie Broden (2025).

Code repository

https://github.com/jansim/multiversum

Package on PyPI

https://pypi.org/project/multiversum/

Package Documentation

https://jansim.github.io/multiversum/

License

The package is released under the Apache 2.0 License.

10. World-Wide-Lab

Description

`World-Wide-Lab` is a platform to collect online data at scale, with a special focus on running online citizen science experiments. It sits right in between web experiment libraries and data analysis software in the scope of running an online web study.

Data collected within `World-Wide-Lab` has been used in Simson et al. (2025a) (Chapter 6) and James et al. (2025).

Code repository

https://github.com/world-wide-lab/world-wide-lab

Documentation

https://worldwidelab.org

License

The software is released under the MIT License.

Part IV.

Closing

11. Conclusion

In this thesis, I argue that the classic ML pipeline should be critically re-evaluated rather than being taken for granted. Across five publications situated at different stages of the ML pipeline, I demonstrate both issues in current practices as well as potential solutions. In Chapters 4 and 5, respectively, I highlight how "lazy" data practices and arbitrary evaluation strategies can be problematic. In Chapter 6, I demonstrate how participatory input can be used to address "lazy" data practices such as the omission of subpopulations and help navigate the ML multiverse described in Chapter 5. Similarly, Chapter 7 aims to improve practices in the field by providing a readily usable corpus of datasets with sensible defaults for pre-processing and an easy way to use diverse collections of datasets. Last, Chapter 8 puts into question practices around the sourcing of novel data, exploring a more empowering form of self-classification.

While the thesis makes significant contributions to addressing issues, a substantial amount of work remains to be done, and in many ways, we have barely scratched the surface of how the ML pipeline may continue to (and possibly need to) evolve. Looking ahead, there are several promising directions for future research building on the thesis.

One of these is that, similar to the past, future milestones in ML and AI will require new, high-quality datasets. As datasets can encode important information about local distributions, cultural norms, and preferences, we must ensure that we accurately represent different regions of the world in the data we source. Addressing these geographical imbalances and representing populations worldwide more accurately will require sourcing new datasets. However, sourcing new data and especially doing so *ethically*, is difficult and will require systemic changes in the field (Zhao et al., 2024; Andrews et al., 2023).

The *FairGround* framework (Chapter 7) provides a first step in this direction, by making geographical imbalances in dataset representation measurable, highlighting blind spots for future data sourcing efforts. Applying the framework to domains beyond classification in FairML may enable a broader awareness of such issues. Unfortunately, creating and annotating a corpus like this is a complex and labor-intensive task. In this light, large language models (LLMs) offer a promising opportunity to (partially) automate the annotation process through the extraction of information from documents. Allowing annotators to change from an active annotation role to one of reviewing pre-filled annotations could increase the feasibility of creating dataset corpora for different domains. Nevertheless, as with the automation of decision-making in ADM, any such implementation must be undertaken with great care. Issues such as automation bias (Mosier et al., 1998; Beck et al., 2025), i.e., the tendency to overly agree with automated suggestions, will require thoughtful handling.

The thesis highlights the opportunities of multiverse analyses for ML and particularly FairML (Chapter 5). How best to handle the resulting data remains an open question, however (Del Giudice and Gangestad, 2021). Future methodological research focusing on exactly this aspect – novel ways to analyze the results of a multiverse analysis – has the potential to provide a valuable contribution

to scientific methodology, ML, and AI, as well as various other fields of research. In the context of ML and AI, this line of work also presents an interesting opportunity to improve the understanding of model multiplicity.

A second difficulty with multiverse analyses is creating their decision space. Which decisions and options to include in the analysis are heavily dependent on what one defines as "plausible" or "arbitrary" (Steegen et al., 2016; Del Giudice and Gangestad, 2021). A potential solution to this may be sourcing decisions and options from the literature. This is yet another area where sufficiently powerful document extraction systems may prove helpful: extracting and structuring choices and decisions along the ML pipeline and data science life cycle from published literature and code would grant the ability to perform more empirically informed multiverse analyses.

Moreover, adopting learnings from participatory design provides a promising avenue to further empower stakeholders. This includes opportunities for stakeholders not only to be included and represented in the data, but also to shape both the collection processes and the models that will be trained on said data. As more powerful systems are developed and deployed, creating structures for participatory and democratic input around these systems will become crucial to better understand which values and preferences ML and AI systems should align with (Huang et al., 2024; Awad et al., 2018). Building on the success in Chapter 6, citizen science platforms may be a promising avenue for enabling and empowering input on model design, alignment, deployment, and beyond.

Even more so, when adopting the lens of the "ladder of participation" (Arnstein, 1969) to frontier AI and ML models, public research itself is often situated towards the lower end of the ladder, with training and deployment of models shaped by industry. Instead, I argue that we should strive to further climb up this ladder, aiming for a more active role in shaping the role of AI and ML in society going forward.

List of Contributing Publications

Simson, J., Fabris, A., & Kern, C. (2024). Lazy Data Practices Harm Fairness Research. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 642–659. doi: 10.1145/3630106.3658931 URL https://doi.org/10.1145/3630106.3658931

Simson, J., Pfisterer, F., & Kern, C. (2024). One Model Many Scores: Using Multiverse Analysis to Prevent Fairness Hacking and Evaluate the Influence of Model Design Decisions. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1305–1320. Association for Computing Machinery. doi: 10.1145/3630106.3658974 URL https://doi.org/10.1145/3630106.3658974

Simson, J., Draxler, F., Mehr, S., & Kern, C. (2025). Preventing Harmful Data Practices by using Participatory Input to Navigate the Machine Learning Multiverse. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pages 1–30, New York, NY, USA. Association for Computing Machinery. doi: 10.1145/3706598.3713482 URL https://doi.org/10.1145/3706598.3713482

Simson, J., Fabris, A., Fröhner, C., Kreuter, F. & Kern, C. (2024). Bias Begins with Data: The FairGround Corpus for Robust and Reproducible Research on Algorithmic Fairness. *Preprint*. doi: 10.48550/arXiv.2510 URL https://doi.org/10.48550/arXiv.2510.22363

Simson, J., Kononykhina, O., & Schierholz, M. (2023). occupationMeasurement: A Comprehensive Toolbox for Interactive Occupation Coding in Surveys. In *Journal of Open Source Software*, 8(88), 5505. doi: 10.21105/joss.05505 URL https://doi.org/10.21105/joss.05505

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A Reductions Approach to Fair Classification. In *Proceedings of the 35th International Conference on Machine Learning*, pages 60–69. PMLR, July 2018. URL https://proceedings.mlr.press/v80/agarwal18a.html.

Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 120–129. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/agarwal19d.html.

Akira Kurosawa. Rashômon, August 1950.

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

Jerone Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, and Alice Xiang. Ethical Considerations for Responsible Data Curation. *Advances in Neural Information Processing Systems*, 36:55320–55360, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ad3ebc951f43d1e9ed20187a7b5bc4ee-Abstract-Datasets_and_Benchmarks.html.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, pages 254–264, May 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Sherry R. Arnstein. A Ladder Of Citizen Participation. *Journal of the American Institute of Planners*, 35(4):216–224, July 1969. ISSN 0002-8991. doi: 10.1080/01944366908977225. URL http://www.tandfonline.com/doi/abs/10.1080/01944366908977225.

Daniel Atherton. Incident number 608. *AI Incident Database*, 2023. URL https://incidentdatabase.ai/cite/608.

Katrin Auspurg and Josef Brüderl. Has the Credibility of the Social Sciences Been Credibly Destroyed? Reanalyzing the "Many Analysts, One Data Set" Project. *Socius*, 7: 23780231211024421, January 2021. ISSN 2378-0231. doi: 10.1177/23780231211024421. URL https://doi.org/10.1177/23780231211024421.

Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The Moral Machine experiment. *Nature*, 563(7729): 59–64, November 2018. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-018-0637-6. URL http://www.nature.com/articles/s41586-018-0637-6.

Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks, April 2022. URL http://arxiv.org/abs/2106.05498.

Solon Barocas and Andrew D. Selbst. Big Data's Disparate Impact. *SSRN Electronic Journal*, 2016. ISSN 1556-5068. doi: 10.2139/ssrn.2477899. URL https://www.ssrn.com/abstract=2477899.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023. ISBN 978-0-262-04861-3. URL https://fairmlbook.org/.

Denis Baylor, Kevin Haas, Konstantinos Katsiapis, Sammy Leong, Rose Liu, Clemens Menwald, Hui Miao, Neoklis Polyzotis, Mitchell Trott, and Martin Zinkevich. Continuous Training for Production ML in the TensorFlow Extended (TFX) Platform. In *2019 USENIX Conference on Operational Machine Learning (OpML 19)*, pages 51–53, 2019. ISBN 978-1-939133-00-7. URL https://www.usenix.org/conference/opml19/presentation/baylor.

Jacob Beck, Stephanie Eckman, Bolei Ma, Rob Chew, and Frauke Kreuter. Order Effects in Annotation Tasks: Further Evidence of Annotation Sensitivity. In Raúl Vázquez, Hande Celikkanat, Dennis Ulmer, Jörg Tiedemann, Swabha Swayamdipta, Wilker Aziz, Barbara Plank, Joris Baan, and Marie-Catherine de Marneffe, editors, *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 81–86, St Julians, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.uncertainlp-1.8/.

Jacob Beck, Stephanie Eckman, Christoph Kern, and Frauke Kreuter. Bias in the Loop: How Humans Evaluate AI-Generated Suggestions, September 2025. URL http://arxiv.org/abs/2509.08514.

Samuel J. Bell, Onno Kampman, Jesse Dodge, and Neil Lawrence. Modeling the Machine Learning Multiverse. *Advances in Neural Information Processing Systems*, 35:18416–18429, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/750337e1301941f81ae31a90e0a1c181-Abstract-Conference.html.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018. URL https://arxiv.org/abs/1810.01943.

James Bergstra and Yoshua Bengio. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13(10):281–305, 2012. ISSN 1533-7928. URL http://jmlr.org/papers/v13/bergstra12a.html.

Bibliothèque nationale de France. Je participe, tu participes, il participe, nous participons, vous participez, ils profitent (avec personnages), May 1968. URL https://commons.wikimedia.org/wiki/File:Mai_1968._Je_participe,_tu_participes,_il_participe,_nous_participons,_vous_participez,_ils_profitent_(avec_personnages)_affiche_(Variante),_non_identifi%C3%A9.jpg.

References

Martin Binder, Florian Pfisterer, Michel Lang, Lennart Schneider, Lars Kotthoff, and Bernd Bischl. Mlr3pipelines - Flexible Machine Learning Pipelines in R. *Journal of Machine Learning Research*, 22(184):1–7, 2021. ISSN 1533-7928. URL http://jmlr.org/papers/v22/21-0281.html.

Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–14, New York, NY, USA, April 2018. Association for Computing Machinery. ISBN 978-1-4503-5620-6. doi: 10.1145/3173574.3173951. URL https://doi.org/10.1145/3173574.3173951.

Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546, January 2021. doi: 10.1109/WACV48630.2021.00158. URL https://ieeexplore.ieee.org/abstract/document/9423393.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. Power to the People? Opportunities and Challenges for Participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8, Arlington VA USA, October 2022a. ACM. ISBN 978-1-4503-9477-2. doi: 10.1145/3551624.3555290. URL https://dl.acm.org/doi/10.1145/3551624.3555290.

Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. The Forgotten Margins of AI Ethics. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 948–958, New York, NY, USA, June 2022b. Association for Computing Machinery. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533157. URL https://dl.acm.org/doi/10.1145/3531146.3533157.

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. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Mining and Knowledge Discovery*, 13(2), March 2023. ISSN 1942-4787, 1942-4795. doi: 10.1002/widm.1484. URL https://onlinelibrary.wiley.com/doi/10.1002/widm.1484.

Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 850–863, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533149. URL https://doi.org/10.1145/3531146.3533149.

Emily Black, Talia Gillis, and Zara Yasmine Hall. D-hacking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 602–615, Rio de Janeiro Brazil, June 2024a. ACM. ISBN 979-8-4007-0450-5. doi: 10.1145/3630106.3658928. URL https://dl.acm.org/doi/10.1145/3630106.3658928.

Emily Black, Logan Koepke, Pauline Kim, Solon Barocas, and Mingwei Hsu. The Legal Duty to Search for Less Discriminatory Algorithms, June 2024b. URL http://arxiv.org/abs/2406.06817.

References

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001. doi: 10.1214/ss/1009213726. URL https://doi.org/10.1214/ss/1009213726.

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, Alejandro Ecker, Achim Edelmann, Maureen A. Eger, Simon Ellerbrock, Anna Forke, Andrea 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. 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, November 2022. doi: 10.1073/pnas.2203150119. URL https://www.pnas.org/doi/10.1073/pnas.2203150119.

Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-making in Child Welfare Services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, Glasgow Scotland Uk, May 2019. ACM. ISBN 978-1-4503-5970-2. doi: 10.1145/3290605.3300271. URL https://dl.acm.org/doi/10.1145/3290605.3300271.

Meta S. Brown. What IT Needs To Know About The Data Mining Process, July 2015. URL https://www.forbes.com/sites/metabrown/2015/07/29/

what-it-needs-to-know-about-the-data-mining-process/.

Bundesagentur für Arbeit. *Klassifikation der Berufe 2010*, volume 1 & 2. Bundesagentur für Arbeit, Nürnberg, 2011.

Jenna Burrell. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1):2053951715622512, June 2016. ISSN 2053-9517. doi: 10.1177/2053951715622512. URL https://doi.org/10.1177/2053951715622512.

Antonio A. Casilli, Paola Tubaro, Maxime Cornet, Clément Le Ludec, Juana Torres-Cierpe, and Matheus Viana Braz. Global inequalities in the production of artificial intelligence: A four-country study on data work, 2024. URL https://arxiv.org/abs/2410.14230.

Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1):4209, March 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-07939-1. URL https://www.nature.com/articles/s41598-022-07939-1.

Simon Caton, Saiteja Malisetty, and Christian Haas. Impact of Imputation Strategies on Fairness in Machine Learning. *Journal of Artificial Intelligence Research*, 74, September 2022. ISSN 1076-9757. doi: 10.1613/jair.1.13197. URL https://doi.org/10.1613/jair.1.13197.

L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 319–328, New York, NY, USA, January 2019. Association for Computing Machinery. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287586. URL https://doi.org/10.1145/3287560.3287586.

Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. CRISP-DM 1.0: Step-by-step data mining guide, 2000.

Chaofan Chen, Kangcheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang. An interpretable model with globally consistent explanations for credit risk. *CoRR*, abs/1811.12615, 2018. URL http://arxiv.org/abs/1811.12615.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 339–348, New York, NY, USA, January 2019. Association for Computing Machinery. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287594. URL https://doi.org/10.1145/3287560.3287594.

Jaemin Cho, Abhay Zala, and Mohit Bansal. DALL-EVAL: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3020–3031, Paris, France, October 2023. IEEE. ISBN 979-8-3503-0718-4. doi: 10.1109/ICCV51070.2023.00283. URL https://ieeexplore.ieee.org/document/10378621/.

Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM*, 63(5):82–89, April 2020. ISSN 0001-0782. doi: 10.1145/3376898. URL https://dl.acm.org/doi/10.1145/3376898.

Christian Sandvig. Seeing the Sort: The Aesthetic and Industrial Defense of "The Algorithm", November 2014. URL https://median.newmediacaucus.org/art-infrastructures-information/seeing-the-sort-the-aesthetic-and-industrial-defense-of-the-algorithm/.

Nancy S. Cole and Michael J. Zieky. The New Faces of Fairness. *Journal of Educational Measurement*, 38(4):369–382, 2001. ISSN 1745-3984. doi: 10.1111/j.1745-3984.2001.tb01132.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-3984.2001.tb01132.x.

A. Feder Cooper, Katherine Lee, Madiha Zahrah Choksi, Solon Barocas, Christopher De Sa, James Grimmelmann, Jon Kleinberg, Siddhartha Sen, and Baobao Zhang. Arbitrariness and Social Prediction: The Confounding Role of Variance in Fair Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22004–22012, March 2024. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v38i20.30203. URL https://doi.org/10.1609/aaai.v38i20.30203.

Eric Corbett, Emily Denton, and Sheena Erete. Power and Public Participation in AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–13, Boston MA USA, October 2023. ACM. ISBN 979-8-4007-0381-2. doi: 10.1145/3617694.3623228. URL https://dl.acm.org/doi/10.1145/3617694.3623228.

Stefano Coretta, Joseph V. Casillas, Simon Roessig, Michael Franke, Byron Ahn, Ali H. Al-Hoorie, Jalal Al-Tamimi, Najd E. Alotaibi, Mohammed K. AlShakhori, Ruth M. Altmiller, Pablo Arantes, Angeliki Athanasopoulou, Melissa M. Baese-Berk, George Bailey, Cheman Baira A Sangma, Eleonora J. Beier, Gabriela M. Benavides, Nicole Benker, Emelia P. BensonMeyer, Nina R. Benway, Grant M. Berry, Liwen Bing, Christina Bjorndahl, Mariška Bolyanatz, Aaron Braver, Violet A. Brown, Alicia M. Brown, Alejna Brugos, Erin M. Buchanan, Tanna Butlin, Andrés Buxó-Lugo, Coline Caillol, Francesco Cangemi, Christopher Carignan, Sita Carraturo, Tiphaine Caudrelier, Eleanor Chodroff, Michelle Cohn, Johanna Cronenberg, Olivier Crouzet, Erica L. Dagar, Charlotte Dawson, Carissa A. Diantoro, Marie Dokovova, Shiloh Drake, Fengting Du, Margaux Dubuis, Florent Duême, Matthew Durward, Ander Egurtzegi, Mahmoud M. Elsherif, Janina Esser, Emmanuel Ferragne, Fernanda Ferreira, Lauren K. Fink, Sara Finley, Kurtis Foster, Paul Foulkes, Rosa Franzke, Gabriel Frazer-McKee, Robert Fromont, Christina García, Jason Geller, Camille L. Grasso, Pia Greca, Martine Grice, Magdalena S. Grose-Hodge, Amelia J. Gully, Caitlin Halfacre, Ivy Hauser, Jen Hay, Robert Haywood, Sam Hellmuth, Allison I. Hilger, Nicole Holliday, Damar Hoogland, Yaqian Huang, Vincent Hughes, Ane Icardo Isasa, Zlatomira G. Ilchovska, Hae-Sung Jeon, Jacq Jones, Mágat N. Junges, Stephanie Kaefer, Constantijn Kaland, Matthew C. Kelley, Niamh E. Kelly, Thomas Kettig, Ghada Khattab, Ruud Koolen, Emiel Krahmer, Dorota Krajewska, Andreas Krug, Abhilasha A. Kumar, Anna Lander, Tomas O. Lentz, Wanyin Li, Yanyu Li, Maria Lialiou, Ronaldo M. LimaJr., Justin J. H. Lo, Julio Cesar Lopez Otero, Bradley Mackay, Bethany MacLeod, Mel Mallard, Carol-Ann Mary McConnellogue, George Moroz, Mridhula Murali, Ladislas Nalborczyk, Filip Nenadić, Jessica Nieder, Dušan Nikolić, Francisco G. S. Nogueira, Heather M. Offerman, Elisa Passoni, Maud Pélissier, Scott J. Perry, Alexandra M. Pfiffner, Michael Proctor, Ryan Rhodes, Nicole Rodríguez, Elizabeth Roepke, Jan P. Röer, Lucia Sbacco, Rebecca Scarborough, Felix Schaeffler, Erik Schleef, Dominic Schmitz, Alexander Shiryaev, Márton Sóskuthy, Malin Spaniol, Joseph A. Stanley, Alyssa Strickler, Alessandro Tavano, Fabian Tomaschek, Benjamin V. Tucker, Rory Turnbull, Kingsley O. Ugwuanyi, Iñigo Urrestarazu-Porta, Ruben van de Vijver, Kristin J. Van Engen, Emiel van Miltenburg, Bruce Xiao Wang, Natasha Warner, Simon Wehrle, Hans

Westerbeek, Seth Wiener, Stephen Winters, Sidney G.-J. Wong, Anna Wood, Jane Wottawa, Chenzi Xu, Germán Zárate-Sández, Georgia Zellou, Cong Zhang, Jian Zhu, and Timo B. Roettger. Multidimensional Signals and Analytic Flexibility: Estimating Degrees of Freedom in Human-Speech Analyses. *Advances in Methods and Practices in Psychological Science*, 6 (3):25152459231162567, July 2023. ISSN 2515-2459. doi: 10.1177/25152459231162567. URL https://doi.org/10.1177/25152459231162567.

André F. Cruz and Moritz Hardt. Unprocessing Seven Years of Algorithmic Fairness, March 2024. URL http://arxiv.org/abs/2306.07261.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. ISSN 1533-7928. URL http://jmlr.org/papers/v23/20-1335.html.

Ankolika De, Shaheen Kanthawala, and Jessica Maddox. Who Gets Heard? Calling Out the "Hard-to-Reach" Myth for Non-WEIRD Populations' Recruitment and Involvement in Research. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pages 855–867, New York, NY, USA, June 2025. Association for Computing Machinery. ISBN 979-8-4007-1482-5. doi: 10.1145/3715275.3732055. URL https://dl.acm.org/doi/10.1145/3715275.3732055.

MaryBeth Defrance, Maarten Buyl, and Tijl De Bie. ABCFair: An Adaptable Benchmark approach for Comparing Fairness Methods. *Advances in Neural Information Processing Systems*, 37:40145–40163, December 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/46bae562da84d63269673808e8eff777-Abstract-Datasets_and_Benchmarks_Track.html.

Marco Del Giudice and Steven W. Gangestad. A Traveler's Guide to the Multiverse: Promises, Pitfalls, and a Framework for the Evaluation of Analytic Decisions. *Advances in Methods and Practices in Psychological Science*, 4(1):2515245920954925, January 2021. ISSN 2515-2459. doi: 10.1177/2515245920954925. URL https://doi.org/10.1177/2515245920954925.

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–23, Boston MA USA, October 2023. ACM. ISBN 979-8-4007-0381-2. doi: 10.1145/3617694.3623261. URL https://dl.acm.org/doi/10.1145/3617694.3623261.

Kerstin Denecke, Elia Gabarron, Rebecca Grainger, Stathis Th. Konstantinidis, Annie Lau, Octavio Rivera-Romero, Talya Miron-Shatz, and Mark Merolli. Artificial Intelligence for Participatory Health: Applications, Impact, and Future Implications: Contribution of the IMIA Participatory Health and Social Media Working Group. *Yearbook of Medical Informatics*, 28

(01):165–173, August 2019. ISSN 0943-4747, 2364-0502. doi: 10.1055/s-0039-1677902. URL http://www.thieme-connect.de/DOI/DOI?10.1055/s-0039-1677902.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848. URL https://ieeexplore.ieee.org/abstract/document/5206848.

Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Remi Denton. CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2342–2351, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3534647. URL https://dl.acm.org/doi/10.1145/3531146.3534647.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring Adult: New Datasets for Fair Machine Learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 6478–6490. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/32e54441e6382a7fbacbbbaf3c450059-Abstract.html.

Jiayun Dong and Cynthia Rudin. Variable Importance Clouds: A Way to Explore Variable Importance for the Set of Good Models, February 2020. URL https://doi.org/10.48550/arXiv.1901.03209.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, NY, USA, January 2012. Association for Computing Machinery. ISBN 978-1-4503-1115-1. doi: 10.1145/2090236.2090255. URL https://doi.org/10.1145/2090236.2090255.

Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data, March 2020. URL http://arxiv.org/abs/2003.06505.

European Parliament, Council, and Commission. Charter of Fundamental Rights of the European Union, December 2007. URL http://data.europa.eu/eli/treaty/char_2007/oj/eng.

Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, September 2022. ISSN 1573-756X. doi: 10.1007/s10618-022-00854-z.

Alessandro Fabris, Nina Baranowska, Matthew J. Dennis, David Graus, Philipp Hacker, Jorge Saldivar, Frederik Zuiderveen Borgesius, and Asia J. Biega. Fairness and Bias in Algorithmic Hiring: A Multidisciplinary Survey. *ACM Trans. Intell. Syst. Technol.*, 16(1):16:1–16:54, January 2025. ISSN 2157-6904. doi: 10.1145/3696457. URL https://dl.acm.org/doi/10.1145/3696457.

Evanthia Faliagka, Kostas Ramantas, Athanasios Tsakalidis, and Giannis Tzimas. Application of machine learning algorithms to an online recruitment system. In *Proc. International Conference on Internet and Web Applications and Services*, pages 215–220, 2012. URL https://personales.upv.es/thinkmind/dl/conferences/iciw/iciw_2012/iciw_2012_7_40_20189.pdf.

Michael Feffer, Hoda Heidari, and Zachary C. Lipton. Moral Machine or Tyranny of the Majority? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(5):5974–5982, June 2023a. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v37i5.25739. URL https://ojs.aaai.org/index.php/AAAI/article/view/25739.

Michael Feffer, Michael Skirpan, Zachary Lipton, and Hoda Heidari. From Preference Elicitation to Participatory ML: A Critical Survey & Guidelines for Future Research. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pages 38–48, New York, NY, USA, August 2023b. Association for Computing Machinery. ISBN 979-8-4007-0231-0. doi: 10.1145/3600211.3604661. URL https://dl.acm.org/doi/10.1145/3600211.3604661.

Li Fei-Fei and Ranjay Krishna. Searching for computer vision north stars. *Daedalus*, 151(2):85–99, 2022. URL https://direct.mit.edu/daed/article-abstract/151/2/85/110602.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 259–268, New York, NY, USA, August 2015. Association for Computing Machinery. ISBN 978-1-4503-3664-2. doi: 10.1145/2783258.2783311. URL https://dl.acm.org/doi/10.1145/2783258.2783311.

Matthias Feurer and Frank Hutter. Hyperparameter Optimization. In Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren, editors, *Automated Machine Learning: Methods, Systems, Challenges*, The Springer Series on Challenges in Machine Learning, pages 3–33. Springer International Publishing, Cham, 2019. ISBN 978-3-030-05318-5. doi: 10.1007/978-3-030-05318-5_1. URL https://doi.org/10.1007/978-3-030-05318-5_1.

Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. Efficient and Robust Automated Machine Learning. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/hash/11d0e6287202fced83f79975ec59a3a6-Abstract.html.

Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 329–338, Atlanta GA USA, January 2019. ACM. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287589. URL https://dl.acm.org/doi/10.1145/3287560.3287589.

Prakhar Ganesh, Afaf Taik, and Golnoosh Farnadi. Systemizing Multiplicity: The Curious Case of Arbitrariness in Machine Learning. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(2):1032–1048, October 2025. ISSN 3065-8365. doi: 10.1609/aies.v8i2.36609. URL https://ojs.aaai.org/index.php/AIES/article/view/36609.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, December 2021. ISSN 0001-0782, 1557-7317. doi: 10.1145/3458723. URL https://dl.acm.org/doi/10.1145/3458723.

Seymour Geisser. *Predictive Inference*. Chapman and Hall/CRC, New York, November 2017. ISBN 978-0-203-74231-0. doi: 10.1201/9780203742310. URL https://doi.org/10.1201/9780203742310.

Andrew Gelman and Eric Loken. The Statistical Crisis in Science. *American Scientist*, 102(6):460, 2014. URL https://sites.stat.columbia.edu/gelman/research/published/ForkingPaths.pdf.

Travis Greene, Galit Shmueli, Jan Fell, Ching-Fu Lin, and Han-Wei Liu. Forks Over Knives: Predictive Inconsistency in Criminal Justice Algorithmic Risk Assessment Tools. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(Supplement_2):S692–S723, December 2022. ISSN 0964-1998. doi: 10.1111/rssa.12966. URL https://doi.org/10.1111/rssa.12966.

Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS symposium on machine learning and the law*, volume 1, page 11. Barcelona, Spain, 2016. URL https://www.mlandthelaw.org/papers/grgic.pdf.

Ulrike Grömping. South German Credit Data: Correcting a Widely Used Data Set. *Reports in Mathematics, Physics and Chemistry*, 04/2019, November 2019. ISSN 2190-3913. URL http://www1.beuth-hochschule.de/FB_II/reports/Report-2019-004.pdf.

Odd Erik Gundersen and Sigbjørn Kjensmo. State of the Art: Reproducibility in Artificial Intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. ISSN 2374-3468. doi: 10.1609/aaai.v32i1.11503. URL https://ojs.aaai.org/index.php/AAAI/article/view/11503.

Aaron Halfaker and R. Stuart Geiger. ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2): 1–37, October 2020. ISSN 2573-0142. doi: 10.1145/3415219. URL https://dl.acm.org/doi/10.1145/3415219.

Moritz Hardt. *The Emerging Science of Machine Learning Benchmarks*. 2025. URL https://mlbenchmarks.org. Online manuscript.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

Galen Harrison, Kevin Bryson, Ahmad Emmanuel Balla Bamba, Luca Dovichi, Aleksander Hermann Binion, Arthur Borem, and Blase Ur. JupyterLab in Retrograde: Contextual Notifications That Highlight Fairness and Bias Issues for Data Scientists. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–19, New York, NY, USA, May 2024. Association for Computing Machinery. ISBN 979-8-4007-0330-0. doi: 10.1145/3613904.3642755. URL https://dl.acm.org/doi/10.1145/3613904.3642755.

Joshua K. Hartshorne, Joshua R. de Leeuw, Noah D. Goodman, Mariela Jennings, and Timothy J. O'Donnell. A thousand studies for the price of one: Accelerating psychological science with Pushkin. *Behavior Research Methods*, February 2019. ISSN 1554-3528. doi: 10.3758/s13428-018-1155-z. URL https://doi.org/10.3758/s13428-018-1155-z.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1939–1948. PMLR, July 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Jon Henley. Dutch government faces collapse over child benefits scandal. *The Guardian*, January 2021. ISSN 0261-3077. URL https://www.theguardian.com/world/2021/jan/14/dutch-government-faces-collapse-over-child-benefits-scandal.

Joseph Henrich, Steven J. Heine, and Ara Norenzayan. Most people are not WEIRD. *Nature*, 466 (7302):29–29, July 2010. ISSN 1476-4687. doi: 10.1038/466029a. URL https://www.nature.com/articles/466029a.

Luke Henriques-Gomes. Robodebt: Five years of lies, mistakes and failures that caused a $1.8bn scandal. *The Guardian*, March 2023. ISSN 0261-3077. URL https://www.theguardian.com/australia-news/2023/mar/11/robodebt-five-years-of-lies-mistakes-and-failures-that-caused-a-18bn-scandal.

Moritz Herrmann, F. Julian D. Lange, Katharina Eggensperger, Giuseppe Casalicchio, Marcel Wever, Matthias Feurer, David Rügamer, Eyke Hüllermeier, Anne-Laure Boulesteix, and Bernd Bischl. Position: Why We Must Rethink Empirical Research in Machine Learning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 18228–18247. PMLR, July 2024. URL https://proceedings.mlr.press/v235/herrmann24b.html.

Hans Hofmann. Statlog (german credit data). UCI Machine Learning Repository, 1994. URL https://doi.org/10.24432/C5NC77.

Giles Hooker. Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, 2007. ISSN 1061-8600. URL https://www.jstor.org/stable/27594267.

Aspen Hopkins and Serena Booth. Machine Learning Practices Outside Big Tech: How Resource Constraints Challenge Responsible Development. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pages 134–145, New York, NY, USA, July 2021. Association for Computing Machinery. ISBN 978-1-4503-8473-5. doi: 10.1145/3461702.3462527. URL https://dl.acm.org/doi/10.1145/3461702.3462527.

Hsiang Hsu and Flavio Calmon. Rashomon Capacity: A Metric for Predictive Multiplicity in Classification. *Advances in Neural Information Processing Systems*, 35:28988–29000, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ba4caa85ecdcafbf9102ab8ec384182d-Paper-Conference.pdf.

Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. Collective Constitutional AI: Aligning a Language Model with Public Input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 1395–1417, New York, NY, USA, June 2024. Association for Computing Machinery. ISBN 979-8-4007-0450-5. doi: 10.1145/3630106.3658979. URL https://dl.acm.org/doi/10.1145/3630106.3658979.

Matthew Hutson. Artificial intelligence faces reproducibility crisis. *Science*, 359(6377):725–726, February 2018. doi: 10.1126/science.359.6377.725. URL https://www.science.org/doi/10.1126/science.359.6377.725.

Frank Hutter, Holger Hoos, and Kevin Leyton-Brown. An Efficient Approach for Assessing Hyperparameter Importance. In *Proceedings of the 31st International Conference on Machine Learning*, pages 754–762. PMLR, January 2014. URL https://proceedings.mlr.press/v32/hutter14.html.

Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren, editors. *Automated Machine Learning: Methods, Systems, Challenges*. Springer Nature, 2019. doi: 10.1007/978-3-030-05318-5. URL https://library.oapen.org/handle/20.500.12657/23012.

ILO. *International Standard Classification of Occupations 2008 (ISCO-08): Structure, Group Definitions and Correspondence Tables*. International Labour Office, May 2012. ISBN 978-92-2-125953-4. URL http://www.ilo.org/global/publications/ilo-bookstore/order-online/books/WCMS_172572/lang--en/index.htm.

John P. A. Ioannidis. Why Most Published Research Findings Are False. *PLOS Medicine*, 2 (8):e124, August 2005. ISSN 1549-1676. doi: 10.1371/journal.pmed.0020124. URL https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124.

Rashidul Islam, Shimei Pan, and James R. Foulds. Can We Obtain Fairness For Free? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 586–596, Virtual Event USA, July 2021. ACM. ISBN 978-1-4503-8473-5. doi: 10.1145/3461702.3462614. URL https://doi.org/10.1145/3461702.3462614.

Logan S. James, Sarah C. Woolley, Jon T. Sakata, Courtney B. Hilton, Michael J. Ryan, and Samuel A. Mehr. Humans share acoustic preferences with other animals, June 2025. URL https://www.biorxiv.org/content/10.1101/2025.06.26.661759v1.

Kenneth Jensen. English: A diagram showing the relationship between the different phases of CRISP-DM and illustrates the recursive nature of a data mining project., April 2012. URL https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png.

Leslie K. John, George Loewenstein, and Drazen Prelec. Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, 23(5):524–532, May 2012. ISSN 0956-7976. doi: 10.1177/0956797611430953. URL https://doi.org/10.1177/0956797611430953.

Julia Sophie Broden. Conformal Prediction and the Many-Worlds of Fairness. Master's thesis, Ludwig-Maximilians-Universität München, München, Germany, August 2025.

Mitchel Kappen and Marnix Naber. Objective and bias-free measures of candidate motivation during job applications. *Scientific Reports*, 11(1):21254, November 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-00659-y. URL https://www.nature.com/articles/s41598-021-00659-y.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2564–2572. PMLR, July 2018. URL https://proceedings.mlr.press/v80/kearns18a.html.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 100–109, New York, NY, USA, 2019. Association for Computing

Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287592. URL https://doi.org/10.1145/3287560.3287592.

Christoph Kern, Frederic Gerdon, Ruben L. Bach, Florian Keusch, and Frauke Kreuter. Humans versus machines: Who is perceived to decide fairer? Experimental evidence on attitudes toward automated decision-making. *Patterns*, 3(10), October 2022. ISSN 2666-3899. doi: 10.1016/j.patter.2022.100591. URL https://www.cell.com/patterns/abstract/S2666-3899(22)00209-4.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic Fairness. *AEA Papers and Proceedings*, 108:22–27, May 2018. ISSN 2574-0768. doi: 10.1257/pandp.20181018. URL https://www.aeaweb.org/articles?id=10.1257/pandp.20181018.

Ron Kohavi. Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 202–207, Portland, Oregon, August 1996. AAAI Press. URL http://www.aaai.org/Papers/KDD/1996/KDD96-033.pdf.

Lars Kotthoff, Chris Thornton, Holger H. Hoos, Frank Hutter, and Kevin Leyton-Brown. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *Journal of Machine Learning Research*, 18(25):1–5, 2017. ISSN 1533-7928. URL http://jmlr.org/papers/v18/16-261.html.

Max Kuhn, Hadley Wickham, and Emil Hvitfeldt. *recipes: Preprocessing and Feature Engineering Steps for Modeling*, 2025. URL https://CRAN.R-project.org/package=recipes. R package version 1.3.1.

Bogdan Kulynych, David Madras, Smitha Milli, Inioluwa Deborah Raji, Angela Zhou, and Richard Zemel. Participatory approaches to machine learning. In *International Conference on Machine Learning Workshop*, volume 7, 2020. URL https://participatoryml.github.io/.

Matthias Kuppler, Christoph Kern, Ruben L. Bach, and Frauke Kreuter. From fair predictions to just decisions? Conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making. *Frontiers in Sociology*, 7, October 2022. ISSN 2297-7775. doi: 10.3389/fsoc.2022.883999. URL https://www.frontiersin.org/journals/sociology/articles/10.3389/fsoc.2022.883999/full.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Sean Laraway, Susan Snycerski, Sean Pradhan, and Bradley E. Huitema. An Overview of Scientific Reproducibility: Consideration of Relevant Issues for Behavior Science/Analysis. *Perspectives on Behavior Science*, 42(1):33–57, March 2019. ISSN 2520-8977. doi: 10.1007/s40614-019-00193-3. URL https://doi.org/10.1007/s40614-019-00193-3.

Mike Laszkiewicz, Imant Daunhawer, Julia E. Vogt, Asja Fischer, and Johannes Lederer. Benchmarking the Fairness of Image Upsampling Methods. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 489–517, New York, NY, USA, June 2024. Association for Computing Machinery. ISBN 979-8-4007-0450-5. doi: 10.1145/3630106.3658921. URL https://dl.acm.org/doi/10.1145/3630106.3658921.

Edward E. Leamer. Sensitivity Analyses Would Help. *The American Economic Review*, 75(3): 308–313, 1985. ISSN 0002-8282. URL https://www.jstor.org/stable/1814801.

Michelle Seng Ah Lee, Luciano Floridi, and Jatinder Singh. Formalising trade-offs beyond algorithmic fairness: Lessons from ethical philosophy and welfare economics. *AI and Ethics*, 1(4):529–544, November 2021. ISSN 2730-5961. doi: 10.1007/s43681-021-00067-y. URL https://doi.org/10.1007/s43681-021-00067-y.

Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):181:1–181:35, November 2019. doi: 10.1145/3359283. URL https://dl.acm.org/doi/10.1145/3359283.

Edo Liberty, Zohar Karnin, Bing Xiang, Laurence Rouesnel, Baris Coskun, Ramesh Nallapati, Julio Delgado, Amir Sadoughi, Yury Astashonok, Piali Das, Can Balioglu, Saswata Chakravarty, Madhav Jha, Philip Gautier, David Arpin, Tim Januschowski, Valentin Flunkert, Yuyang Wang, Jan Gasthaus, Lorenzo Stella, Syama Rangapuram, David Salinas, Sebastian Schelter, and Alex Smola. Elastic Machine Learning Algorithms in Amazon SageMaker. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, SIGMOD '20, pages 731–737, New York, NY, USA, May 2020. Association for Computing Machinery. ISBN 978-1-4503-6735-6. doi: 10.1145/3318464.3386126. URL https://doi.org/10.1145/3318464.3386126.

Yang Liu, Alex Kale, Tim Althoff, and Jeffrey Heer. Boba: Authoring and Visualizing Multiverse Analyses. *IEEE Transactions on Visualization and Computer Graphics*, 27(2): 1753–1763, February 2021. ISSN 1941-0506. doi: 10.1109/TVCG.2020.3028985. URL https://ieeexplore.ieee.org/abstract/document/9216579.

Carol Long, Hsiang Hsu, Wael Alghamdi, and Flavio Calmon. Individual Arbitrariness and Group Fairness. *Advances in Neural Information Processing Systems*, 36:68602–68624, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/d891d240b5784656a0356bf4b00f5cdd-Paper-Conference.pdf.

Alexandra Sasha Luccioni, Frances Corry, Hamsini Sridharan, Mike Ananny, Jason Schultz, and Kate Crawford. A Framework for Deprecating Datasets: Standardizing Documentation, Identification, and Communication. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 199–212, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533086. URL https://dl.acm.org/doi/10.1145/3531146.3533086.

Sean McGregor. Ai incident database, 2025. URL https://incidentdatabase.ai/. Retrieved October 2025 from https://incidentdatabase.ai/.

Kristof Meding and Thilo Hagendorff. Fairness Hacking: The Malicious Practice of Shrouding Unfairness in Algorithms. *Philosophy & Technology*, 37(1):4, January 2024. ISSN 2210-5441. doi: 10.1007/s13347-023-00679-8. URL https://doi.org/10.1007/s13347-023-00679-8.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6):115:1–115:35, July 2021. ISSN 0360-0300. doi: 10.1145/3457607. URL https://doi.org/10.1145/3457607.

Michelle M. Mello and Sherri Rose. Denial—Artificial Intelligence Tools and Health Insurance Coverage Decisions. *JAMA Health Forum*, 5(3):e240622, March 2024. ISSN 2689-0186. doi: 10.1001/jamahealthforum.2024.0622. URL https://doi.org/10.1001/jamahealthforum.2024.0622.

Anna P. Meyer, Aws Albarghouthi, and Loris D'Antoni. The Dataset Multiplicity Problem: How Unreliable Data Impacts Predictions. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 193–204, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3593988. URL https://doi.org/10.1145/3593013.3593988.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, 8(Volume 8, 2021):141–163, March 2021. ISSN 2326-8298, 2326-831X. doi: 10.1146/annurev-statistics-042720-125902. URL https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-042720-125902.

Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014. doi: 10.1016/j.dss.2014.03.001. URL https://doi.org/10.1016/j.dss.2014.03.001.

Kathleen L. Mosier, Linda J. Skitka, Susan Heers, and Mark Burdick. Automation Bias: Decision Making and Performance in High-Tech Cockpits. *The International Journal of Aviation Psychology*, 8(1):47–63, January 1998. ISSN 1050-8414. doi: 10.1207/s15327108ijap0801_3. URL https://doi.org/10.1207/s15327108ijap0801_3.

Amitabha Mukerjee, Rita Biswas, Kalyanmoy Deb, and Amrit P. Mathur. Multi–objective Evolutionary Algorithms for the Risk–return Trade–off in Bank Loan Management. *International Transactions in Operational Research*, 9(5):583–597, 2002. ISSN 1475-3995. doi: 10.1111/1475-3995.00375. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-3995.00375.

Michael J. Muller and Sarah Kuhn. Participatory design. *Communications of the ACM*, 36 (6):24–28, June 1993. ISSN 0001-0782, 1557-7317. doi: 10.1145/153571.255960. URL https://dl.acm.org/doi/10.1145/153571.255960.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, October 2019. doi: 10.1126/science.aax2342. URL https://www.science.org/doi/10.1126/science.aax2342.

Chinasa T. Okolo, Nicola Dell, and Aditya Vashistha. Making AI Explainable in the Global South: A Systematic Review. In *Proceedings of the 5th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, COMPASS '22, pages 439–452, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9347-8. doi: 10.1145/3530190.3534802. URL https://dl.acm.org/doi/10.1145/3530190.3534802.

Randal S. Olson and Jason H. Moore. Identifying and Harnessing the Building Blocks of Machine Learning Pipelines for Sensible Initialization of a Data Science Automation Tool. In Rick Riolo, Bill Worzel, Brian Goldman, and Bill Tozier, editors, *Genetic Programming Theory and Practice XIV*, pages 211–223. Springer International Publishing, Cham, 2018. ISBN 978-3-319-97088-2. doi: 10.1007/978-3-319-97088-2_14. URL https://doi.org/10.1007/978-3-319-97088-2_14.

References

OPEN SCIENCE COLLABORATION. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, August 2015. doi: 10.1126/science.aac4716. URL https://www.science.org/doi/10.1126/science.aac4716.

Chirag J. Patel, Belinda Burford, and John P.A Ioannidis. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of clinical epidemiology*, 68(9):1046–1058, September 2015. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2015.05.029. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4555355/.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. ISSN 1533-7928. URL http://jmlr.org/papers/v12/pedregosa11a.html.

Periklis Perikleous, Andreas Kafkalias, Zenonas Theodosiou, Pinar Barlas, Evgenia Christoforou, Jahna Otterbacher, Gianluca Demartini, and Andreas Lanitis. How Does the Crowd Impact the Model? A Tool for Raising Awareness of Social Bias in Crowdsourced Training Data. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4951–4954, Atlanta GA USA, October 2022. ACM. ISBN 978-1-4503-9236-5. doi: 10.1145/3511808.3557178. URL https://dl.acm.org/doi/10.1145/3511808.3557178.

Nederlandse Autoriteit Persoonsgegevens and Belastingdienst. Toeslagen-de verwerking van de nationaliteit van aanvragers van kinderopvangtoeslag, 2021. URL https://www.autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/onderzoek_belastingdienst_kinderopvangtoeslag.pdf.

Dana Pessach and Erez Shmueli. A Review on Fairness in Machine Learning. *ACM Computing Surveys*, 55(3):51:1–51:44, February 2022. ISSN 0360-0300. doi: 10.1145/3494672. URL https://doi.org/10.1145/3494672.

Florian Pfisterer. *Democratizing machine learning*. Dissertation, Ludwig-Maximilians-Universität München, October 2022. URL https://edoc.ub.uni-muenchen.de/30947/.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359, December 2011. ISSN 1573-0565. doi: 10.1007/s10994-011-5256-5. URL https://doi.org/10.1007/s10994-011-5256-5.

Kit T. Rodolfa, Erika Salomon, Lauren Haynes, Iván Higuera Mendieta, Jamie Larson, and Rayid Ghani. Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 142–153, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372863. URL https://doi.org/10.1145/3351095.3372863.

Cynthia Rudin, Chudi Zhong, Lesia Semenova, Margo Seltzer, Ronald Parr, Jiachang Liu, Srikar Katta, Jon Donnelly, Harry Chen, and Zachery Boner. Position: Amazing things come from having many good models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 42783–42795, Vienna, Austria, July 2024. JMLR.org.

Abhraneel Sarma, Alex Kale, Michael Jongho Moon, Nathan Taback, Fanny Chevalier, Jessica Hullman, and Matthew Kay. Multiverse: Multiplexing Alternative Data Analyses in R Notebooks. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15, Hamburg Germany, April 2023. ACM. ISBN 978-1-4503-9421-5. doi: 10.1145/3544548.3580726. URL https://dl.acm.org/doi/10.1145/3544548.3580726.

Malte Schierholz. Eine Hilfsklassifikation mit Tätigkeitsbeschreibungen für Zwecke der Berufskodierung. *AStA Wirtschafts- und Sozialstatistisches Archiv*, 12(3-4):285–298, December 2018. ISSN 1863-8155, 1863-8163. doi: 10.1007/s11943-018-0231-2. URL http://link.springer.com/10.1007/s11943-018-0231-2.

Malte Schierholz, Miriam Gensicke, Nikolai Tschersich, and Frauke Kreuter. Occupation coding during the interview. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(2):379–407, 2018. ISSN 1467-985X. doi: 10.1111/rssa.12297. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12297.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html.

Candice Schumann, Femi Olanubi, Auriel Wright, Ellis Monk, Courtney Heldreth, and Susanna Ricco. Consensus and Subjectivity of Skin Tone Annotation for ML Fairness. *Advances in Neural Information Processing Systems*, 36:30319–30348, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/60d25b3210c92f5ba2002a8e1f1adf1c-Abstract-Datasets_and_Benchmarks.html.

Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the Existence of Simpler Machine Learning Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1827–1858, June 2022. doi: 10.1145/3531146.3533232. URL https://doi.org/10.1145/3531146.3533232.

Harald Semmelrock, Simone Kopeinik, Dieter Theiler, Tony Ross-Hellauer, and Dominik Kowald. Reproducibility in Machine Learning-Driven Research, July 2023. URL http://arxiv.org/abs/2307.10320.

Ali Akbar Septiandri, Marios Constantinides, Mohammad Tahaei, and Daniele Quercia. WEIRD FAccTs: How Western, Educated, Industrialized, Rich, and Democratic is FAccT? In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 160–171, Chicago IL USA, June 2023. ACM. ISBN 979-8-4007-0192-4. doi: 10.1145/3593013.3593985. URL https://dl.acm.org/doi/10.1145/3593013.3593985.

R. Silberzahn, E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, F. Bai, C. Bannard, E. Bonnier, R. Carlsson, F. Cheung, G. Christensen, R. Clay, M. A. Craig, A. Dalla Rosa, L. Dam, M. H. Evans, I. Flores Cervantes, N. Fong, M. Gamez-Djokic, A. Glenz, S. Gordon-McKeon, T. J. Heaton, K. Hederos, M. Heene, A. J. Hofelich Mohr, F. Högden, K. Hui, M. Johannesson, J. Kalodimos, E. Kaszubowski, D. M. Kennedy, R. Lei, T. A. Lindsay,

S. Liverani, C. R. Madan, D. Molden, E. Molleman, R. D. Morey, L. B. Mulder, B. R. Nijstad, N. G. Pope, B. Pope, J. M. Prenoveau, F. Rink, E. Robusto, H. Roderique, A. Sandberg, E. Schlüter, F. D. Schönbrodt, M. F. Sherman, S. A. Sommer, K. Sotak, S. Spain, C. Spörlein, T. Stafford, L. Stefanutti, S. Tauber, J. Ullrich, M. Vianello, E.-J. Wagenmakers, M. Witkowiak, S. Yoon, and B. A. Nosek. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*, 1(3):337–356, September 2018. ISSN 2515-2459. doi: 10.1177/2515245917747646. URL https://doi.org/10.1177/2515245917747646.

Raphael Silberzahn and Eric L. Uhlmann. Crowdsourced research: Many hands make tight work. *Nature*, 526(7572):189–191, October 2015. ISSN 1476-4687. doi: 10.1038/526189a. URL https://www.nature.com/articles/526189a.

Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 22(11):1359–1366, November 2011. ISSN 0956-7976, 1467-9280. doi: 10.1177/0956797611417632. URL http://journals.sagepub.com/doi/10.1177/0956797611417632.

Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. Specification curve analysis. *Nature Human Behaviour*, 4(11):1208–1214, November 2020. ISSN 2397-3374. doi: 10.1038/s41562-020-0912-z. URL https://www.nature.com/articles/s41562-020-0912-z.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html.

Raphael Sonabend, Franz J Király, Andreas Bender, Bernd Bischl, and Michel Lang. mlr3proba: an r package for machine learning in survival analysis. *Bioinformatics*, 37(17):2789–2791, 2021.

Clay Spinuzzi. The methodology of participatory design. *Technical communication*, 52(2):163–174, 2005. URL https://www.ingentaconnect.com/content/stc/tc/2005/00000052/00000002/art00005.

Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society*, 9(2):20539517221115189, July 2022. ISSN 2053-9517. doi: 10.1177/20539517221115189. URL https://doi.org/10.1177/20539517221115189.

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, 11(5):702–712, September 2016. ISSN 1745-6916. doi: 10.1177/1745691616658637. URL https://doi.org/10.1177/1745691616658637.

Edward Telles and Liza Steele. Pigmentocracy in the Americas: How is Educational Attainment Related to Skin Color? *AmericasBarometer Insights: 2012*, 73, 2012. URL https://www.vanderbilt.edu/lapop/insights/IO873en.pdf.

David Thiel and Jeffrey Hancock. Identifying and Eliminating CSAM in Generative ML Training Data and Models, 2023. URL https://purl.stanford.edu/kh752sm9123.

Chris Thornton, Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 847–855, New York, NY, USA, August 2013. Association for Computing Machinery. ISBN 978-1-4503-2174-7. doi: 10.1145/2487575.2487629. URL https://doi.org/10.1145/2487575.2487629.

United Nations. Universal declaration of human rights, 1948. https://www.un.org/en/about-us/universal-declaration-of-human-rights.

United States District Court, District of Minnesota. Lokken v UnitedHealth Group Inc. case: 0:23-cv-03514, 2023.

Aleksandra Urman, Mykola Makhortykh, and Aniko Hannak. WEIRD Audits? Research Trends, Linguistic and Geographical Disparities in the Algorithm Audits of Online Platforms - A Systematic Literature Review. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pages 375–390, New York, NY, USA, June 2025. Association for Computing Machinery. ISBN 979-8-4007-1482-5. doi: 10.1145/3715275.3732026. URL https://dl.acm.org/doi/10.1145/3715275.3732026.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, pages 1–7, Gothenburg Sweden, May 2018. ACM. ISBN 978-1-4503-5746-3. doi: 10.1145/3194770.3194776. URL https://dl.acm.org/doi/10.1145/3194770.3194776.

Eric-Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, and Han L. J. van der Maas. Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, 100(3):426–432, 2011. ISSN 1939-1315. doi: 10.1037/a0022790.

Jamelle Watson-Daniels, Solon Barocas, Jake M. Hofman, and Alexandra Chouldechova. 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, pages 297–311, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 979-8-4007-0192-4. doi: 10.1145/3593013.3593998. URL https://doi.org/10.1145/3593013.3593998.

Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and Improving Fairness of AI Systems. *Journal of Machine Learning Research*, 24(257):1–8, 2023. ISSN 1533-7928. URL http://jmlr.org/papers/v24/23-0389.html.

Hilde Weerts, Florian Pfisterer, Matthias Feurer, Katharina Eggensperger, Edward Bergman, Noor Awad, Joaquin Vanschoren, Mykola Pechenizkiy, Bernd Bischl, and Frank Hutter. Can Fairness be Automated? Guidelines and Opportunities for Fairness-aware AutoML. *Journal of Artificial Intelligence Research*, 79:639–677, February 2024. ISSN 1076-9757. doi: 10.1613/jair.1.14747. URL https://www.jair.org/index.php/jair/article/view/14747.

Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. Data collection and quality challenges in deep learning: A data-centric AI perspective. *The VLDB Journal*, 32(4):791–813,

July 2023. ISSN 0949-877X. doi: 10.1007/s00778-022-00775-9. URL https://doi.org/10.1007/s00778-022-00775-9.

Hadley Wickham. Tidy Data. *Journal of Statistical Software*, 59(1):1–23, September 2014. ISSN 1548-7660. doi: 10.18637/jss.v059.i10. URL https://www.jstatsoft.org/index.php/jss/article/view/v059i10.

Hadley Wickham. *R for Data Science*. O'Reilly Media, Incorporated, Sebastopol, 2nd edition, 2023. ISBN 978-1-4920-9740-2 978-1-4920-9737-2. URL https://r4ds.hadley.nz/.

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. Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43):1686, November 2019. ISSN 2475-9066. doi: 10.21105/joss.01686. URL https://joss.theoj.org/papers/10.21105/joss.01686.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, January 1992. ISSN 0893-6080. doi: 10.1016/S0893-6080(05)80023-1. URL https://www.sciencedirect.com/science/article/pii/S0893608005800231.

Doris Xin, Hui Miao, Aditya Parameswaran, and Neoklis Polyzotis. Production Machine Learning Pipelines: Empirical Analysis and Optimization Opportunities. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, pages 2639–2652, New York, NY, USA, June 2021. Association for Computing Machinery. ISBN 978-1-4503-8343-1. doi: 10.1145/3448016.3457566. URL https://dl.acm.org/doi/10.1145/3448016.3457566.

Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. Exploring the Whole Rashomon Set of Sparse Decision Trees. *Advances in Neural Information Processing Systems*, 35:14071–14084, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/5afaa8b4dd18eb1eed055d2d821b58ae-Abstract-Conference.html.

Bin Yu and Rebecca L. Barter. *Veridical Data Science (Online Version)*. MIT Press, 2024. URL https://vdsbook.com/. Online manuscript.

Bin Yu and Karl Kumbier. Veridical data science. *Proceedings of the National Academy of Sciences*, 117(8):3920–3929, February 2020. doi: 10.1073/pnas.1901326117. URL https://doi.org/10.1073/pnas.1901326117.

Matei Zaharia, Andrew Chen, Aaron Davidson, Ali Ghodsi, Sue Ann Hong, Andy Konwinski, Siddharth Murching, Tomas Nykodym, Paul Ogilvie, Mani Parkhe, et al. Accelerating the machine learning lifecycle with mlflow. *IEEE Data Eng. Bull.*, 41(4):39–45, 2018. URL http://sites.computer.org/debull/A18dec/p39.pdf.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/zemel13.html.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 335–340, New York, NY, USA, December 2018. Association for Computing Machinery. ISBN 978-1-4503-6012-8. doi: 10.1145/3278721.3278779. URL https://dl.acm.org/doi/10.1145/3278721.3278779.

Dora Zhao, Morgan K. Scheuerman, Pooja Chitre, Jerone T. Andrews, Georgia Panagiotidou, Shawn Walker, Kathleen H. Pine, and Alice Xiang. A Taxonomy of Challenges to Curating Fair Datasets. *Advances in Neural Information Processing Systems*, 37:97826–97858, December 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b142e78db191e19b17e60c1425a28b52-Abstract-Datasets_and_Benchmarks_Track.html.

Marc-André Zöller and Marco F. Huber. Benchmark and Survey of Automated Machine Learning Frameworks. *Journal of Artificial Intelligence Research*, 70:409–472, January 2021. ISSN 1076-9757. doi: 10.1613/jair.1.11854. URL https://www.jair.org/index.php/jair/article/view/11854.

Alain F. Zuur, Elena N. Ieno, Neil J. Walker, Anatoly A. Saveliev, and Graham M. Smith. Zero-Truncated and Zero-Inflated Models for Count Data. In *Mixed Effects Models and Extensions in Ecology with R*, pages 261–293. Springer, New York, NY, 2009. ISBN 978-0-387-87458-6. doi: 10.1007/978-0-387-87458-6_11. URL https://doi.org/10.1007/978-0-387-87458-6_11.

Eidesstattliche Versicherung (Affidavit)

(Siehe Promotionsordnung vom 12. Juli 2011, § 8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 21.01.2026 Jan Simson