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.

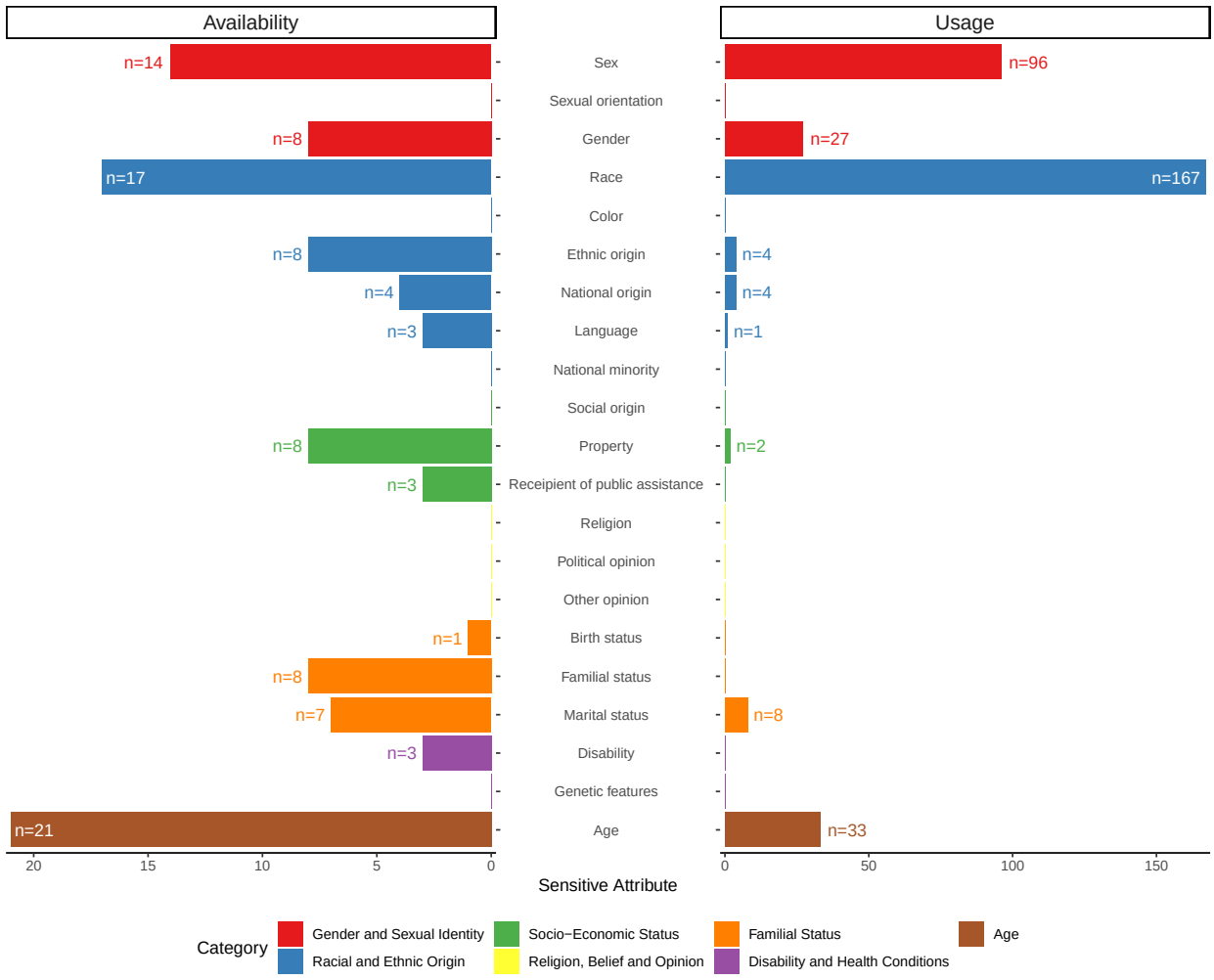

Figure 1: There is a large discrepancy between the list of attributes considered protected under international legislation and their availability or usage in datasets. Bar chart displaying the availability (left) and usage (right) of protected attributes in the literature for all categories of protected 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

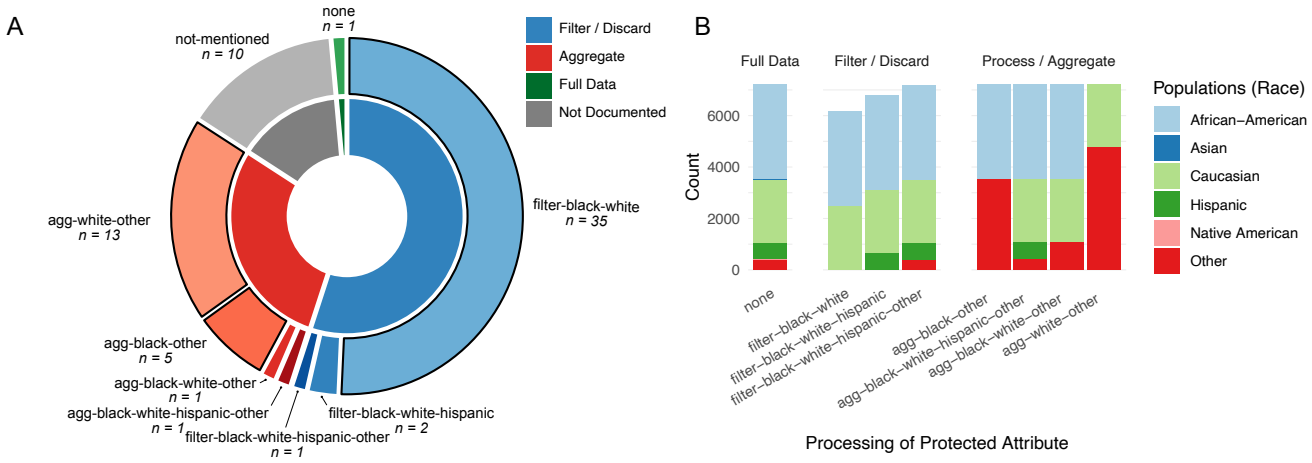

Figure 2: Data from smaller populations is almost *always* either discarded or aggregated within the annotated literature. (A) Prevalence of processing strategies for the COMPAS dataset within the annotated literature and (B) resulting base rates of the protected attribute from these different processing 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.

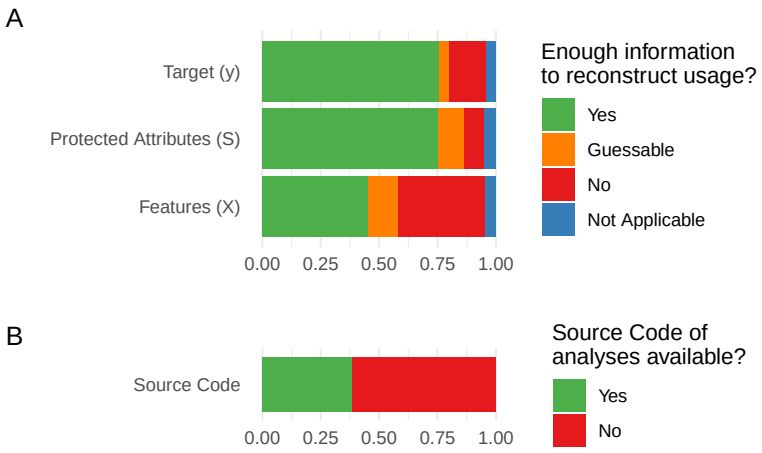

Figure 3: A large section of the annotated literature lacks sufficient information to reproduce analyses. Bar diagrams showing whether publications in the annotated literature contain (A) sufficient information to reconstruct usage of the predicted target variables y, the protected features S and the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

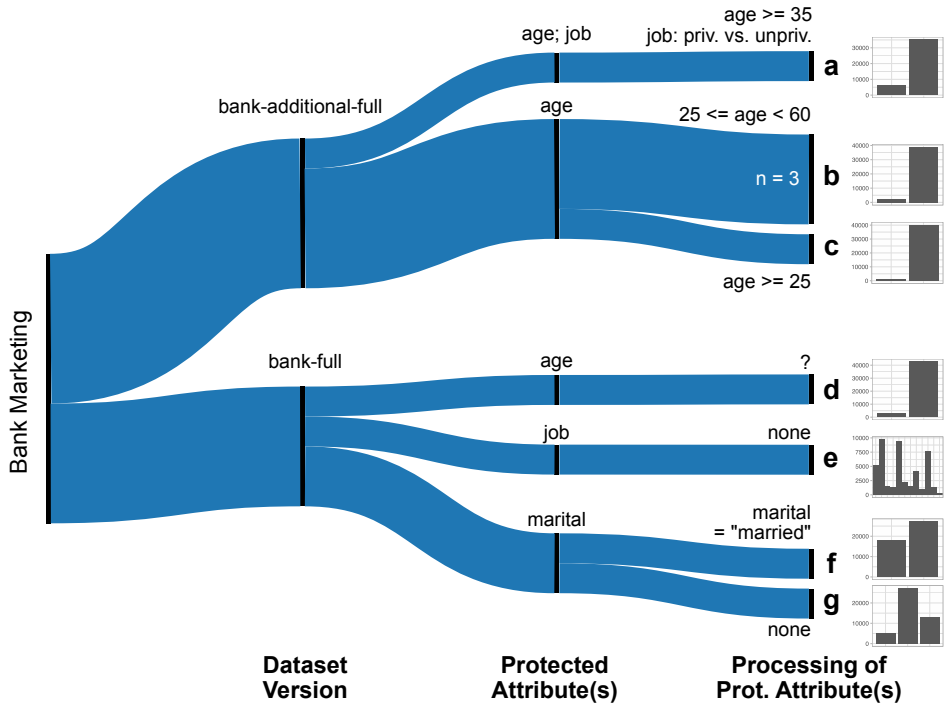

Figure 4: The "same" dataset is used in many different ways within the literature. Sankey diagram illustrating the usage of the Bank dataset within the annotated literature. Each split corresponds to a choice where differences were observed in the literature. Each unique combination of choices or scenario 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.

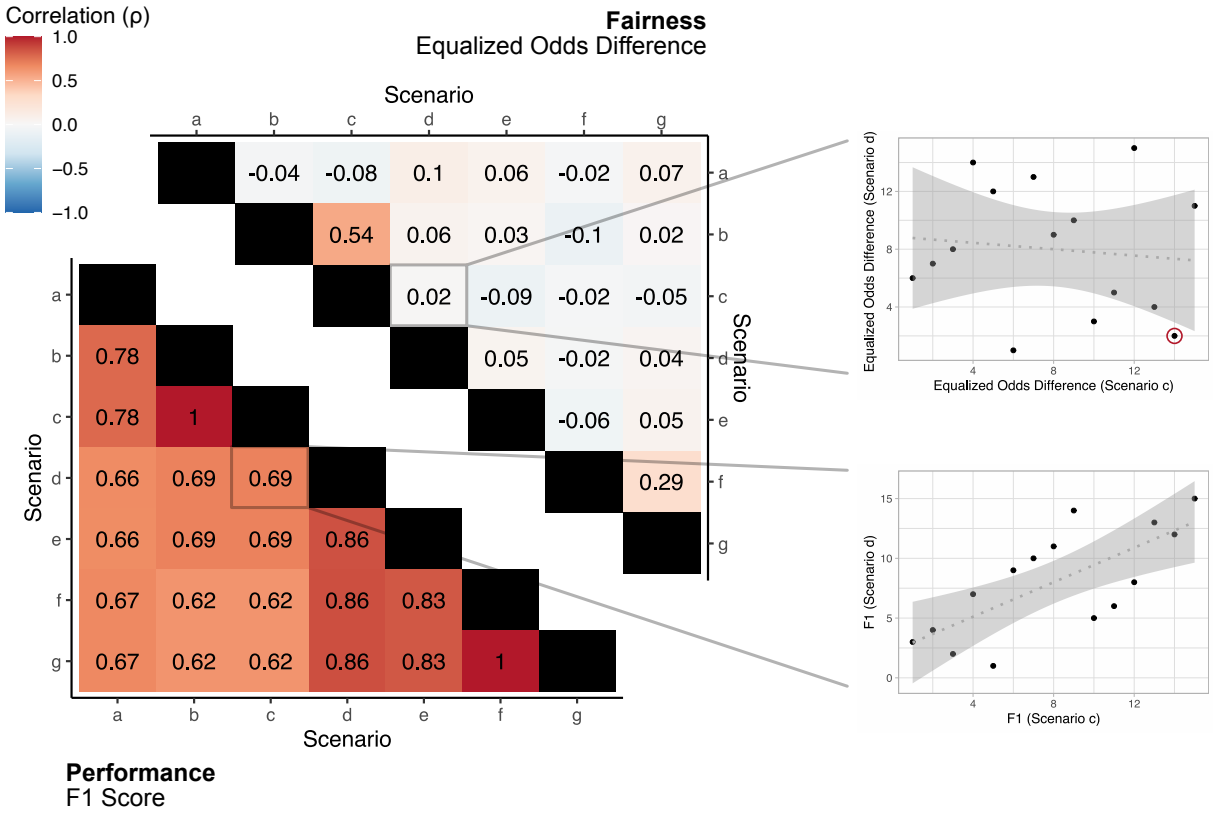

Figure 5: While a practitioner would choose roughly similar models based on performance across the different scenarios, they would choose very different ones based on fairness. Spearman's ρ correlations of model ranks on a measure of fairness (Equalized Odds Difference) and performance (F1 score) between 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 $\bar{\rho}_{F1} = 0.747$. 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.

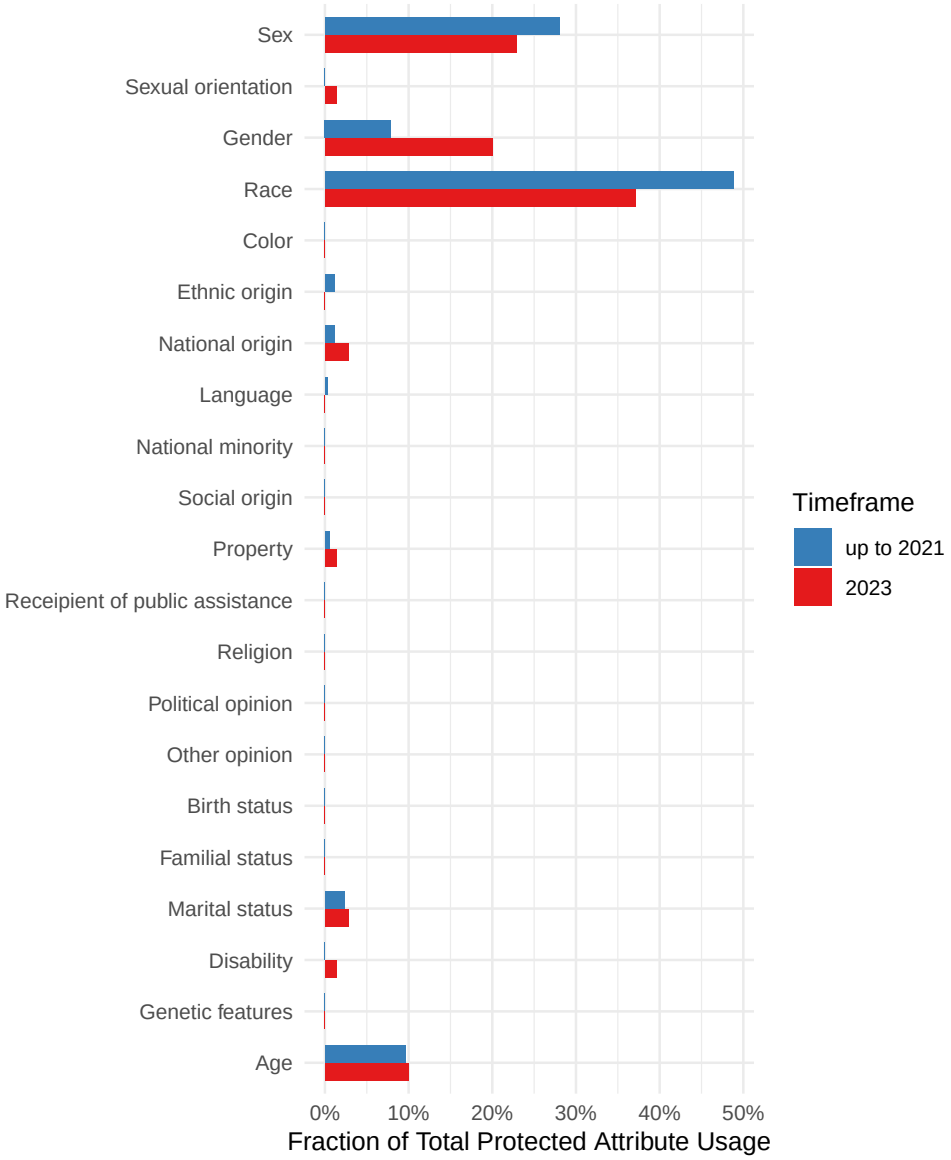

Figure 6: The usage of protected attributes remained similar in 2023. Relative usage of protected attributes in the annotated literature up to 2021 and within the subset of literature we examined in 2023. Usage within the annotated literature corresponds to the right half of Figure 1.

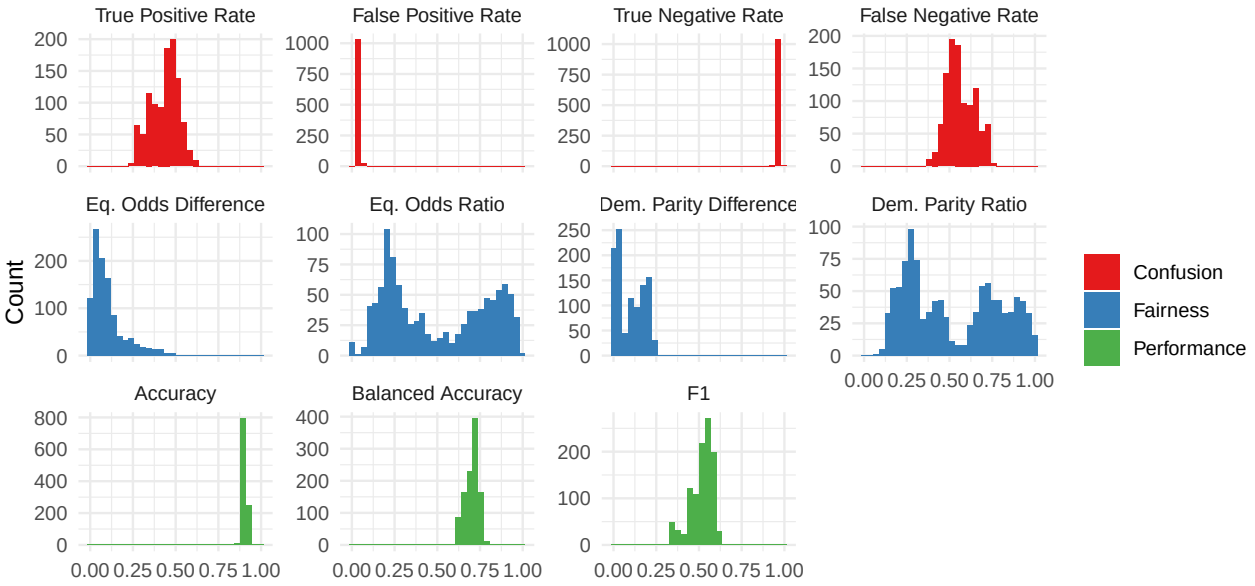

Figure 7: There is a large degree of overall variation, especially on fairness metrics. Histograms displaying the overall variation on different metrics within and across different scenarios and repetitions of the analysis.

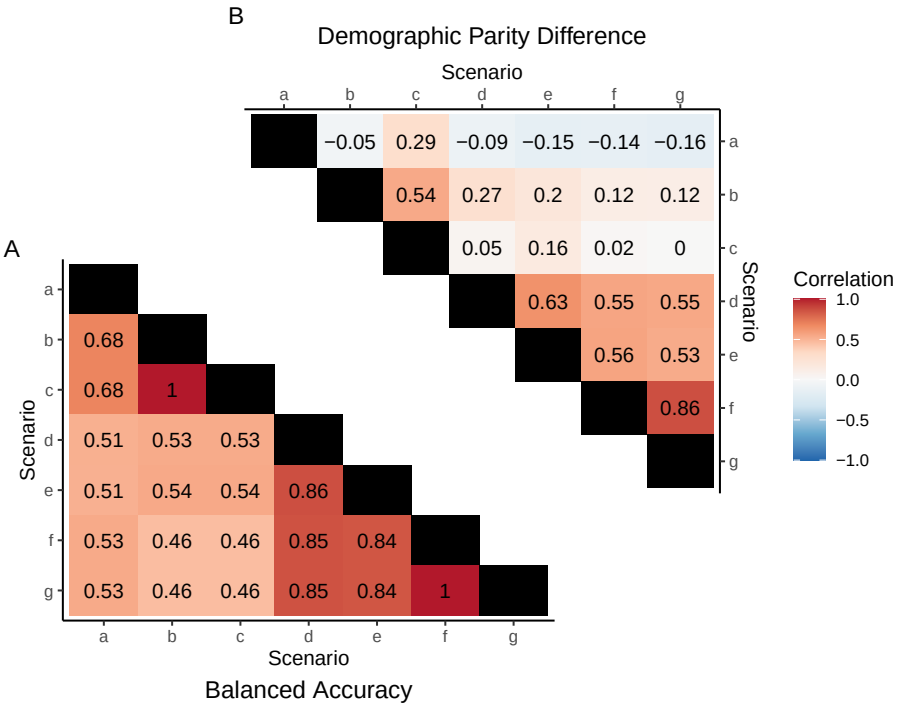

Figure 8: Spearman's ρ correlations of model ranks on (A) Balanced Accuracy and (B) Demographic Parity Difference between different scenarios. Letters correspond to the scenarios described in Figure 4.

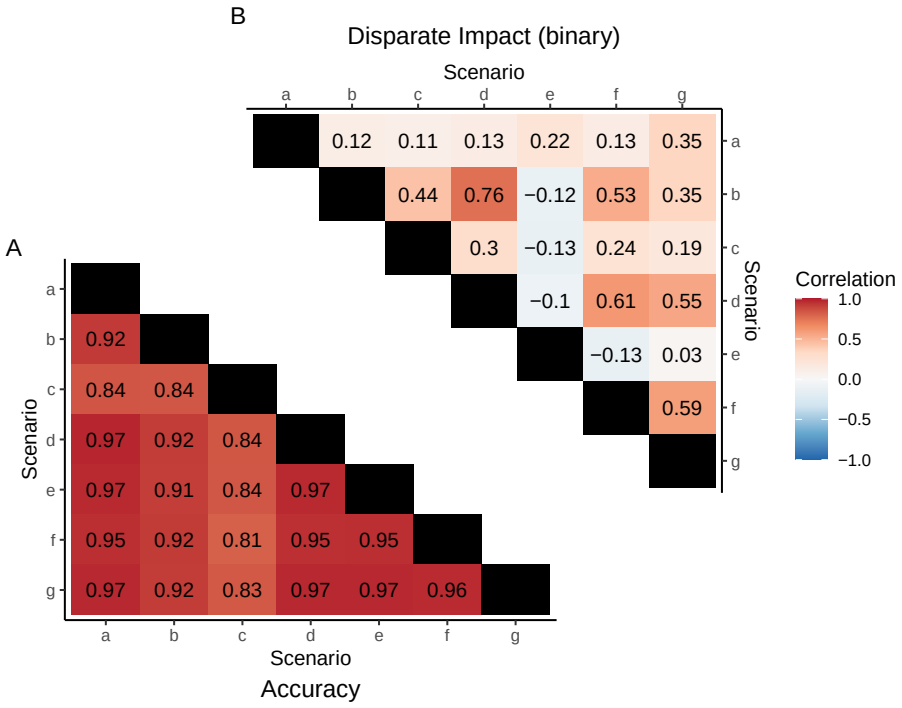

Figure 9: Spearman's ρ correlations of model ranks on (A) Raw Accuracy and (B) Disparate Impact (binary) between different scenarios when reproducing our analysis from Section 5 using an existing selection of fairness-aware algorithms and methodology [46]. Letters correspond to the scenarios described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