Data Mining and Knowledge Discovery
https://doi.org/10.1007/s10618-022-00854-z

Algorithmic fairness datasets: the story so far

Alessandro Fabris[1] · Stefano Messina[1] · Gianmaria Silvello[1] · Gian Antonio Susto[1]

Received: 27 August 2021 / Accepted: 5 July 2022

Abstract

Data-driven algorithms are studied and deployed in diverse domains to support critical decisions, directly impacting people's well-being. As a result, a growing community of researchers has been investigating the equity of existing algorithms and proposing novel ones, advancing the understanding of risks and opportunities of automated decision-making for historically disadvantaged populations. Progress in fair machine learning and equitable algorithm design hinges on data, which can be appropriately used only if adequately documented. Unfortunately, the algorithmic fairness community, as a whole, suffers from a collective data documentation debt caused by a lack of information on specific resources (*opacity*) and scatteredness of available information (*sparsity*). In this work, we target this data documentation debt by surveying over two hundred datasets employed in algorithmic fairness research, and producing standardized and searchable documentation for each of them. Moreover we rigorously identify the three most popular fairness datasets, namely Adult, COMPAS, and German Credit, for which we compile in-depth documentation. This unifying documentation effort supports multiple contributions. Firstly, we summarize the merits and limitations of Adult, COMPAS, and German Credit, adding to and unifying recent scholarship, calling into question their suitability as general-purpose fairness benchmarks. Secondly, we document hundreds of available alternatives, annotating their domain and supported fairness tasks, along with additional properties of interest for

Responsible editor: Toon Calders.

This is the extended version of (Fabris et al. 2022).

✉ Alessandro Fabris
fabrisal@dei.unipd.it

Gianmaria Silvello
silvello@dei.unipd.it

Gian Antonio Susto
gianantonio.susto@unipd.it

[1] Dipartimento di Ingegneria dell'Informazione, Università di Padova, Via Giovanni Gradenigo 6B, 35131 Padua, Italy

Published online: 17 September 2022

fairness practitioners and researchers, including their format, cardinality, and the sensitive attributes they encode. We summarize this information, zooming in on the tasks, domains, and roles of these resources. Finally, we analyze these datasets from the perspective of five important data curation topics: anonymization, consent, inclusivity, labeling of sensitive attributes, and transparency. We discuss different approaches and levels of attention to these topics, making them tangible, and distill them into a set of best practices for the curation of novel resources.

Keywords Algorithmic fairness · Datasets · Documentation debt

1 Introduction

Following the widespread study and application of data-driven algorithms in contexts that are central to people's well-being, a large community of researchers has coalesced around the growing field of algorithmic fairness, investigating algorithms through the lens of justice, equity, bias, power, and harms. A line of work gaining traction in the field, intersecting with critical data studies, human–computer interaction, and computer-supported cooperative work, focuses on data transparency and standardized documentation processes to describe key characteristics of datasets (Gebru et al. 2018; Holland et al. 2018; Bender and Friedman 2018; Geiger et al. 2020; Jo and Gebru 2020; Miceli et al. 2021). Most prominently, Gebru et al. (2018) and Holland et al. (2018) proposed two complementary documentation frameworks, called *Datasheets for Datasets* and *Dataset Nutrition Labels*, to improve data curation practices and favour more informed data selection and utilization for dataset users. Overall, this line of work has contributed to an unprecedented attention to dataset documentation in Machine Learning (ML), including a novel track focused on datasets at the Conference on Neural Information Processing Systems (NeurIPS), an initiative to support dataset tracking in repositories for scholarly articles,[1] and dedicated works producing retrospective documentation for existing datasets (Bandy and Vincent 2021; Garbin et al. 2021), auditing their properties (Prabhu and Birhane 2020) and tracing their usage (Peng et al. 2021).

In recent work, Bender et al. (2021) propose the notion of *documentation debt*, in relation to training sets that are undocumented and too large to document retrospectively. We extend this definition to the collection of datasets employed in a given field of research. We see two components at work contributing to the documentation debt of a research community. On one hand, *opacity* is the result of poor documentation affecting single datasets, contributing to misunderstandings and misuse of specific resources. On the other hand, when relevant information exists but does not reach interested parties, there is a problem of documentation *sparsity*. One example that is particularly relevant for the algorithmic fairness community is represented by the German Credit dataset (UCI Machine Learning Repository 1994), a popular resource in this field. Many works of algorithmic fairness, including recent ones, carry out experiments on this dataset using sex as a protected attribute (He et al. 2020b; Yang

[1] https://medium.com/paperswithcode/datasets-on-arxiv-1a5a8f7bd104.

et al. 2020a; Baharlouei et al. 2020; Lohaus et al. 2020; Martinez et al. 2020; Wang et al. 2021; Perrone et al. 2021; Sharma et al. 2021), while existing yet overlooked documentation shows that this feature cannot be reliably retrieved (Grömping 2019). Moreover, the mere fact that a dataset exists and is relevant to a given task or a given domain may be unknown. The BUPT Faces datasets, for instance, were presented as the second existing resource for face analysis with race annotations (Wang and Deng 2020). However several resources were already available at the time, including Labeled Faces in the Wild (Han and Jain 2014), UTK Face (Zhang et al. 2017b), Racial Faces in the Wild (Wang et al. 2019e), and Diversity in Faces (Merler et al. 2019).[2]

To tackle the documentation debt of the algorithmic fairness community, we survey the datasets used in over 500 articles on fair ML and equitable algorithmic design, presented at seven major conferences, considering each edition in the period 2014–2021, and more than twenty domain-specific workshops in the same period. We find over 200 datasets employed in studies of algorithmic fairness, for which we produce compact and standardized documentation, called *data briefs*. Data briefs are intended as a lightweight format to document fundamental properties of data artifacts used in algorithmic fairness, including their purpose, their features, with particular attention to sensitive ones, the underlying labeling procedure, and the envisioned ML task, if any. To favor domain-based and task-based search from dataset users, data briefs also indicate the domain of the processes that produced the data (e.g., radiology) and list the fairness tasks studied on a given dataset (e.g. fair ranking). For this endeavour, we have contacted creators and knowledgeable practitioners identified as primary points of contact for the datasets. We received feedback (incorporated into the final version of the data briefs) from 79 curators and practitioners, whose contribution is acknowledged at the end of this article. Moreover, we identify and carefully analyze the three datasets most often utilized in the surveyed articles (Adult, COMPAS, and German Credit), retrospectively producing a datasheet and a nutrition label for each of them. From these documentation efforts, we extract a summary of the merits and limitations of popular algorithmic fairness benchmarks, a categorization of domains and fairness tasks for existing datasets, and a set of best practices for the curation of novel resources.

Overall, we make the following contributions.

- **Unified analysis of popular fairness benchmarks**. We produce *datasheets* and *nutrition labels* for Adult, COMPAS, and German Credit, from which we extract a summary of their merits and limitations. We add to and unify recent scholarship on these datasets, calling into question their suitability as general-purpose fairness benchmarks due to contrived prediction tasks, noisy data, severe coding mistakes, and age.
- **Survey of existing alternatives**. We compile standardized and compact documentation for over two hundred resources used in fair ML research, annotating their domain, the tasks they support, and the roles they play in works of algorithmic fairness. By assembling sparse information on hundreds of datasets into a single document, we aim to support multiple goals by researchers and practitioners,

[2] Hereafter, for brevity, we only report dataset names. The relevant references and additional information can be found in Appendix A.

including domain-oriented and task-oriented search by dataset users. Contextually, we provide a novel categorization of tasks and domains investigated in algorithmic fairness research (summarized in Tables 2 and 3).

- **Best practices for the curation of novel resources**. We analyze different approaches to anonymization, consent, inclusivity, labeling, and transparency across these datasets. By comparing existing approaches and discussing their advantages, we make the underlying concerns visible and practical, and extract best practices to inform the curation of new datasets and post-hoc remedies to existing ones.

The rest of this work is organized as follows. Section 2 introduces related works. Section 3 presents the methodology and inclusion criteria of this survey. Section 4 analyzes the perks and limitations of the most popular datasets, namely Adult (Sect. 4.1), COMPAS (Sect. 4.2), and German Credit (Sect. 4.3), and provides an overall summary of their merits and limitations as fairness benchmarks (Sect. 4.4). Section 5 discusses alternative fairness resources from the perspective of the underlying domains (Sect. 5.1), the fair ML tasks they support (Sect. 5.2), and the roles they play (Sect. 5.3). Section 6 presents important topics in data curation, discussing existing approaches and best practices to avoid re-identification (Sect. 6.1), elicit informed consent (Sect. 6.2), consider inclusivity (Sect. 6.3), collect sensitive attributes (Sect. 6.4), and document datasets (Sect. 6.5). Section 7 summarizes the broader benefits of our documentation effort and envisioned uses for the research community. Finally, Sect. 8 contains concluding remarks and recommendations. Interested readers may find the data briefs in Appendix A, followed by the detailed documentation produced for Adult (B), COMPAS (C), and German Credit (D).

2 Related work

2.1 Algorithmic fairness surveys

Multiple surveys about algorithmic fairness have been published in the literature (Mehrabi et al. 2021; Caton and Haas 2020; Pessach and Shmueli 2020). These works typically focus on describing and classifying important measures of algorithmic fairness and methods to enhance it. Some articles also discuss sources of bias (Mehrabi et al. 2021), software packages and projects which address fairness in ML (Caton and Haas 2020), or describe selected sub-fields of algorithmic fairness (Pessach and Shmueli 2020). Datasets are typically not emphasized in these works, which is also true of domain-specific surveys on algorithmic fairness, focused e.g. on ranking (Pitoura et al. 2021), Natural Language Processing (NLP) (Sun et al. 2019) and computational medicine (Sun et al. 2019). As an exception, Pessach and Shmueli (2020) and Zehlike et al. (2021) list and briefly describe 12 popular algorithmic fairness datasets, and 19 datasets employed in fair ranking research, respectivey.

2.2 Data studies

The work most closely related (and concurrently carried out) to ours is Le Quy et al. (2022). The authors perform a detailed analysis of 15 tabular datasets used in works of algorithmic fairness, listing important metadata (e.g. domain, protected attributes, collection period and location), and carrying out an exploratory analysis of the probabilistic relationship between features. Our work complements it by placing more emphasis on (1) a rigorous methodology for the inclusion of resources, (2) a wider selection of (over 200) datasets spanning different data types, including text, image, timeseries, and tabular data, (3) a fine-grained evaluation of domains and tasks associated with each dataset, and (4) the analysis and distillation of best practices for data curation. It will be interesting to see how different goals of the research community, such as selection of appropriate resources for experimentation and data studies, can benefit from the breadth and depth of both works.

Other works analyzing multiple datasets along specific lines have been published in recent years. Crawford and Paglen (2021) focus on resources commonly used as training sets in computer vision, with attention to associated labels and underlying taxonomies. Fabbrizzi et al. (2021) also consider computer vision datasets, describing types of bias affecting them, along with methods for discovering and measuring bias. Peng et al. (2021) analyze ethical concerns in three popular face and person recognition datasets, stemming from derivative datasets and models, lack of clarity of licenses, and dataset management practices. Geiger et al. (2020) evaluate transparency in the documentation of labeling practices employed in over 100 datasets about Twitter. Leonelli and Tempini (2020) study practices of collection, cleaning, visualization, sharing, and analysis across a variety of research domains. Romei and Ruggieri (2014) survey techniques and data for discrimination analysis, focused on measuring, rather than enforcing, equity in human processes.

A different, yet related, family of articles provides deeper analyses of single datasets. Prabhu and Birhane (2020) focus on Imagenet (ILSVRC 2012) which they analyze along the lines of consent, problematic content, and individual re-identification. Kizhner et al. (2020) study issues of representation in the Google Arts and Culture project across countries, cities and institutions. Some works provide datasheets for a given resource, such as CheXpert (Garbin et al. 2021) and the BookCorpus (Bandy and Vincent 2021). Among popular fairness datasets, COMPAS has drawn scrutiny from multiple works, analysing its numerical idiosyncrasies (Barenstein 2019) and sources of bias (Bao et al. 2021). Ding et al. (2021) study numerical idiosyncrasies in the Adult dataset, and propose a novel version, for which they provide a datasheet. Grömping (2019) discuss issues resulting from coding mistakes in German Credit.

Our work combines the breadth of multi-dataset and the depth of single-dataset studies. On one hand, we survey numerous resources used in works of algorithmic fairness, analyzing them across multiple dimensions. On the other hand, we identify the most popular resources, compiling their *datasheet* and *nutrition label*, and summarize their perks and limitations. Moreover, by making our data briefs available, we hope to contribute a useful tool to the research community, favouring further data studies and analyses, as outlined in Sect. 7.

2.3 Documentation frameworks

Several data documentation frameworks have been proposed in the literature; three popular ones are described below. *Datasheets for Datasets* (Gebru et al. 2018) are a general-purpose qualitative framework with over fifty questions covering key aspects of datasets, such as motivation, composition, collection, preprocessing, uses, distribution, and maintenance. Another qualitative framework is represented by *Data statements* (Bender and Friedman 2018), which is tailored for NLP, requiring domain-specific information on language variety and speaker demographics. *Dataset Nutrition Labels* (Holland et al. 2018) describe a complementary, quantitative framework, focused on numerical aspects such as the marginal and joint distribution of variables.

Popular datasets require close scrutiny; for this reason we adopt these frameworks, producing three datasheets and nutrition labels for Adult, German Credit, and COMPAS. This approach, however, does not scale to a wider documentation effort with limited resources. For this reason, we propose and produce *data briefs*, a lightweight documentation format designed for algorithmic fairness datasets. Data briefs, described in Appendix A, include fields specific to fair ML, such sensitive attributes and tasks for which the dataset has been used in the algorithmic fairness literature.

3 Methodology

In this work, we consider (1) every article published in the proceedings of domain-specific 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) every article published in proceedings of well-known machine learning and data mining 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); (3) every article available from Past Network Events and Older Workshops and Events of the FAccT network.[3] We consider the period from 2014, the year of the first workshop on Fairness, Accountability, and Transparency in Machine Learning, to June 2021, thus including works presented at FAccT, ICLR, AIES, and CVPR in 2021.[4]

To target works of algorithmic fairness, we select a subsample of these articles whose titles contain either of the following strings, where the star symbol represents the wildcard character: `*fair*` (targeting e.g. fairness, unfair), `*bias*` (biased, debiasing), `discriminat*` (discrimination, discriminatory), `*equal*` (equality, unequal), `*equit*` (equity, equitable), disparate (disparate impact), `*parit*` (par-

[3] https://facctconference.org/network/.

[4] We are working on an update covering more recent work, including articles presented at the ACM conference on Equity and Access in Algorithms, Mechanisms, and Optimization.

ity, disparities). These selection criteria are centered around equity-based notions of fairness, typically operationalized by measuring disparity in some algorithmic property across individuals or groups of individuals. Through manual inspection by two authors, we discard articles where these keywords are used with a different meaning. Discarded works, for instance, include articles on handling pose distribution bias (Zhao et al. 2021), compensating selection bias to improve accuracy without attention to sensitive attributes (Kato et al. 2019), enhancing desirable discriminating properties of models (Chen et al. 2018a), or generally focused on model performance (Li et al. 2018; Zhong et al. 2019). This leaves us with 558 articles.

From the articles that pass this initial screening, we select datasets treated as important data artifacts, either being used to train/test an algorithm or undergoing a data audit, i.e., an in-depth analysis of different properties. We produce a data brief for these datasets by (1) reading the information provided in the surveyed articles, (2) consulting the provided references, and (3) reviewing scholarly articles or official websites found by querying popular search engines with the dataset name. We discard the following:

- Word Embeddings (WEs). We only consider the corpora they are trained on, provided WEs are trained as part of a given work and not taken off the shelf;
- toy datasets, i.e., simulations with no connection to real-world processes, unless they are used in more than one article, which we take as a sign of importance in the field;
- auxiliary resources that are only used as a minor source of ancillary information, such as the percentage of US residents in each state;
- datasets for which the available information is insufficient. This happens very seldom when points (1), (2), and (3) outlined above result in little to no information about the curators, purposes, features, and format of a dataset. For popular datasets, this is never the case.

For each of the 226 datasets satisfying the above criteria, we produce a data brief, available in Appendix A with a description of the underlying coding procedure. From this effort, we rigorously identify the three most popular resources, whose perks and limitations are summarized in the next section.

4 Most popular datasets

Figure 1 depicts the number of articles using each dataset, showing that dataset utilization in surveyed scholarly works follows a long tail distribution. Over 100 datasets are only used once, also because some of these resources are not publicly available. Complementing this long tail is a short head of nine resources used in ten or more articles. These datasets are Adult (118 usages), COMPAS (81), German Credit (35), Communities and Crime (26), Bank Marketing (19), Law School (17), CelebA (16), MovieLens (14), and Credit Card Default (11). The tenth most used resource is the toy dataset from Zafar et al. (2017c), used in 7 articles. In this section, we summarize positive and negative aspects of the three most popular datasets, namely Adult, COM-

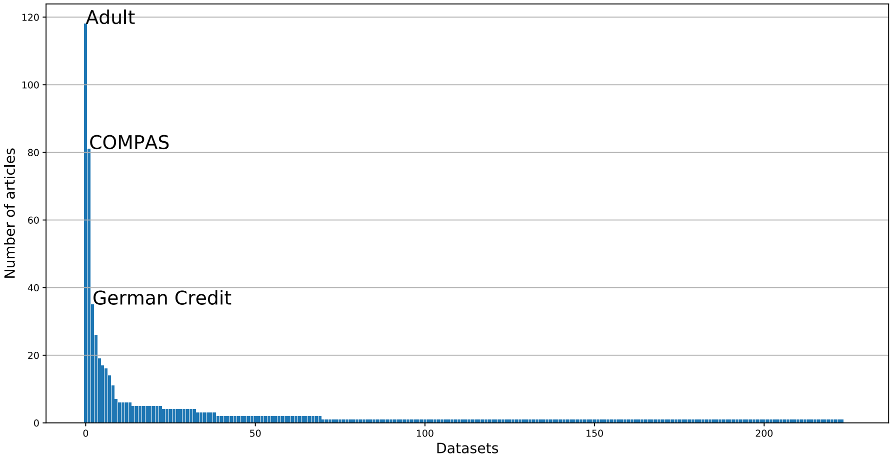

Fig. 1 Utilization of datasets in fairness research follows a long tail distribution

PAS, and German Credit, informed by extensive documentation in Appendices B, C, and D.

4.1 Adult

The Adult dataset was created as a resource to benchmark the performance of machine learning algorithms on socially relevant data. Each instance is a person who responded to the March 1994 US Current Population Survey, represented along demographic and socio-economic dimensions, with features describing their profession, education, age, sex, race, personal, and financial condition. The dataset was extracted from the census database, preprocessed, and donated to UCI Machine Learning Repository in 1996 by Ronny Kohavi and Barry Becker. A binary variable encoding whether respondents' income is above $50,000 was chosen as the target of the prediction task associated with this resource.

Adult inherits some positive sides from the best practices employed by the US Census Bureau. Although later filtered somewhat arbitrarily, the original sample was designed to be representative of the US population. Trained and compensated interviewers collected the data. Attributes in the dataset are self-reported and provided by consensual respondents. Finally, the original data from the US Census Bureau is well documented, and its variables can be mapped to Adult by consulting the original documentation (US Dept. of Commerce Bureau of the Census 1995), except for a variable denominated `fnlwgt`, whose precise meaning is unclear.

A negative aspect of this dataset is the contrived prediction task associated with it. Income prediction from socio-economic factors is a task whose social utility appears rather limited. Even discounting this aspect, the arbitrary $50,000 threshold for the binary prediction task is high, and model properties such as accuracy and fairness are very sensitive to it (Ding et al. 2021). Furthermore, there are several sources of noise affecting the data. Roughly 7% of the data points have missing values, plausibly due to issues with data recording and coding, or respondents' inability to recall information. Moreover, the tendency in household surveys for respondents to under-report their income is a common concern of the Census Bureau (Moore et al. 2000). Another source of noise is top-coding of the variable "capital-gain" (saturation to $99,999) to avoid the re-identification of certain individuals (US Dept. of Commerce Bureau of the Census 1995). Finally, the dataset is rather old; sensitive attribute "race" contains the outdated "Asian Pacific Islander" class. It is worth noting that a set of similar resources was recently made available, allowing more current socio-economic studies of the US population (Ding et al. 2021).

4.2 COMPAS

This dataset was created for an external audit of racial biases in the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) risk assessment tool developed by Northpointe (now Equivant), which estimates the likelihood of a defendant becoming a recidivist. Instances represent defendants scored by COMPAS in Broward County, Florida, between 2013–2014, reporting their demographics, criminal record, custody and COMPAS scores. Defendants' public criminal records were obtained from the Broward County Clerk's Office website matching them based on date of birth, first and last names. The dataset was augmented with jail records and COMPAS scores provided by the Broward County Sheriff's Office. Finally, public incarceration records were downloaded from the Florida Department of Corrections website. Instances are associated with two target variables (is_recid and is_violent_recid), indicating whether defendants were booked in jail for a criminal offense (potentially violent) that occurred after their COMPAS screening but within two years.

On the upside, this dataset is recent and captures some relevant aspects of the COMPAS risk assessment tool and the criminal justice system in Broward County. On the downside, it was compiled from disparate sources, hence clerical errors and mismatches are present (Larson et al. 2016). Moreover, in its official release (ProPublica 2016), the COMPAS dataset features redundant variables and data leakage due to spuriously time-dependent recidivism rates (Barenstein 2019). For these reasons, researchers must perform further preprocessing in addition to the standard one by ProPublica. More subjective choices are required of researchers interested in counterfactual evaluation of risk-assessment tools, due to the absence of a clear indication of whether defendants were detained or released pre-trial (Mishler et al. 2021). The lack of a standard preprocessing protocol beyond the one by ProPublica (ProPublica 2016), which is insufficient to handle these factors, may cause issues of reproducibility and difficulty in comparing methods. Moreover, according to Northpointe's response to the ProPublica's study, several risk factors considered by the COMPAS algorithm are absent from the dataset (Dieterich et al. 2016). As an additional concern, race categories lack Native Hawaiian or Other Pacific Islander, while Hispanic is redefined as race instead of ethnicity (Bao et al. 2021). Finally, defendants' personal information (e.g. race and criminal history) is available in conjunction with obvious identifiers, making re-identification of defendants trivial.

Table 1 Limitations of popular algorithmic fairness datasets

	Adult	COMPAS	German credit
Age	Old (1994)	Recent (2013–2016)	Very old (1973–1975)
Prediction task	Contrived (income > 50K$)	Realistic (recidivism)	Realistic (creditworthiness)
Sensitive attributes	Outdated racial categories	Outdated racial categories	Sex cannot be retrieved
Sources of noise	Top-coding; tendency to under-report income	Data leakage; label bias; clerical errors	Incorrect code table
Sample representativeness	US working population	Convenience sample (Broward County)	Artificial sample (credit granted, negative class oversampled)
Preprocessing needed	Handling missing values (7%)	Handling missing values (80%); removing redundant features; ground truth on detainment	None
Additional concerns	Accuracy and fairness are sensitive to arbitrary 50K$ threshold	Potential for misguided discussion on criminal justice	Interpretability and exploratory analyses are invalid

COMPAS also represents a case of a broad phenomenon which can be termed *data bias*. With terminology from Friedler et al. (2021), when it comes to datasets encoding complex human phenomena, there is often a disconnect between the *construct space* (what we aim to measure) and the *observed space* (what we end up observing). This may be especially problematic if the difference between construct and observation is uneven across individuals or groups. COMPAS, for example, is a dataset about criminal offense. Offense is central to the prediction target Y, aimed at encoding recidivism, and to the available covariates X, summarizing criminal history. However, the COMPAS dataset (observed space) is an imperfect proxy for the criminal patterns it should summarize (construct space). The prediction labels Y actually encode re-arrest, instead of re-offense (Larson et al. 2016), and are thus clearly influenced by spatially differentiated policing practices (Fogliato et al. 2021). This is also true of criminal history encoded in COMPAS covariates, again mediated by arrest and policing practices which may be racially biased (Bao et al. 2021; Mayson 2018). As a result, the true fairness of an algorithm, just like its accuracy, may differ significantly from what is reported on biased data. For example, algorithms that achieve equality of true positive rates across sensitive groups on COMPAS are deemed fair under the *equal opportunity* measure (Hardt et al. 2016). However, if both the training set on which this objective is enforced and the test set on which it is measured are affected by race-dependent noise described above, those algorithms are only “fair” in an abstract observed space, but not in the real construct space we ultimately care about (Friedler et al. 2021).

Overall, these considerations paint a mixed picture for a dataset of high social relevance that was extremely useful to catalyze attention on algorithmic fairness issues, displaying at the same time several limitations in terms of its continued use as a flexible benchmark for fairness studies of all sorts. In this regard, Bao et al. (2021) suggest avoiding the use of COMPAS to demonstrate novel approaches in algorithmic fairness, as considering the data without proper context may lead to misleading conclusions, which could misguidedly enter the broader debate on criminal justice and risk assessment.

4.3 German credit

The German Credit dataset was created to study the problem of computer-assisted credit decisions at a regional Bank in southern Germany. Instances represent loan applicants from 1973 to 1975, who were deemed creditworthy and were granted a loan, bringing about a natural selection bias. Within this sample, bad credits are oversampled to favour a balance in target classes (Grömping 2019). The data summarizes applicants’ financial situation, credit history, and personal situation, including housing and number of liable people. A binary variable encoding whether each loan recipient punctually payed every installment is the target of a classification task. Among the covariates, marital status and sex are jointly encoded in a single variable. Many documentation mistakes are present in the UCI entry associated with this resource (UCI Machine Learning Repository 1994). A revised version with correct variable encodings, called

South German Credit, was donated to UCI Machine Learning Repository (2019) with an accompanying report (Grömping 2019).

The greatest upside of this dataset is the fact that it captures a real-world application of credit scoring at a bank. On the downside, the data is half a century old, significantly limiting the societally useful insights that can be gleaned from it. Most importantly, the popular release of this dataset (UCI Machine Learning Repository 1994) comes with highly inaccurate documentation which contains wrong variable codings. For example, the variable reporting whether loan recipients are foreign workers has its coding reversed, so that, apparently, fewer than 5% of the loan recipients in the dataset would be German. Luckily, this error has no impact on numerical results obtained from this dataset, as it is irrelevant at the level of abstraction afforded by raw features, with the exception of potentially counterintuitive explanations in works of interpretability and exploratory analysis (Le Quy et al. 2022). This coding error, along with others discussed in Grömping (2019) was corrected in a novel release of the dataset (UCI Machine Learning Repository 2019). Unfortunately and most importantly for the fair ML community, retrieving the sex of loan applicants is simply not possible, unlike the original documentation suggested. This is due to the fact that one value of this feature was used to indicate both women who are divorced, separated, or married, and men who are single, while the original documentation reported each feature value to correspond to same-sex applicants (either male-only or female-only). This particular coding error ended up having a non-negligible impact on the fair ML community, where many works studying group fairness extract sex from the joint variable and use it as a sensitive attribute, even years after the redacted documentation was published (Wang et al. 2021; Le Quy et al. 2022). These coding mistakes are part of a documentation debt whose influence continues to affect the algorithmic fairness community.

4.4 Summary

Adult, COMPAS, and German Credit are the most used datasets in the surveyed algorithmic fairness literature, despite the limitations summarized in Table 1. Their status as de facto fairness benchmarks is probably due to their use in seminal works (Pedreshi et al. 2008; Calders et al. 2009) and influential articles (Angwin et al. 2016) on algorithmic fairness. Once this fame was created, researchers had clear incentives to study novel problems and approaches on these datasets, which, as a result, have become even more established benchmarks in the algorithmic fairness literature (Bao et al. 2021). On close scrutiny, the fundamental merit of these datasets lies in originating from human processes, encoding protected attributes, and having different base rates for the target variable across sensitive groups. Their use in recent works on algorithmic fairness can be interpreted as a signal that the authors have basic awareness of default data practices in the field and that the data was not made up to fit the algorithm. Overarching claims of significance in real-world scenarios stemming from experiments on these datasets should be met with skepticism. Experiments that claim extracting a sex variable from the German Credit dataset should be considered noisy at best. As for alternatives, Bao et al. (2021) suggest employing well-designed simulations. A complementary avenue is to seek different datasets that are relevant for the problem at hand.

We hope that the two hundred data briefs accompanying this work will prove useful in this regard, favouring both domain-oriented and task-oriented searches, according to the classification discussed in the next section.

5 Existing alternatives

In this section, we discuss existing fairness resources from three different perspectives. In Sect. 5.1 we describe the different domains spanned by fairness datasets. In Sect. 5.2 we provide a categorization of fairness tasks supported by the same resources. In Sect. 5.3 we discuss the different roles played by these datasets in fairness research, such as supporting training and benchmarking.

5.1 Domain

Algorithmic fairness concerns arise in any domain where Automated Decision Making (ADM) systems may influence human well-being. Unsurprisingly, the datasets in our survey reflect a variety of areas where ADM systems are studied or deployed, including criminal justice, education, search engines, online marketplaces, emergency response, social media, medicine, and hiring. In Fig. 2, we report a subdivision of the surveyed datasets in different macrodomains.[5] We mostly follow the area-category taxonomy by Scimago,[6] departing from it where appropriate. For example, we consider computer vision and linguistics macrodomains of their own, for the purposes of algorithmic fairness, as much fair ML work has been published in both disciplines. Below we present a description of each macrodomain and its main subdomains, summarized in detail in Table 2.

Computer Science. Datasets from this macrodomain are very well represented, comprising *information systems, social media, library and information sciences, computer networks, and signal processing*. *Information systems* heavily feature datasets on search engines for various items such as text, images, worker profiles, and real estate, retrieved in response to queries issued by users (Occupations in Google Images, Scientist+Painter, Zillow Searches, Barcelona Room Rental, Burst, TaskRabbit, Online Freelance Marketplaces, Bing US Queries, Symptoms in Queries). Other datasets represent problems of item recommendation, covering products, businesses, and movies (Amazon Recommendations, Amazon Reviews, Google Local, MovieLens, FilmTrust). The remaining datasets in this subdomain represent knowledge bases (Freebase15k-237, Wikidata) and automated screening systems (CVs from Singapore, Pymetrics Bias Group). Datasets from *social media* that are not focused on links and relationships between people are also considered part of computer science in this survey. These resources are often focused on text, powering tools, and analyses of hate speech and toxicity (Civil Comments, Twitter Abusive Behavior, Twitter Offensive Language, Twitter Hate Speech Detection, Twitter Online Harassment), dialect (Twit-

[5] The total exceeds 226 due to multiple domains being applicable to some dataset.

[6] See the "subject area" and "subject category" drop down menus from https://www.scimagojr.com/journalrank.php, accessed on March 15, 2022.

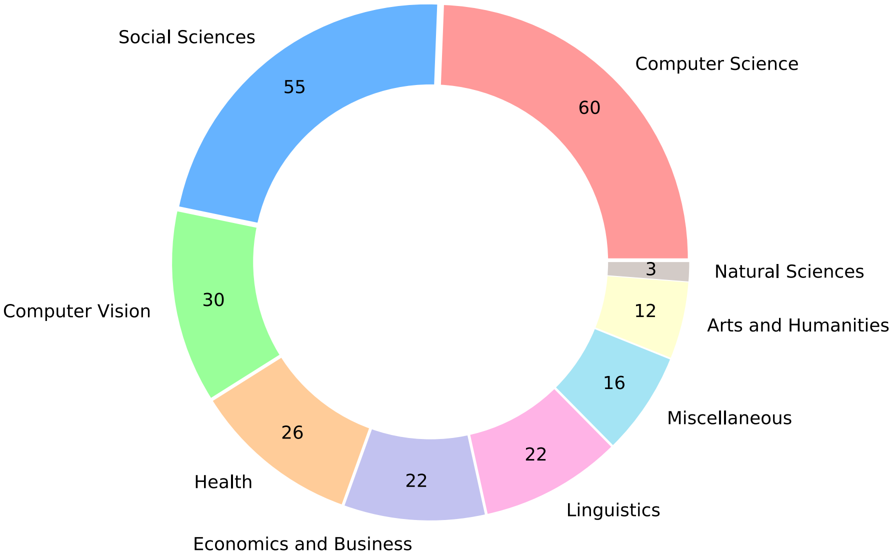

Fig. 2 Datasets employed in fairness research span diverse domains. See Table 2 for a detailed breakdown

terAAE), and political leaning (Twitter Presidential Politics). Twitter is by far the most represented platform, while datasets from Facebook (German Political Posts), Steeemit (Steemit), Instagram (Instagram Photos), Reddit (RtGender, Reddit Comments), Fitocracy (RtGender), and YouTube (YouTube Dialect Accuracy) are also present. Datasets from *library and information sciences* are mainly focused on academic collaboration networks (Cora Papers, CiteSeer Papers, PubMed Diabetes Papers, ArnetMiner Citation Network, 4area, Academic Collaboration Networks), except for a dataset about peer review of scholarly manuscripts (Paper-Reviewer Matching).

Social Sciences. Datasets from social sciences are also plentiful, spanning *law, education, social networks, demography, social work, political science, transportation, sociology* and *urban studies*. *Law* datasets are mostly focused on recidivism (Crowd Judgement, COMPAS, Recidivism of Felons on Probation, State Court Processing Statistics, Los Angeles City Attorney's Office Records) and crime prediction (Strategic Subject List, Philadelphia Crime Incidents, Stop, Question and Frisk, Real-Time Crime Forecasting Challenge, Dallas Police Incidents, Communities and Crime), with a granularity spanning the range from individuals to communities. In the area of *education* we find datasets that encode application processes (Nursery, IIT-JEE), student performance (Student, Law School, UniGe, ILEA, US Student Performance, Indian Student Performance, EdGap, Berkeley Students), including attempts at automated grading (Automated Student Assessment Prize), and placement information after school (Campus Recruitment). Some datasets on student performance support studies of differences across schools and educational systems, for which they report useful features (Law School, ILEA, EdGap), while the remaining datasets are more focused on differences in the individual condition and outcome for students, typically within the same institution. Datasets about *social networks* mostly concern online social networks (Facebook Ego-networks, Facebook Large Network, Pokec Social Network, Rice Facebook Network, Twitch Social Networks, University Facebook Networks), except for High School Contact and Friendship Network, also featuring offline relations. *Demography* datasets comprise census data from different countries (Dutch Census, Indian Census, National Longitudinal Survey of Youth, Section 203 determinations, US Census Data (1990)). Datasets from *social work* cover complex personal and social problems, including child maltreatment prevention (Allegheny Child Welfare), emergency response (Harvey Rescue), and drug abuse prevention (Homeless Youths' Social Networks, DrugNet). Resources from *political science* describe registered voters (North Carolina Voters), electoral precincts (MGGG States), polling (2016 US Presidential Poll), and sortition (Climate Assembly UK). *Transportation* data summarizes trips and fares from taxis (NYC Taxi Trips, Shanghai Taxi Trajectories), ride-hailing (Chicago Ridesharing, Ride-hailing App), and bike sharing services (Seoul Bike Sharing), along with public transport coverage (Equitable School Access in Chicago). *Sociology* resources summarize online (Libimseti) and offline dating (Columbia University Speed Dating). Finally, we assign SafeGraph Research Release to *urban studies*.

Computer Vision. This is an area of early success for artificial intelligence, where fairness typically concerns learned representations and equality of performance across classes. The surveyed articles feature several popular datasets on image classification (ImageNet, MNIST, Fashion MNIST, CIFAR), visual question answering (Visual Question Answering), segmentation and captioning (MS-COCO, Open Images Dataset). We find over ten face analysis datasets (Labeled Faces in the Wild, UTK Face, Adience, FairFace, IJB-A, CelebA, Pilot Parliaments Benchmark, MS-Celeb-1M, Diversity in Faces, Multi-task Facial Landmark, Racial Faces in the Wild, BUPT Faces), including one from experimental psychology (FACES), for which fairness is most often intended as the robustness of classifiers across different subpopulations, without much regard for downstream benefits or harms to these populations. Synthetic images are popular to study the relationship between fairness and disentangled representations (dSprites, Cars3D, shapes3D). Similar studies can be conducted on datasets with spurious correlations between subjects and backgrounds (Waterbirds, Benchmarking Attribution Methods) or gender and occupation (Athletes and health professionals). Finally, the Image Embedding Association Test dataset is a fairness benchmark to study biases in image embeddings across religion, gender, age, race, sexual orientation, disability, skin tone, and weight. It is worth noting that this significant proportion of computer vision datasets is not an artifact of including CVPR in the list of candidate conferences, which contributed just five additional datasets (Multi-task Facial Landmark, Office31, Racial Faces in the Wild, BUPT Faces, Visual Question Answering).

Health. This macrodomain, comprising medicine, psychology and pharmacology displays a notable diversity of subdomains interested by fairness concerns. Specialties represented in the surveyed datasets are mostly medical, including *public health* (Antelope Valley Networks, Willingness-to-Pay for Vaccine, Kidney Matching, Kidney Exchange Program), *cardiology* (Heart Disease, Arrhythmia, Framingham), *endocrinology* (Diabetes 130-US Hospitals, Pima Indians Diabetes Dataset), *health policy* (Heritage Health, MEPS-HC). Specialties such as *radiology* (National

Lung Screening Trial, MIMIC-CXR-JPG, CheXpert) and *dermatology* (SIIM-ISIC Melanoma Classification, HAM10000) feature several image datasets for their strong connections with medical imaging. Other specialties include *critical care medicine* (MIMIC-III), *neurology* (Epileptic Seizures), *pediatrics* (Infant Health and Development Program), *sleep medicine* (Apnea), *nephrology* (Renal Failure), *pharmacology* (Warfarin) and *psychology* (Drug Consumption, FACES). These datasets are often extracted from care data of multiple medical centers to study problems of automated diagnosis. Resources derived from longitudinal studies, including Framingham and Infant Health and Development Program are also present. Works of algorithmic fairness in this domain are typically concerned with obtaining models with similar performance for patients across race and sex.

Linguistics. In addition to the textual resources we already described, such as the ones derived from social media, several datasets employed in algorithmic fairness literature can be assigned to the domain of linguistics and Natural Language Processing (NLP). There are many examples of resources curated to be fairness benchmarks for different tasks, including machine translation (Bias in Translation Templates), sentiment analysis (Equity Evaluation Corpus), coreference resolution (Winogender, Winobias, GAP Coreference), named entity recognition (In-Situ), language models (BOLD) and word embeddings (WEAT). Other datasets have been considered for their size and importance for pretraining text representations (Wikipedia dumps, One billion word benchmark, BookCorpus, WebText) or their utility as NLP benchmarks (GLUE, Business Entity Resolution). Speech recognition resources have also been considered (TIMIT).

Economics and Business. This macrodomain comprises datasets from *economics*, *finance*, *marketing*, and *management information systems*. *Economics* datasets mostly consist of census data focused on wealth (Adult, US Family Income, Poverty in Colombia, Costarica Household Survey) and other resources which summarize employment (ANPE), tariffs (US Harmonized Tariff Schedules), insurance (Italian Car Insurance), and division of goods (Spliddit Divide Goods). *Finance* resources feature data on microcredit and peer-to-peer lending (Mobile Money Loans, Kiva, Prosper Loans Network), mortgages (HMDA), loans (German Credit, Credit Elasticities), credit scoring (FICO) and default prediction (Credit Card Default). *Marketing* datasets describe marketing campaigns (Bank Marketing), customer data (Wholesale) and advertising bids (Yahoo! A1 Search Marketing). Finally, datasets from *management information systems* summarize information about automated hiring (CVs from Singapore, Pymetrics Bias Group) and employee retention (IBM HR Analytics).

Miscellaneous. This macrodomain contains several datasets originating from the *news* domain (Yow news, Guardian Articles, Latin Newspapers, Adressa, Reuters 50 50, New York Times Annotated Corpus, TREC Robust04). Other resources include datasets on food (Sushi), sports (Fantasy Football, FIFA 20 Players, Olympic Athletes) , and toy datasets (Toy Dataset 1–4).

Arts and Humanities. In this area we mostly find *literature* datasets, which contain text from literary works (Shakespeare, Curatr British Library Digital Corpus, Victorian Era Authorship Attribution, Nominees Corpus, Riddle of Literary Quality), which are typically studied with NLP tools. Other datasets in this domain include domain-specific information systems about books (Goodreads Reviews), *movies* (MovieLens) and *music* (Last.fm, Million Song Dataset, Million Playlist Dataset).

Natural Sciences. This domain is represented with three datasets from *biology* (iNaturalist), *biochemestry* (PP-Pathways) and *plant science*, with the classic Iris dataset.

As a whole, many of these datasets encode fundamental human activities where algorithms and ADM systems have been studied and deployed. Alertness and attention to equity seems especially important in specific domains, including social sciences, computer science, medicine, and economics. Here the potential for impact may result in large benefits, but also great harm, particularly for vulnerable populations and minorities, more likely to be neglected during the design, training, and testing of an ADM. After concentrating on domains, in the next section we analyze the variety of tasks studied in works of algorithmic fairness and supported by these datasets.

5.2 Task and setting

Researchers and practitioners are showing an increasing interest in algorithmic fairness, proposing solutions for many different *tasks*, including fair classification, regression, and ranking. At the same time, the academic community is developing an improved understanding of important challenges that run across different tasks in the algorithmic fairness space (Chouldechova and Roth 2020), also thanks to practitioner surveys (Holstein et al. 2019) and studies of specific legal challenges (Andrus et al. 2021). To exemplify, the presence of noise corrupting labels for sensitive attributes represents a challenge that may apply across different tasks, including fair classification, regression, and ranking. We refer to these challenges as *settings*, describing them in the second part of this section. While our work focuses on fair ML datasets, it is cognizant of the wide variety of tasks tackled in the algorithmic fairness literature, which are captured in a specific field of our data briefs. In this section we provide an overview of common tasks and settings studied on these datasets, showing their variety and diversity. Table 3 summarizes these tasks, listing the three most used datasets for each task. When describing a task, we explicitly highlight datasets that are particularly relevant to it, even when outside of the top three.

5.2.1 Task

Fair classification (Calders and Verwer 2010; Dwork et al. 2012) is the most common task by far. Typically, it involves equalizing some measure of interest across subpopulations, such as the recall, precision, or accuracy for different racial groups. On the other hand, individually fair classification focuses on the idea that similar individuals (low distance in the covariate space) should be treated similarly (low distance in the outcome space), often formalized as a Lipschitz condition. Unsurprisingly, the most common datasets for fair classification are the most popular ones overall (Sect. 4), i.e., Adult, COMPAS, and German Credit.

Fair regression (Berk et al. 2017) concentrates on models that predict a real-valued target, requiring the average loss to be balanced across groups. Individual fairness in

Table 2 A selection of datasets through the lens of the domain taxonomy

Domain	Sample datasets
Computer science	
Social media	
Toxicity and hate speech	Civil Comments, Wikipedia Toxic Comments, Twitter offensive language
Political leaning	Twitter Presidential Politics
Dialect	TwitterAAE
Library and information sciences	
Collaboration networks	Paper-reviewer matching, 4area, ArnetMiner Citation network
Peer review	Paper-reviewer matching
Information systems	
Search engines	Online freelance marketplaces, Bing US Queries, Symptoms in Queries
Recommender systems	Amazon Recommendations, Amazon Reviews, MovieLens
Knowledge bases	Freebase15k-237, Wikidata
Computer networks	KDD Cup 99
Pattern recognition	Internet Ads
Signal processing	Vehicle
Social sciences	
Urban studies	SafeGraph Research Release
Social networks	University Facebook Networks, Pokec Social Network, Rice Facebook Network
Demography	US Census Data (1990), Dutch Census, National Longitudinal Survey of Youth
Sociology	Columbia University Speed Dating, Libimseti
Law	
Recidivism prediction	COMPAS, recidivism of felons on probation, state court processing statistics
Crime prediction	Communities and crime, stop, question and frisk, strategic subject list
Political science	
Registered voters	North Carolina Voters
Electoral precincts	MGGG States
Polling	2016 US Presidential Poll
Sortition	Climate Assembly UK
Education	
Application processes	Nursery, IIT-JEE
Student performance	Student, Law School, UniGe
Post-education placement	Campus Recruitment

Table 2 continued

Domain	Sample datasets
Social work	
Child maltreatment prevention	Allegheny Child Welfare
Emergency response	Harvey Rescue
Drug abuse prevention	Homeless Youths' Social Networks, DrugNet
Transportation	
Taxi trips	NYC Taxi Trips, Shanghai Taxi Trajectories
Ride hailing	Chicago Ridesharing, Ride-hailing App
Bike sharing	Seoul Bike Sharing
Public transport	Equitable School Access in Chicago
Computer vision	
General purpose	ImageNet, MNIST, CIFAR
Face analysis	CelebA, Pilot Parliaments Benchmar, FairFace
Synthetic	dSprites, Cars3D, shapes3D
Health	
Sleep medicine	Apnea
Critical care medicine	MIMIC-III
Public health	Kidney exchange program, willingness-to-pay for vaccine, kidney matching
Cardiology	Arrhythmia, heart disease, framingham
Neurology	Epileptic seizures
Pediatrics	Infant Health and Development Program (IHDP)
Dermatology	HAM10000, SIIM-ISIC melanoma classification
Medicine	Stanford medicine research data repository
Pharmacology	Warfarin
Endocrinology	Diabetes 130-US Hospitals, Pima Indians diabetes dataset (PIDD)
Nephrology	Renal failure
Radiology	CheXpert, MIMIC-CXR-JPG, national lung screening trial (NLST)
Health policy	Heritage Health, MEPS-HC
Applied psychology	Drug consumption
Experimental psychology	FACES
Economics and business	
Economics	
Census	Adult, US Family Income, Poverty in Colombia
Employment	ANPE
Tariffs	US Harmonized tariff schedule
Insurance	Italian car insurance
Division of goods	Spliddit divide goods

Table 2 continued

Domain	Sample datasets
Finance	
Peer-to-peer lending	Mobile money loans, kiva, prosper loans network
Mortgages	HMDA
Credit scoring	FICO
Other credit	German credit, credit card default, credit elasticities
Marketing	
Marketing campaigns	Bank Marketing
Advertising bids	Yahoo! A1 search marketing, wholesale
Management information systems	
Automated hiring	Pymetrics bias group, CVs from Singapore
Employee retention	IBM HR analytics
Linguistics	
General purpose	Wikipedia dumps, one billion word benchmark, BookCorpus
Fairness benchmarks	Bias in translation templates, equity evaluation corpus, Winogender
Arts and Humanities	
Music	Million playlist dataset (MPD), million song dataset (MSD), Last.fm
Literature	Goodreads reviews, riddle of literary quality, nominees corpus
Movies	MovieLens, FilmTrust
Natural sciences	
Biology	iNaturalist Datasets
Biochemestry	PP-Pathways
Plant science	Iris
Miscellaneous	
News	TREC Robust04, New York Times Annotated Corpus, Reuters 50 50
Sports	Fantasy football, FIFA 20 Players, Olympic Athletes
Food	Sushi

this context may require losses to be as uniform as possible across all individuals. Fair regression is a less popular task, often studied on the Communities and Crime dataset, where the task is predicting the rate of violent crimes in different communities.

Fair ranking (Yang and Stoyanovich 2017) requires ordering candidate items based on their relevance to a current need. Fairness in this context may concern both the people producing the items that are being ranked (e.g. artists) and those consuming the items (users of a music streaming platform). It is typically studied in applications of recommendation (MovieLens, Amazon Recommendations, Last.fm, Million Song Dataset, Adressa) and search engines (Yahoo! c14B Learning to Rank, Microsoft Learning to Rank, TREC Robust04).

Table 3 Most used datasets by algorithmic fairness task and setting

Task	Datasets
Fair classification	Adult; COMPAS; German Credit
Fair regression	Communities and Crime; Law School; Student
Fair ranking	MovieLens; German Credit; Kiva
Fair matching	NYC Taxi Trips; Libimseti; Columbia University Speed Dating
Fair risk assessment	COMPAS; Allegheny Child Welfare; Infant Health and Development Program (IHDP)
Fair representation learning	Adult; COMPAS; dSprites
Fair clustering	Adult; Bank Marketing; Diabetes 130-US Hospitals
Fair anomaly detection	Adult; MNIST; Credit Card Default
Fair districting	MGGG States
Fair task assignment	Crowd Judgement; COMPAS
Fair spatio-temporal process learning	Real-Time Crime Forecasting Challenge; Dallas Police Incidents; Harvey Rescue
Fair graph diffusion/augmentation	University Facebook Networks; Antelope Valley Networks; Rice Facebook Network
Fair resource allocation/subset selection	ML Fairness Gym; US Federal Judges; Climate Assembly UK
Fair data summarization	Adult; Student; Credit Card Default
Fair data generation	CelebA; MovieLens; shapes3D
Fair graph mining	MovieLens; Freebase15k-237; PP-Pathways
Fair pricing	Willingness-to-Pay for Vaccine; Credit Elasticities; Italian Car Insurance
Fair advertising	Yahoo! A1 Search Marketing; North Carolina Voters; Instagram Photos
Fair routing	Shanghai Taxi Trajectories
Fair entity resolution	Winogender; Winobias; Business Entity Resolution
Fair sentiment analysis	Popular Baby Names; Equity Evaluation Corpus (EEC); TwitterAAE
Bias in word embeddings	Wikipedia dumps; Word Embedding Association Test (WEAT); Popular Baby Names
Bias in language models	TwitterAAE; BOLD; GLUE
Fair machine translation	Bias in Translation Templates
Fair speech recognition	YouTube Dialect Accuracy; TIMIT
Setting	**Datasets**
Rich-subgroup fairness	Adult; COMPAS; Communities and Crime
Fairness under unawareness	Adult; COMPAS; HMDA
Limited-label fairness	Adult; German Credit; COMPAS
Robust fairness	COMPAS; Adult; MEPS-HC
Dynamical fairness	FICO; ML Fairness Gym; COMPAS
Preference-based fairness	Adult; COMPAS; Toy Dataset 1
Multi-stage fairness	Adult; Heritage Health; Twitter Offensive Language

Table 3 continued

Setting	Datasets
Fair few-shot learning	Communities and Crime; Toy Dataset 1; Mobile Money Loans
Fair private learning	UTK Face; CheXpert; FairFace
Fair federated learning	Vehicle; Sentiment140; Shakespeare
Fair incremental learning	ImageNet; CIFAR
Fair active learning	Adult; German Credit; Heart Disease
Fair selective classification	CheXpert; CelebA; Civil Comments

Fair matching (Kobren et al. 2019) is similar to ranking as they are both tasks defined on two-sided markets. This task, however, is focused on highlighting and matching pairs of items on both sides of the market, without emphasis on the ranking component. Datasets for this task are from diverse domains, including dating (Libimseti, Columbia University Speed Dating) transportation (NYC Taxi Trips, Ride-hailing App) and organ donation (Kidney Matching, Kidney Exchange Program).

Fair risk assessment (Coston et al. 2020) studies algorithms that score instances in a dataset according to a predefined type of risk. Relevant domains include healthcare and criminal justice. Key differences with respect to classification are an emphasis on real-valued scores rather than labels, and awareness that the risk assessment process can lead to interventions impacting the target variable. For this reason, fairness concerns are often defined in a counterfactual fashion. The most popular dataset for this task is COMPAS, followed by datasets from medicine (IHDP, Stanford Medicine Research Data Repository), social work (Allegheny Child Welfare), Economics (ANPE) and Education (EdGap).

Fair representation learning (Creager et al. 2019) concerns the study of features learnt by models as intermediate representations for inference tasks. A popular line of work in this space, called *disentaglement*, aims to learn representations where a single factor of import corresponds to a single feature. Ideally, this approach should select representations where sensitive attributes cannot be used as proxies for target variables. Cars3D and dSprites are popular datasets for this task, consisting of synthetic images depicting controlled shape types under a controlled set of rotations. Post-processing approaches are also applicable to obtain fair representations from biased ones via debiasing.

Fair clustering (Chierichetti et al. 2017) is an unsupervised task concerned with the division of a sample into homogenous groups. Fairness may be intended as an equitable representation of protected subpopulations in each cluster, or in terms of average distance from the cluster center. While Adult is the most common dataset for problems of fair clustering, other resources often used for this task include Bank Marketing, Diabetes 130-US Hospitals, Credit Card Default and US Census Data (1990).

Fair anomaly detection (Zhang and Davidson 2021), also called **outlier detection** (Davidson and Ravi 2020), is aimed at identifying surprising or anomalous points in a dataset. Fairness requirements involve equalizing salient quantities (e.g. acceptance rate, recall, precision, distribution of anomaly scores) across populations of interest. This problem is particularly relevant for members of minority groups, who, in the absence of specific attention to dataset inclusivity, are less likely to fit the norm in the feature space.

Fair districting (Schutzman 2020) is the division of a territory into electoral districts for political elections. Fairness notions brought forth in this space are either outcome-based, requiring that seats earned by a party roughly match their share of the popular vote, or procedure-based, ignoring outcomes and requiring that counties or municipalities are split as little as possible. MGGG States is a reference resource for this task, providing precinct-level aggregated information about demographics and political leaning of voters in US districts.

Fair task assignment and **truth discovery** (Goel and Faltings 2019; Li et al. 2020d) are different subproblems in the same area, focused on the subdivision of work and the aggregation of answers in crowdsourcing. Here fairness may be intended concerning errors in the aggregated answer, requiring errors to be balanced across subpopulations of interest, or in terms of the work load imposed to workers. A dataset suitable for this task is Crowd Judgement, containing crowd-sourced recidivism predictions.

Fair spatio-temporal process learning (Shang et al. 2020) focuses on the estimation of models for processes which evolve in time and space. Surveyed applications include crime forecasting (Real-Time Crime Forecasting Challenge, Dallas Police Incidents) and disaster relief (Harvey Rescue), with fairness requirements focused on equalization of performance across different neighbourhoods and special attention to their racial composition.

Fair graph diffusion (Farnad et al. 2020) models and optimizes the propagation of information and influence over networks, and its probability of reaching individuals of different sensitive groups. Applications include obesity prevention (Antelope Valley Networks) and drug-use prevention (Homeless Youths' Social Networks). **Fair graph augmentation** (Ramachandran et al. 2021) is a similar task, defined on graphs which define access to resources based on existing infrastructure (e.g. transportation), which can be augmented under a budget to increase equity. This task has been proposed to improve school access (Equitable School Access in Chicago) and information availability in social networks (Facebook100).

Fair resource allocation/subset selection (Babaioff et al. 2019; Huang et al. 2020) can often be formalized as a classification problem with constraints on the number of positives. Fairness requirements are similar to those of classification. Subset selection may be employed to choose a group of people from a wider set for a given task (US Federal Judges, Climate Assembly UK). Resource allocation concerns the division of goods (Spliddit Divide Goods) and resources (ML Fairness Gym, German Credit).

Fair data summarization (Celis et al. 2018) refers to presenting a summary of datasets that is equitable to subpopulations of interest. It may involve finding a small subset representative of a larger dataset (strongly linked to subset selection) or selecting the most important features (dimensionality reduction). Approaches for this task have been applied to select a subset of images (Scientist+Painter) or customers (Bank Marketing), that represent the underlying population across sensitive demographics.

Fair data generation (Xu et al. 2018) deals with generating "fair" data points and labels, which can be used as training or test sets. Approaches in this space may be used to ensure an equitable representation of protected categories in data generation processes learnt from biased datasets (CelebA, IBM HR Analytics), and to evaluate biases in existing classifiers (MS-Celeb-1M). Data generation may also be limited to synthesizing artificial sensitive attributes (Burke et al. 2018a).

Fair graph mining (Kang et al. 2020) focuses on representations and prediction tasks on graph structures. Fairness may be defined either as a lack of bias in representations, or with respect to a final inference task defined on the graph. Fair graph mining approaches have been applied to knowledge bases (Freebase15k-237, Wikidata), collaboration networks (CiteSeer Paper, Academic Collaboration Networks) and social network datasets (Facebook Large Network, Twitch Social Networks).

Fair pricing (Kallus and Zhou 2021) concerns learning and deploying an optimal pricing policy for revenue while maintaining equity of access to services and consumer welfare across sensitive groups. Datasets employed in fair pricing are from the economics (Credit Elasticities, Italian Car Insurance), transportation (Chicago Ridesharing), and public health domains (Willingness-to-Pay for Vaccine).

Fair advertising (Celis et al. 2019a) is also concerned with access to goods and services. It comprises both bidding strategies and auction mechanisms which may be modified to reduce discrimination with respect to the gender or race composition of the audience that sees an ad. One publicly available dataset for this subtask is Yahoo! A1 Search Marketing.

Fair routing (Qian et al. 2015) is the task of suggesting an optimal path from a starting location to a destination. For this task, experimentation has been carried out on a semi-synthetic traffic dataset (Shanghai Taxi Trajectories). The proposed fairness measure requires equalizing the driving cost per customer across all drivers.

Fair entity resolution (Cotter et al. 2019) is a task focused on deciding whether multiple records refer to the same entity, which is useful, for instance, for the construction and maintenance of knowledge bases. Business Entity Resolution is a proprietary dataset for fair entity resolution, where constraints of performance equality across chain and non-chain businesses can be tested. Winogender and Winobias are publicly available datasets developed to study gender biases in pronoun resolution.

Fair sentiment analysis (Kiritchenko and Mohammad 2018) is a well-established instance of fair classification, where text snippets are typically classified as positive, negative, or neutral depending on the sentiment they express. Fairness is intended with respect to the entities mentioned in the text (e.g. men and women). The central idea is that the estimated sentiment for a sentence should not change if female entities (e.g. “her”, “woman”, “Mary”) are substituted with their male counterparts (“him”, “man”, “James”). The Equity Evaluation Corpus is a benchmark developed to assess gender and race bias in sentiment analysis models.

Bias in Word Embeddings (WEs) (Bolukbasi et al. 2016) is the study of undesired semantics and stereotypes captured by vectorial representations of words. WEs are typically trained on large text corpora (Wikipedia dumps) and audited for associations between gendered words (or other words connected to sensitive attributes) and stereotypical or harmful concepts, such as the ones encoded in WEAT.

Bias in Language Models (LMs) (Bordia and Bowman 2019) is, quite similarly, the study of biases in LMs, which are flexible models of human language based on contextualized word representations, which can be employed in a variety of linguistics and NLP tasks. LMs are trained on large text corpora from which they may learn spurious correlations and stereotypes. The BOLD dataset is an evaluation benchmark for LMs, based on prompts that mention different socio-demographic groups. LMs complete these prompts into full sentences, which can be tested along different dimensions (sentiment, regard, toxicity, emotion and gender polarity).

Fair Machine Translation (MT) (Stanovsky et al. 2019) concerns automatic translation of text from a source language into a target one. MT systems can exhibit gender biases, such as a tendency to translate gender-neutral pronouns from the source language into gendered pronouns of the target language in accordance with gender stereotypes. For example, a "nurse" mentioned in a gender-neutral context in the source sentence may be rendered with feminine grammar in the target language. Bias in Translation Templates is a set of short templates to test such biases.

Fair speech recognition (Tatman 2017) requires accurate annotation of spoken language into text across different demographics. YouTube Dialect Accuracy is a dataset developed to audit the accuracy of YouTube's automatic captions across two genders and five dialects of English. Similarly, TIMIT is a classic speech recognition dataset annotated with American English dialect and gender of speaker.

5.2.2 Setting

As noted at the beginning of this section, there are several *settings* (or challenges) that run across different tasks described above. Some of these settings are specific to fair ML, such as ensuring fairness across an exponential number of groups, or in the presence of noisy labels for sensitive attributes. Other settings are connected with common ML challenges, including few-shot and privacy-preserving learning. Below we describe common settings encountered in the surveyed articles. Most of these settings are tested on fairness datasets which are popular overall, i.e. Adult, COMPAS, and German Credit. We highlight situations where this is not the case, potentially due to a given challenge arising naturally in some other dataset.

Rich-subgroup fairness (Kearns et al. 2018) is a setting where fairness properties are required to hold not only for a limited number of protected groups, but across an exponentially large number of subpopulations. This line of work represents an attempt to bridge the normative reasoning underlying individual and group fairness.

Fairness under unawareness is a general expression to indicate problems where sensitive attributes are missing (Chen et al. 2019a), encrypted (Kilbertus et al. 2018) or corrupted by noise (Lamy et al. 2019). These problems respond to real-world challenges related to the confidential nature of protected attributes, that individuals may wish to hide, encrypt, or obfuscate. This setting is most commonly studied on highly popular fairness dataset (Adult, COMPAS), moderately popular ones (Law School and Credit Card Default), and a dataset about home mortgage applications in the US (HMDA).

Limited-label fairness comprises settings with limited information on the target variable, including situations where labelled instances are few (Ji et al. 2020), noisy (Wang et al. 2021), or only available in aggregate form (Sabato and Yom-Tov 2020).

Robust fairness problems arise under perturbations to the training set (Huang and Vishnoi 2019), adversarial attacks (Nanda et al. 2021) and dataset shift (Singh et al. 2021). This line of research is often connected with work in robust machine learning, extending the stability requirements beyond accuracy-related metrics to fairness-related ones.

Dynamical fairness (Liu et al. 2018; D'Amour et al. 2020) entails repeated decisions in changing environments, potentially affected by the very algorithm that is being studied. Works in this space study the co-evolution of algorithms and populations on which they act over time. For example, an algorithm that achieves equality of acceptance rates across protected groups in a static setting may generate further incentives for the next generation of individuals from historically disadvantaged groups. Popular resources for this setting are FICO and the ML Fairness GYM.

Preference-based fairness (Zafar et al. 2017b) denotes work informed, explicitly or implicitly, by the preferences of stakeholders. For people subjected to a decision this is related to notions of envy-freeness and loss aversion (Ali et al. 2019b); alternatively, policy-makers can express indications on how to trade-off different fairness measures (Zhang et al. 2020c), or experts can provide demonstrations of fair outcomes (Galhotra et al. 2021).

Multi-stage fairness (Madras et al. 2018b) refers to settings where several decision makers coexist in a compound decision-making process. Decision makers, both humans and algorithmic, may act with different levels of coordination. A fundamental question in this setting is how to ensure fairness under composition of different decision mechanisms.

Fair few-shot learning (Zhao et al. 2020b) aims at developing fair ML solutions in the presence of a small amount of data samples. The problem is closely related to, and possibly solved by, **fair transfer learning** (Coston et al. 2019) where the goal is to exploit the knowledge gained on a problem to solve a different but related one. Datasets where this setting arises naturally are Communities and Crime, where one may restrict the training set to a subset of US states, and Mobile Money Loans, which consists of data from different African countries.

Fair private learning (Bagdasaryan et al. 2019; Jagielski et al. 2019) studies the interplay between privacy-preserving mechanisms and fairness constraints. Works in this space consider the equity of machine learning models designed to avoid leakage of information about individuals in the training set. Common domains for datasets employed in this setting are face analysis (UTK Face, FairFace, Diversity in Face) and medicine (CheXpert, SIIM-ISIC Melanoma Classification, MIMIC-CXR-JPG).

Additional settings that are less common include **fair federated learning** (Li et al. 2020b), where algorithms are trained across multiple decentralized devices, **fair incremental learning** (Zhao et al. 2020a), where novel classes may be added to the learning problem over time, **fair active learning** (Noriega-Campero et al. 2019), allowing for the acquisition of novel information during inference and **fair selective classification** (Jones et al. 2021), where predictions are issued only if model confidence is above a certain threshold.

Overall, we found a variety of tasks defined on fairness datasets, ranging from generic, such as *fair classification*, to narrow and specifically defined on certain datasets, such as *fair districting* on MGGG States and *fair truth discovery* on Crowd Judgement. Orthogonally to this dimension, many settings or challenges may arise to complicate these tasks, including noisy labels, system dynamics, and privacy concerns.

Quite clearly, algorithmic fairness research has been expanding in both directions, by studying a variety of tasks under diverse and challenging settings. In the next section, we analyze the roles played in scholarly works by the surveyed datasets.

5.3 Role

The datasets used in algorithmic fairness research can play different roles. For example, some may be used to train novel algorithms, while others are suited to test existing algorithms from a specific point of view. Chapter 7 of Barocas et al. (2019), describes six different roles of datasets in machine learning. We adopt their framework to analyse fair ML datasets, adding to the taxonomy two roles that are specific to fairness research.

A source of real data. While synthetic datasets and simulations may be suited to demonstrate specific properties of a novel method, the usefulness of an algorithm is typically established on data from the real world. More than a sign of immediate applicability to important challenges, good performance on real-world sources of data signals that the researchers did not make up the data to suit the algorithm. This is likely the most common role for fairness datasets, especially common for the ones hosted on the UCI ML repository, including Adult, German Credit, Communities and Crime, Diabetes 130-US Hospitals, Bank Marketing, Credit Card Default, US Census Data (1990). These resources owe their popularity in fair ML research to being a product of human processes and to encoding protected attributes. Quite simply, they are sources of real human data.

A catalyst of domain-specific progress. Datasets can spur algorithmic insight and bring about domain-specific progress. Civil Comments is a great example of this role, powering the Jigsaw Unintended Bias in Toxicity Classification challenge. The challenge responds to a specific need in the space of automated moderation against toxic comments in online discussion. Early attempts at toxicity detection resulted in models which associate mentions of frequently attacked identities (e.g. gay) with toxicity, due to spurious correlations in training sets. The dataset and associated challenge tackle this issue by providing toxicity ratings for comments, along with labels encoding whether members of a certain group are mentioned, favouring measurement of undesired bias. Many other datasets can play a similar role, including, Winogender, Winobias and the Equity Evaluation Corpus. In a broader sense, COMPAS and the accompanying study (Angwin et al. 2016) have been an important catalyst, not for a specific task, but for fairness research overall.

A way to numerically track progress on a problem. This role is common for machine learning benchmarks that also provide human performance baselines. Algorithmic methods approaching or surpassing these baselines are often considered a sign that the task is "solved" and that harder benchmarks are required (Barocas et al. 2019). Algorithmic fairness is a complicated, context-dependent, contested construct whose correct measurement is continuously debated. Due to this reason, we are unaware of any dataset having a similar role in the algorithmic fairness literature.

A resource to compare models. Practitioners interested in solving a specific problem may take a large set of algorithms and test them on a group of datasets that are representative of their problem, in order to select the most promising ones. For well-established

ML challenges, there are often leaderboards providing a concise comparison between algorithms for a given task, which may be used for model selection. This setting is rare in the fairness literature, also due to inherent difficulties in establishing a single measure of interest in the field. One notable exception is represented by Friedler et al. (2019), who employed a suite of four datasets (Adult, COMPAS, German Credit, Ricci) to compare the performance of four different approaches to fair classification.

A source of pre-training data. Flexible, general-purpose models are often pre-trained to encode useful representations, which are later fine-tuned for specific tasks in the same domain. For example, large text corpora are often employed to train language models and word embeddings which are later specialized to support a variety of downstream NLP applications. Wikipedia dumps, for instance, are often used to train word embeddings and investigate their biases (Brunet et al. 2019; Liang and Acuna 2020; Papakyriakopoulos et al. 2020). Several algorithmic fairness works aim to study and mitigate undesirable biases in learnt representations. Corpora like Wikipedia dumps are used to obtain representations via realistic pretraining procedures that mimic common machine learning practice as closely as possible.

A source of training data. Models for a specific task are typically learnt from training sets that encode relations between features and target variable in a representative fashion. One example from the fairness literature is Large Movie Review, used to train sentiment analysis models, later audited for fairness (Liang and Acuna 2020). For fairness audits, one alternative would be resorting to publicly available models, but sometimes a close control on the training corpus and procedure is necessary. Indeed, it is interesting to study issues of model fairness in relation to biases present in the respective training corpora, which can help explain the causes of bias (Brunet et al. 2019). Some works measure biases in internal model representations before and after fine-tuning on a training set, and regard the difference as a measure of bias in the training set. Babaeianjelodar et al. (2020) employ this approach to measure biases in RtGender, Civil Comments, and datasets from GLUE.

A representative summary of a service. Much important work in the fairness literature is focused on measuring fairness and harms in the real world. This line of work includes audits of products and services, which rely on datasets extracted from the application of interest. Datasets created for this purpose include Amazon Recommendations, Pymetrics Bias Group, Occupations in Google Images, Zillow Searches, Online Freelance Marketplaces, Bing US Queries, YouTube Dialect Accuracy. Several other datasets were originally created for this purpose and later repurposed in the fairness literature as sources of real data, including Stop Question and Frisk, HMDA, Law School, and COMPAS.

An important source of data. Some datasets acquire a pivotal role in research and industry, to the point of being considered a de-facto standard for a given purpose. This status warrants closer scrutiny of the dataset, through which researchers aim to uncover potential biases and problematic aspects that may impact models and insights derived from the dataset. ImageNet, for instance, is a dataset with millions of images across thousands of categories. Since its release in 2011, this resource has been used to train, benchmark, and compare hundreds of computer vision models. Given its status in machine learning research, ImageNet has been the subject of two quantitative investigations analyzing its biases and other problematic aspects in the person subtree, uncovering issues of representation (Yang et al. 2020b) and non-consensuality (Prabhu and Birhane 2020). A different data bias audit was carried out on SafeGraph Research Release. SafeGraph data captures mobility patterns in the US, with data from nearly 50 million mobile devices obtained and maintained by Safegraph, a private data company. Their recent academic release has become a fundamental resource for pandemic research, to the point of being used by the Centers for Disease Control and Prevention to measure the effectiveness of social distancing measures (Moreland et al. 2020). To evaluate its representativeness for the overall US population, Coston et al. (2021) have studied selection biases in this dataset.

In algorithmic fairness research, datasets play similar roles to the ones they play in machine learning according to Barocas et al. (2019), including training, catalyzing attention, and signalling awareness of common data practices. One notable exception is that fairness datasets are not used to track algorithmic progress on a problem over time, likely due to the fact that there is no consensus on a single measure to be reported. On the other hand, two roles peculiar to fairness research are summarizing a service or product that is being audited, and representing an important resource whose biases and ethical aspects are particularly worthy of attention. We note that these roles are not mutually exclusive and that datasets can play multiple roles. COMPAS, for example, was originally curated to perform an audit of pretrial risk assessment tools and was later used extensively in fair ML research as a source of real human data, becoming, overall, a catalyst for fairness research and debate.

In sum, existing fairness datasets originate from a variety of domains, support diverse tasks, and play different roles in the algorithmic fairness literature. We hope our work will contribute to establishing principled data practices in the field, to guide an optimal usage of these resources. In the next section we continue our discussion on the key features of these datasets with a change of perspective, asking which lessons can be learnt from existing resources for the curation of novel ones.

6 Best practices for dataset curation

In this section, we analyze the surveyed datasets from different perspectives, typical of critical data studies, human–computer interaction, and computer-supported cooperative work. In particular, we discuss concerns of re-identification (Sect. 6.1), consent (Sect. 6.2), inclusivity (Sect. 6.3), sensitive attribute labeling (Sect. 6.4) and transparency (Sect. 6.5). We describe a range of approaches and consideration to these topics, ranging from negligent to conscientious. Our aim is to make these concerns and related desiderata more visible and concrete, to help inform responsible curation of novel fairness resources, whose number has been increasing in recent years (Fig. 3).

6.1 Re-identification

Motivation. Data re-identification (or de-anonymization) is a practice through which instances in a dataset, theoretically representing people in an anonymized fashion, are successfully mapped back to the respective individuals. Their identity is thus discov-

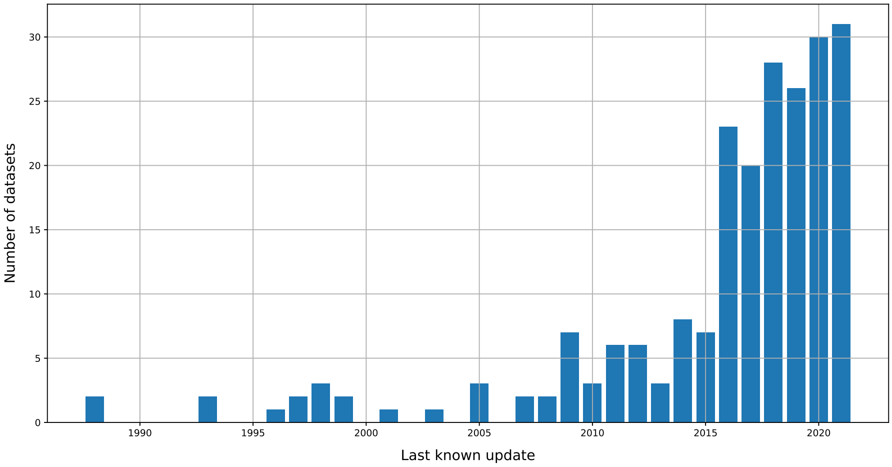

Fig. 3 Most datasets employed in algorithmic fairness were created or updated after 2015, with a clear growth in recent years

ered and associated with the information encoded in the dataset features. Examples of external re-identification attacks include de-anonymization of movie ratings from the Netflix prize dataset (Narayanan and Shmatikov 2008), identification of profiles based on social media group membership (Wondracek et al. 2010), and identification of people depicted in verifiably pornographic categories of ImageNet (Prabhu and Birhane 2020). These analyses were carried out as "attacks" by external teams for demonstrative purposes, but dataset curators and stakeholders may undertake similar efforts internally (McKenna 2019b).

There are multiple harms connected to data re-identification, especially the ones featured in algorithmic fairness research, due to their social significance. Depending on the domain and breadth of information provided by a dataset, malicious actors may acquire information about mobility patterns, consumer habits, political leaning, psychological traits, and medical conditions of individuals, just to name a few. The potential for misuse is tremendous, including phishing attacks, blackmail, threat, and manipulation (Kröger et al. 2021). Face recognition datasets are especially prone to successful re-identification as, by definition, they contain information strongly connected with a person's identity. The problem also extends to general purpose computer vision datasets. In a recent dataset audit, Prabhu and Birhane (2020) found images of beach voyeurism and other non-consensual depictions in ImageNet, and were able to identify the victims using reverse image search engines, highlighting downstream risks of blackmail and other forms of abuse.

Disparate consideration. In this work, we find that fairness datasets are proofed against re-identification with a full range of measures and care. Perhaps surprisingly, some datasets allow for straightforward re-identification of individuals, providing their full names. We do not discuss these resources here to avoid amplifying the harms discussed above. Other datasets afford plausible re-identification, providing social media handles and aliases, such as Twitter Abusive Behavior, Sentiment140, Facebook Large

Network, and Google Local. Columbia University Speed Dating may also fall in this category due to a restricted population from which the sample is drawn, and provision of age, field of study and ZIP code where participants grew up in addition. In contrast, many datasets come with strong guarantees against de-anonymization, which is especially typical of health data, such as MIMIC-III and Heritage Health (El Emam et al. 2012). Indeed, health is a domain where a culture of patient record confidentiality is widely established and there is a strong attention to harm avoidance. Also datasets describing scholarly works and academic collaboration networks (Academic Collaboration Networks, PubMed Diabetes Papers, Cora, CiteSeer) are typically de-identified, with numerical IDs substituting names. This is possibly a sign of attention to anonymization from curators when the data represents potential colleagues. As a consequence, researchers are protected from related harms, but posterior annotation of sensitive attributes similarly to Biega et al. (2019) becomes difficult or impossible. One notable exception is ArnetMiner Citation Network, derived from an online platform which is especially focused on data mining from academic social networks and profiling of researchers.

Mitigating factors. A wide range of factors, summarized in Table 4. may help to reduce the risk of re-identification. A first set of approaches concerns the distribution of data artefacts. Some datasets are simply kept private, minimizing risks in this regard. These include UniGe, US Student Performance, Apnea, Symptoms in Queries and Pymetrics Bias Group, the last two being proprietary datasets that are not disclosed to preserve intellectual property. Twitter Online Harrassment is available upon request to protect the identities of Twitter users that were included. Another interesting approach are mixed release strategies: NLSY has some publicly available data, while access to further information that may favour re-identification (e.g. ZIP code and census tract) is restricted. For crawl-based datasets, it is possible to keep a resource private while providing code to recreate it (Bias in Bios). While this may alleviate some concerns, it will not deter motivated actors. As a post-hoc remedy, proactive removal of problematic instances is also a possibility, as shown by recent work on ImageNet (Yang et al. 2020b).

Another family of approaches is based on redaction, aggregation, and injection of noise. Obfuscation typically involves the distribution of proprietary company data at a level of abstraction which maintains utility to a company while hindering reconstruction of the underlying human-readable data, which also makes re-identification highly unlikely (Yahoo! c14B Learn to Rank, Microsoft Learning to Rank). Noise injection can take many forms, such as top-coding (Adult), i.e., saturation of certain variables, and blurring (Chicago Ridesharing), i.e., disclosure at coarse granularity. Targeted scrubbing of identifiable information is also rather common, with ad-hoc techniques applied in different domains. For example, the curators of ASAP, a dataset featuring student essays, removed personally identifying information from the essays using named entity recognition and several heuristics. Finally, aggregation of data into subpopulations of interest also supports the anonymity of the underlying individuals (FICO).

So far we have covered datasets that feature human data derived from real-world processes. Toy datasets, on the other hand, are perfectly safe from this point of view, however their social relevance is inevitably low. In this work we survey four popular

Table 4 Mitigating factors against re-identification

Mitigating factor	Example datasets
Controlled distribution	
Private dataset	UniGe, Pymetrics Bias Group
Availability upon request	Twitter Online Harrassment
Mixed release strategy	NLSY
Code-based reconstruction	Bias in Bios
Data perturbation	
Obfuscation	Yahoo! c14B Learn to Rank, Microsoft Learning to Rank
Top-coding	Adult
Blurring	Chicago Ridesharing
Targeted scrubbing	ASAP
Aggregation	FICO
Synthesis	
Synthetic data	Toy Dataset 1–4
Semi-synthetic data	Antelope Valley Networks, Kidney Matching
Hypothetical profiles	Italian Car Insurance
Age	German Credit

ones, taken from Zafar et al. (2017c); Donini et al. (2018); Lipton et al. (2018); Singh and Joachims (2019). Semi-synthetic datasets aim for the best of both worlds by generating artificial data from models that emulate the key characteristics of the underlying processes, as is the case with Antelope Valley Networks, Kidney Matching, and the generative adversarial network trained by McDuff et al. (2019) on MS-Celeb-1M. Data synthesis may also be applied to augment datasets with artificial sensitive attributes in a principled fashion [MovieLens—(Burke et al. 2018a)]. Finally, resources designed to externally probe services, algorithms, and platforms, to estimate the direct effect of a feature of interest (e.g. gender, race), may rely on hypothetical profiles (Bertrand and Mullainathan 2004; Fabris et al. 2021). This approach can support evaluations of *fairness through unawareness* (Grgic-Hlaca et al. 2016), of which Italian Car Insurance is an example.

One last important factor is the *age* of a dataset. Re-identification of old information about individuals requires matching with auxiliary resources from the same period, which are less likely to be maintained than comparable resources from recent years. Moreover, even if successful, the consequences of re-identification are likely mitigated by dataset age, as old information about individuals is less likely to support harm against them. The German Credit dataset, for example, represents loan applicants from 1973 to 1975, whose re-identification and subsequent harm appears less likely than re-identification for more recent datasets in the same domain.

Anonymization vs social relevance. Utility and privacy are typically considered conflicting objectives for a dataset (Wieringa et al. 2021). If we define social relevance as the breadth and depth of societally useful insights that can be derived from a dataset, a similar conflict with privacy becomes clear. Old datasets hardly afford any insight that is actionable and relevant to current applications. Insight derived from synthetic datasets is inevitably questionable. Noise injection increases uncertainty and reduces the precision of claims. Obfuscation hinders subsequent annotation of sensitive attributes. Conservative release strategies increase friction and deter from obtaining and analyzing the data. The most socially relevant fairness datasets typically feature confidential information (e.g. criminal history and financial situation) in conjunction with sensitive attributes of individuals (e.g. race and sex). For these reasons, the social impact afforded by a dataset and the safety against re-identification of included individuals are potentially conflicting objectives that require careful balancing. In the next section we discuss informed consent, another important aspect for the privacy of data subjects.

6.2 Consent

Motivation. In the context of data, *informed consent* is an agreement between a data processor and a subject, aimed at allowing collection and use of personal information while guaranteeing some control to the subject. It is emphasized in Article 7 and Recitals (42) and (43) of the General Data Protection Regulation (European Union 2016), requiring it to be freely given, specific, informed, and unambiguous. Paullada et al. (2020) note that in the absence of individual control on personal information, anyone with access to the data can process it with little oversight, possibly against the interest and well-being of data subjects. Consent is thus an important tool in a healthy data ecosystem that favours development, trust, and dignity.

Negative examples. A separate framework, often conflated with consent, is copyright. Licenses such as Creative Commons discipline how academic and creative works can be shared and built upon, with proper credit attribution. According to the Creative Commons organization, however, their licenses are not suited to protect privacy and cover research ethics (Merkley 2019). In computer vision, and especially in face recognition, consent and copyright are often considered and discussed jointly, and Creative Commons licenses are frequently taken as an all-inclusive permit encompassing intellectual property, consent, and ethics (Prabhu and Birhane 2020). Merler et al. (2019), for example, mention privacy and copyright concerns in the construction of Diversity in Faces. These concerns are apparently jointly solved by obtaining images from YFCC-100M, due to the fact that “a large portion of the photos have Creative Commons license”. Indeed lack of consent is a widespread and far-reaching problem in face recognition datasets (Keyes et al. 2019). Prabhu and Birhane (2020) find several examples of non-consensual images in large scale computer vision datasets. A particularly egregious example covered in this survey is MS-Celeb-1M, released in 2016 as the largest publicly available training set for face recognition in the world (Guo et al. 2016b). As suggested by its name, the dataset should feature only celebrities, “to enable our training, testing, and re-distributing under certain licenses” (Guo et al. 2016b). However, the dataset was later found to feature several people who are in no way celebrities, and must simply maintain an online presence. The dataset was retracted for this reason (Murgia 2019).

Positive examples. FACES, an experimental psychology dataset on emotion-related stimuli, represents a positive exception in the face analysis domain. Due its small cardinality, it was possible to obtain informed consent from every participant. One domain where informed consent doctrine has been well-established for decades is medicine; fairness datasets from this space are typically sensitive to the topic. Experiments such as randomized controlled trials always require consent elicitation and often discuss the process in the respective articles. Infant Health and Development Program (IHDP), for instance, is a dataset used to study fair risk assessment. It was collected through the IHDP program, carried out between 1985 and 1988 in the US to evaluate the effectiveness of comprehensive early intervention in reducing developmental and health problems in low birth weight premature infants. Brooks-Gunn et al. (1992) clearly state that “of the 1302 infants who met enrollment criteria, 274 (21%) had parents who refused consent and 43 were withdrawn before entry into the assigned group”. Longitudinal studies require trust and continued participation. They typically produce insights and data thanks to participants who have read and signed an informed consent form. Examples of such datasets include Framingham, stemming from a study on cardiovascular disease, and the National Longitudinal Survey of Youth, following the lives of representative samples of US citizens, focusing on their labor market activities and other significant life events. Field studies and derived datasets (DrugNet, Homeless Youths’ Social Networks) are also attentive to informed consent.

The FRIES framework. According to the Consentful Tech Project,[7] consent should be *Freely given*, *Reversible*, *Informed*, *Enthusiastic*, and *Specific* (FRIES). Below we expand on these points and discuss some fairness datasets through the FRIES lens. Pokec Social Network summarizes the networks of Pokec users, a popular social network in Slovakia and Czech Republic. Due to default privacy settings being predefined as public, a wealth of information for each profile was collected by curators, including information on demographics, politics, education, marital status, and children (Takac and Zabovsky 2012). While privacy settings are a useful tool to control personal data, default public settings are arguably misleading and do not amount to *freely given* consent. In the presence of more conservative predefined settings, a user can explicitly choose to publicly share their information. This may be interpreted as consent to share one’s information here and now with other users; more loose interpretations favouring data collection and distribution are also possible, but they seem rather lacking in *specificity*. It is far from clear that choosing public profile settings entails consent to become part of a study and a publicly available dataset for years to come.

This stands in contrast with Framingham and other datasets derived from medical studies, where consent may be provided or refused with fine granularity (Levy et al. 2010). In this regard, let us consider a consent form from a recent Framingham exam (Framingham Heart Study 2021). The form comes with five different consent boxes which cover participation in examination, use of resulting data, participation in genetic studies, sharing of data with external entities, and notification of findings to subject. Before the consent boxes, a well-structured document informs participants on the

[7] https://www.consentfultech.io/.

reasons for this study, clarifies that they can choose to drop out without penalties at any point, provides a point of contact, explains what will happen in the study and what are the risks to the subject. Some examples of accessible language and open explanations include the following:

- "You have the right to refuse to allow your data and samples to be used or shared for further research. Please check the appropriate box in the selection below."
- "There is a potential risk that your genetic information could be used to your disadvantage. For example, if genetic research findings suggest a serious health problem, that could be used to make it harder for you to get or keep a job or insurance."
- "However, we cannot guarantee total privacy. […] Once information is given to outside parties, we cannot promise that it will be kept private."

Moreover, the consent form is accessible from a website that promises to deliver a Spanish version, showing attention to linguistic minorities. Overall, this approach seems geared towards trust and truly informed consent.

In some cases, consent is made unapplicable by necessity. Allegheny Child Welfare, for instance, stems from an initiative by the Allegheny County's Department of Human Services to develop assistive tools to support child maltreatment hotline screening decisions. Individuals who resort to this service are in a situation of need and emergency that makes *enthusiastic* consent highly unlikely. Similar considerations arise in any situations where data subjects are in a state of need and can only access a service by providing their data. A clear example is Harvey Rescue, the result of crowdsourced efforts to connect rescue parties with people requesting help in the Houston area. Moreover, the provision of data is mandatory in some cases, such as the US census, which conflicts with meaningful, let alone enthusiastic, consent.

Finally, consent should be *reversible*, giving individuals a chance to revoke it and be removed from a dataset. This is an active area of research, studying specific tools for consent management (Albanese et al. 2020) and approaches for retroactive removal of an instance from a model's training set (Ginart et al. 2019). Unfortunately, even when discontinued or redacted, some datasets remain available through backchannels and derivatives. MS-Celeb-1M is, again, a negative example in this regard. The dataset was removed by Microsoft after widespread criticism and claims of privacy infringement. Despite this fact, it remains available via academic torrents (Peng et al. 2021). Moreover, MS-Celeb-1M was used as a source of images for several datasets derived from it, including the BUPT Faces and Racial Faces in the Wild datasets covered in this survey. This fact demonstrates that harms related to data artefacts are not simply remedied via retirement or redaction. Ethical considerations about consent and potential harms to people must be more than an afterthought and need to enter the discussion during design.

6.3 Inclusivity

Motivation. Issues of representation, inclusion and diversity are central to the fair ML community. Due to historical biases stemming from structural inequalities, some populations and their perspectives are underrepresented in certain domains and in related data artefacts (Jo and Gebru 2020). For example, the person subtree of ImageNet contains images that skew toward male, young and light skin individuals (Yang et al. 2020b). Female entities were found to be underrepresented in popular datasets for coreference resolution (Zhao et al. 2018). Even datasets that match natural group proportions may support the development of biased tools with low accuracy for minorities.

Recent works have demonstrated the disparate performance of tools on sensitive subpopulations in domains such as health care (Obermeyer and Mullainathan 2019), speech recognition (Tatman 2017), and computer vision (Buolamwini and Gebru 2018). Inclusivity and diversity are often considered a primary solution in this regard, both in training sets, which support the development of better models, and test sets, capable of flagging such issues.

Positive examples. Ideally, inclusivity should begin with a clear definition of data collection objectives (Jo and Gebru 2020). Indeed, we find that diversity and representation are strong points of datasets that were creted to assess biases in services, products and algorithms (BOLD, HMDA, FICO, Law School, Scientist+Painter, CVs from Singapore, YouTube Dialect Accuracy, Pilot Parliaments Benchmark), which were designed and curated with special attention to sensitive groups. We also find instances of ex-post remedies to issues of diversity. As an example, the curators of ImageNet proposed a demographic balancing solution based on a web interface that removes the images of overrepresented categories (Yang et al. 2020b). A natural alternative is the collection of novel instances, a solution adopted for Framingham. This dataset stems from a study of key factors that contribute to cardiovascular disease, with participants recruited in Framingham, Massachusetts over multiple decades. Recent cohorts were especially designed to reflect a greater racial and ethnic diversity in the town (Tsao and Vasan 2015).

Negative examples. Among the datasets we surveyed, we highlight one whose low inclusivity is rather obvious. WebText is a 40 GB text dataset that supported training of the GPT-2 language model (Radford et al. 2019). The authors crawled every document reachable from outbound Reddit links that collected at least 3 *karma*. While this was considered a useful heuristic to achieve size and quality, it ended up skewing this resource towards content appreciated by Reddit users, who are predominantly male, young, and enjoy good internet access. This should act as reminder that size does not guarantee diversity (Bender et al. 2021), and that sampling biases are almost inevitable.

Inclusivity is nuanced. While inclusivity surely requires an attention to subpopulations, a more precise definition may depend on context and application. Based on the task at hand, an ideal sample may feature all subpopulations with equal presence, or proportionally to their share in the overall population. Let us call these the *equal* and *proportional* approach to diversity. The equal approach is typical of datasets that are meant to be evaluation benchmarks (Pilot Parliaments Benchmark, Winobias) and allow for statistically significant statements on performance differences across groups. On the other hand, the proportional approach is rather common in datasets collected by census offices, such as US Census Data (1990), and in resources aimed precisely at studying issues of representation in services and products (Occupations in Google Images).

Open-ended collection of data is ideal to ensure that various cultures are represented in the manner in which they would like to be seen (Jo and Gebru 2020). Unfortunately, we found no instance of datasets where sensitive labels were self-reported according to open-ended responses. On the contrary, individuals with non-conforming gender identities were excluded from some datasets and analyses. Bing US Queries is a proprietary dataset used to study differential user satisfaction with the Bing search engine across different demographic groups. It consists of a subset of Bing users who provided their gender at registration according to a binary categorization, which misrepresents or simply excludes non-binary users from the subset. Moreover, a dataset may be inclusive and encode gender in a non-binary gender fashion (Climate Assembly UK), but, if used in conjunction with an auxiliary dataset where gender has binary encoding, a common solution is removing instances whose gender is neither female nor male (Flanigan et al. 2020).

Inclusivity does not guarantee benefits. To avoid downstream harms, inclusion by itself is insufficient. The context in which people and sensitive groups are represented should always be taken into account. Despite its overall skew towards male subjects, ImageNet has a high female-to-male ratio in classes such as `bra`, `bikini` and `maillot`, which often feature images that are voyeuristic, pornographic, and non-consensual (Prabhu and Birhane 2020). Similarly, in MS-COCO, a famous dataset for object recognition, there is roughly a 1:3 female-to-male ratio, increasing to 0.95 for images of kitchens (Hendricks et al. 2018). This sort of representation is unlikely to benefit women in any way and, on the contrary, may contribute to reinforce stereotypes and support harmful biases.

Another clear (but often ignored) disconnect between the inclusion of a group and benefits to it is represented by the task at hand and, more in general, by possible uses afforded by a dataset. In this regard, we find many datasets from the face recognition domain, which are presented as resources geared towards inclusion (Diversity in Faces, BUPT Faces, UTK Face, FairFace, Racial Faces in the Wild). Attention to subpopulations in this context is still called "diversity" (Diversity in Faces, FairFace, Racial Faces in the Wild) or "social awareness" (BUPT Faces), but is driven by business imperatives and goals of robustness for a technology that can very easily be employed for surveillance purposes, and become detrimental to vulnerable populations included in datasets. In a similar vein, the FACES dataset has been used to measure age bias in emotion detection, a task whose applications and benefits for individuals remain dubious.

Overall, attention to subpopulations is an upside of many datasets we surveyed. However, inclusion, representation, and diversity can be defined in different ways according to the problem at hand. Individuals would rather be included on their own terms, and decide whether and how they should be represented. The problems of diversity and robustness have some clear commonalities, as the former can be seen as a means towards the latter, but it seems advisable to maintain a clear separation between the two, and to avoid equating either one with fairness. Algorithmic fairness will not be "solved" by simply collecting more data, or granting equal performance across different groups identified by a given sensitive attribute.

Table 5 Approaches to demographic data procurement

Approach	Example datasets
Self-reported labels	Bing US Queries, MovieLens, Libimset, Adult, HMDA, Law School, Sushi, Willingness-to-Pay for Vaccine
Expert labels	Pilot parliaments benchmark
Non-expert labels	CelebFaces attributes, diversity in faces, fairface, occupations in Google images
ML algorithm	Racial faces in the wild, instagram photos, BUPT faces, UTK face
ML algorithm + annotators	FairFace, Open Images Dataset
Rule-/knowledge-based algorithm	RtGender, Bias in Bios, Demographics on Twitter, TwitterAAE

6.4 Sensitive attribute labelling

Motivation. Datasets are often taken as factual information that supports objective computation and pattern extraction. The etymology of the word "data", meaning "given", is rather revealing in this sense. On the contrary, research in human–computer interaction, computer-supported cooperative work, and critical data studies argues that this belief is superficial, limited and potentially harmful (Muller et al. 2019; Crawford and Paglen 2021).

Data is, quite simply, a human-influenced entity (Miceli et al. 2021), determined by a chain of discretionary decisions on measurement, sampling and categorization, which shape how and by whom data will be collected and annotated, according to which taxonomy and based on which guidelines. Data science professionals, often more cognizant of the context surrounding data than theoretical researchers, report significant awareness of how curation and annotation choices influence their data and its relation with the underlying phenomena (Muller et al. 2019). In an interview, a senior text classification researcher responsible for ground truth annotation shows consciousness of their own influence on datasets by stating "I am the ground truth." (Muller et al. 2019).

Sensitive attributes, such as race and gender, are no exception in this regard. Inconsistencies in racial annotation are rather common within the same system (Lum et al. 2020) and, even more so, across different systems (Scheuerman et al. 2020; Khan and Fu 2021). External annotation (either human or algorithmic) is essentially based on co-occurrence of specific traits with membership in a group, thus running the risk of encoding and reinforcing stereotypes. Self-reported labels overcome this issue, although they are still based on an imposed taxonomy, unless provided in an open-ended fashion. In this section, we discuss the practices through which sensitive attributes are annotated in datasets used in algorithmic fairness research, which are summarized in Table 5.

Procurement of sensitive attributes. Self-reported labels for sensitive attributes are typical of datasets that represent users of a service, who may report their demographics during registration (Bing US Queries, MovieLens, Libimseti), or that were gathered through surveys (HMDA, Adult, Law School, Sushi, Willingness-to-Pay for Vaccine). These are all resources for which collection of protected attributes was envisioned at design, potentially as an optional step. However, when sensitive attributes are not available, their annotation may be possible through different mechanisms.

A common approach is having sensitive attributes labelled by non-experts, often workers hired on crowdsourcing platforms. CelebFaces Attributes Dataset (CelebA) features images of celebrities from the CelebFaces dataset, augmented with annotations of landmark location and categorical attributes, including gender, skin tone and age, which were annotated by a "professional labeling company" (Liu et al. 2015). In a similar fashion, Diversity in Faces consists of images labeled with gender and age by workers hired through the Figure Eight crowd-sourcing platform, while the creators of FairFace hired workers on Amazon Mechanical Turk to annotate gender, race, and age in a public image dataset. This practice also raises concerns of fair compensation of labour, which are not discussed in this work.

Some creators employ algorithms to obtain sensitive labels. Face datasets curators often resort to the Face++ API (Racial Faces in the Wild, Instagram Photos, BUPT Faces) or other algorithms (UTK Face, FairFace). In essence labeling is classifying, hence measuring and reporting accuracy for this procedure would be in order, but rarely happens. Creators occasionally note that automated labels were validated (FairFace) or substantially enhanced (Open Images Dataset) by human annotators, and very seldom report inter-annotator agreement (Occupations in Google Images).

Other examples of external labels include the geographic origin of candidates in resumes (CVs from Singapore), political leaning of US Twitter profiles (Twitter Political Searches), English dialect of tweets (TwitterAAE), and gender of subjects featured in image search results for professions (Occupations in Google Images). Annotation may also rely on external knowledge bases such as Wikipedia,[8] as is the case with RtGender. In situations where text written by individuals is available, rule-based approaches exploiting gendered nouns ("woman") or pronouns ("she") are also applicable (Bias in Bios, Demographics on Twitter).

Some datasets may simply have no sensitive attribute. These are often used in works of individual fairness, but may occasionally support studies of group fairness. For example, dSprites is a synthetic computer vision dataset where regular covariates may play the role of sensitive variables (Locatello et al. 2019). Alternatively, datasets can be augmented with simulated demographics, as done by Madnani et al. (2017) who randomly assigned a native language to test-takers in ASAP, or through the technique of Burke et al. (2018a), which they demonstrate on MovieLens.

Face datasets. Posterior annotation is especially common in computer vision datasets. The Pilot Parliaments Benchmark, for instance, was devised as a testbed for face analysis algorithms. It consists of images of parliamentary representatives from three African and three European countries, that were labelled by a surgical dermatologist with the Fitzpatrick skin type of the subjects (Fitzpatrick 1988). This is a dermatological scale for skin color, which can be retrieved from people's appearance. On the contrary, annotations of race or ethnicity from a photo are simplistic at best, and it should be clear that they actually capture *perceived race* from the perspective

[8] https://en.wikipedia.org/wiki/Category:American_female_tennis_players.

of the annotator (FairFace, BUPT Faces). Careful nomenclature is an important first step to improve the transparency of a dataset and make the underlying context more visible.[9]

Similarly to Scheuerman et al. (2020), we find that documentation accompanying face recognition datasets hardly ever describes how specific taxonomies for gender and race were chosen, conveying a false impression of objectivity. A description of the annotation process is typically present, but minimal. For Multi-task Facial Landmark, for instance, we only know that "The ground truths of the related tasks are labeled manually" (Zhang et al. 2014).

Annotation trade-offs. It is worth re-emphasizing that sensitive label assignment is a classification task that rests on assumptions. Annotation of race and gender in images, for example, is based on the idea that they can be accurately ascertained from pictures, which is an oversimplification of these constructs. The envisioned classes (e.g. binary gender) are another subjective choice stemming from the point of view of dataset curators and may reflect narrow or outdated conceptions and potentially harm the data subjects. In this regard a quote from the curators of MS-Celeb-1M, who do not annotate race, but consider it for their sampling strategy, is particularly striking: "We cover all the major races in the world (Caucasian, Mongoloid, and Negroid)" (Guo et al. 2016b). For these reasons, external annotation of sensitive attributes is controversial and inevitably influenced by dataset curators.

On the other hand, external annotation may be the only way to test specific biases. Occupations in Google Images, for instance, is an image dataset collected to study gender and skin tone diversity in image search results for various professions. The creators hired workers on Amazon Mechanical Turk to label the gender (male, female) and Fitzpatrick skin tone (Type 1–6) of the primary person in each image. The Pilot Parliaments Benchmark was also annotated externally to obtain a benchmark for the evaluation of face analysis technology, with a balanced representation of gender and skin type. Different purposes can motivate data collection and annotation of sensitive attributes. Purposes and aims should be documented clearly, while also reflecting on other uses and potential for misuse of a dataset (Gebru et al. 2018). Dataset curators may use documentation to discuss these aspects and specify limitations for the intended use of a resource (Peng et al. 2021). In the next section we focus on documentation and why it represents a key component of data curation.

6.5 Transparency

Motivation. Transparent and accurate documentation is a fundamental part of data quality. Its absence may lead to serious issues, including lack of reproducibility, concerns of scientific validity, ethical problems, and harms (Barocas et al. 2019). Clear documentation can shine a light on inevitable choices made by dataset creators and on the context surrounding the data. In the absence of this information, the curation mechanism mediating reality and data is hidden; the data becomes one with its context,

[9] In this article, we typically discuss sensitive attributes following the naming convention in the accompanying documentation of a dataset, avoiding a critical terminology discussion .

to the point that interpretation of numerical results can be misleading and overarching (Bao et al. 2021).

The "ground truth" labels (typically indicated with the letter y), which are the target of prediction tasks in some datasets, such as indications of recidivism in COMPAS, are especially sensitive in this regard. Indeed, not only accuracy and related quality metrics, but also measures of algorithmic fairness such as sufficiency and separation (Barocas et al. 2019) are based on y labels and the ability of ML algorithms to replicate them, implicitly granting them a special status of truthfulness. In reality, however, these labels may be biased and incorrect due to multiple causes, including, very frequently, a disconnect between what we aim to measure in an ideal construct space (e.g., offense in the case of COMPAS) and what we can actually measure in the observed space (e.g., arrest) (Friedler et al. 2021). Fair ML algorithms (measures) can only partly overcome (catch) these biases, and actually run the risk of further reifying them. Proper documentation does not solve this issue, but equips practitioners and researchers with the necessary awareness to handle these biases.

More broadly, good documentation should discuss and explain features, providing context about who collected and annotated the data, how, and for which purpose (Gebru et al. 2018; Denton et al. 2020). This provides dataset users with information they can leverage to select appropriate datasets for their tasks and avoid unintentional misuse (Gebru et al. 2018). Other actors, such as reviewers, may also access the official documentation of a dataset to ensure that it is employed in compliance with its stated purpose, guidelines, and terms of use (Peng et al. 2021).

Positive examples. In this survey, we find examples of excellent documentation in datasets related to studies and experiments, including CheXpert, Framingham and NLSY. Indeed, datasets curated by medical institutions and census offices are often well-documented. The ideal source of good documentation are descriptor articles published in conjunction with a dataset (e.g. MIMIC-III), typically offering stronger guarantees than web pages in terms of quality and permanence. Official websites hosting and distributing datasets are also important to collect updates, errata, and additional information that may not be available at the time of release. The Million Song Dataset and Goodreads Reviews, for instance, are available on websites which contain a useful overview of the respective dataset, a list of updates, code samples, pointers to documentation, and contacts for further questions.

Negative examples. On the other hand, some datasets are opaque and poorly documented. Among publicly available ones, Arrhythmia is distributed with a description of the features but no context about the purposes, actors, and subjects involved in the data collection. Similarly, the whole curation process and composition of Multi-task Facial Landmark is described in a short paragraph, explaining it consists of 10,000 outdoor face images from the web that were labelled manually with gender. Most face datasets suffer from opaque documentation, especially concerning the choice of sensitive labels and their annotation. For semi-synthetic resources, proper documentation is especially important, to let users understand the broader applicability and implications of numerical analyses performed on a dataset. IBM HR Analytics is a resource about employee attrition, which the hosting website describes as containing "fictional data", without any additional information. Nonetheless, this data was plausibly generated in a principled fashion and (even partial) disclosure of the underlying data generation mechanism would benefit dataset users.

Retrospective documentation. Good documentation may also be produced retrospectively (Bandy and Vincent 2021; Garbin et al. 2021). German Credit is an interesting example of a dataset that was poorly documented for decades, until the recent publication of a report correcting severe coding mistakes (Grömping 2019). As mentioned in Sect. 4.3, from the old documentation it seemed possible to retrieve the sex of data subjects from a feature jointly encoding sex and marital status. The *dataset archaeology* work by Grömping (2019) shows that this is not the case, which has particular relevance for many algorithmic fairness works using this dataset with sex as a protected feature, as this feature is simply not available. Numerical results obtained in this setting may be an artefact of the wrong coding with which the dataset has been, and still is, officially distributed in the UCI Machine Learning Repository (1994). Until the report and the new redacted dataset (UCI Machine Learning Repository 2019) become well-known, the old version will remain prevalent and more mistakes will be made. In other words, while the documentation debt for this particular dataset has been retrospectively addressed (*opacity*), many algorithmic fairness works published after the report continue to use the German Credit dataset with sex as a protected attribute (He et al. 2020b; Yang et al. 2020a; Baharlouei et al. 2020; Lohaus et al. 2020; Martinez et al. 2020; Wang et al. 2021). This is an issue of documentation *sparsity*, where the right information exists but does not reach interested parties, including researchers and reviewers.

Documentation is a fundamental part of data curation, with most responsibility resting on creators. However, dataset users can also play a role in mitigating the documentation debt by proactively looking for information about the resources they plan to use. Brief summaries discussing and motivating the chosen datasets can be included in scholarly articles, at least in supplementary materials when conflicting with page limitations. Indeed, documentation debt is a problem for the whole research community, which can be addressed collectively with retrospective contributions and clarifications. We argue that it is also up to individual researchers to seek contextual information for situating the data they want to use.

7 Broader relevance to the community

Along with the analyses presented in this work, through the lens of tasks supported, domains spanned, and roles played by algorithmic fairness datasets, we are releasing the underlying data briefs, as a further contribution for the research community. Data briefs are a short documentation format providing essential information on datasets used in fairness research. Data briefs are composed of ten fields, detailed in appendix A, derived from shared vocabularies such as Data Catalog Vocabulary (DCAT); to be compliant with the FAIR data principles (Wilkinson et al. 2016), we also defined a schema called `fdo` to model the relationships between the terms and to make the links to external vocabularies explicit. We leverage `fdo` to format the data briefs as a Resource Description Framework (RDF) (Miller 1998), and to make them available as linked open data, thus supporting data reuse, interoperability, and interpretability.[10] Our final aim is to release, update, and maintain a web app, which can be queried by researchers and practitioners to find the most relevant datasets, according to their specific needs.[11] We envision several benefits for the algorithmic fairness and data studies communities, such as:

- Informing the choice of datasets for experimental evaluations of fair ML methods, including domain-oriented and task-oriented search.
- Directing studies of data bias, and other quantitative and qualitative analyses, including retrospective documentation efforts, towards popular (or otherwise important) resources.
- Identifying areas and sub-problems that are understudied in the algorithmic fairness literature.
- Supporting multi-dataset studies, focused on resources united by a common characteristic, such as encoding a given sensitive attribute (Scheuerman et al. 2020), concerning computer vision (Fabbrizzi et al. 2021), or being popular in the fairness literature (Le Quy et al. 2022).

8 Conclusions and recommendations

Algorithmic fairness is a young research area, undergoing a fast expansion, with diverse contributions in terms of methodology and applications. Progress in the field hinges on different resources, including, very prominently, datasets. In this work, extending (Fabris et al. 2022), we have surveyed hundreds of datasets used in the fair ML and algorithmic equity literature to help the research community reduce its documentation debt, improve the utilization of existing datasets, and the curation of novel ones.

With respect to existing resources, we have shown that the most popular datasets in the fairness literature (Adult, COMPAS, and German Credit) have limited merits beyond originating from human processes and encoding protected attributes. On the other hand, several negative aspects call into question their current status of general-purpose fairness benchmarks, including contrived prediction tasks, noisy data, severe coding mistakes, limitations in encoding sensitive attributes, and age.

We have documented over two hundred datasets to provide viable alternatives, annotating their domain, the tasks they support, and discussing the roles they play in works of algorithmic fairness. We have shown that the processes generating the data belong to many different domains, including, for instance, criminal justice, education, search engines, online marketplaces, emergency response, social media, medicine, hiring, and finance. At the same time, we have described a variety of tasks studied on these resources, ranging from generic, such as *fair classification*, to narrow such as *fair districting* and *fair truth discovery*. Overall, such diversity of domains and tasks provides a glimpse into the variety of human activities and applications that can be impacted by automated decision making, and that can benefit from algorithmic

[10] Schema publicly available at https://fairnessdatasets.dei.unipd.it/schema/; RDF publicly available at https://zenodo.org/record/6518370#.YnOSKFTMJhF.

[11] This resource will be released at https://fairnessdatasets.dei.unipd.it/.

fairness research. Tasks and domain annotations are made available in our data briefs to facilitate the work of researchers and practitioners interested in the study of algorithmic fairness applied to specific domains or tasks. By assembling sparse information on hundreds of datasets into a single document, we aim to provide a useful reference to support both domain-oriented and task-oriented dataset search.

At the same time, we have analyzed issues connected to re-identification, consent, inclusivity, labeling, and transparency running across these datasets. By describing a range of approaches and attentiveness to these topics, we aim to make them more visible and concrete. On one hand, this may prove valuable to inform post-hoc data interventions aimed at mitigating potential harms caused by existing datasets. On the other hand, as novel datasets are increasingly curated, published, and adopted in fairness research, it is important to motivate these concerns, make them tangible, and distill existing approaches into best practices, which we summarize below, for future endeavours of data curation. Our recommendations complement (and do not replace) a growing body of work studying key aspects in the life cycle of datasets (Gebru et al. 2018; Jo and Gebru 2020; Prabhu and Birhane 2020; Crawford and Paglen 2021; Peng et al. 2021).

Social relevance of data, intended as the breadth and depth of societally useful insights afforded by datasets, is a central requirement in fairness research. Unfortunately, this may conflict with user privacy, favouring re-identification or leaving consideration of consent in the background. Consent should be considered during the initial design of a dataset, in accordance with existing frameworks, such as the FRIES framework outlined in the Consentful Tech project. Moreover, different strategies are available to alleviate concerns of re-identification, including noise injection, conservative release, and (semi)synthetic data generation. Algorithmic fairness is motivated by aims of justice and harm avoidance for people, which should be extended to data subjects.

Inclusivity is also important for social relevance, as it allows for a wider representation, and supports analyses that take into account important groups. However, inclusivity is insufficient in itself. Possible uses afforded by a dataset should always be considered, evaluating costs and benefits for the data subjects and the wider population. In the absence of these considerations, acritical inclusivity runs the risk of simply supporting system robustness across sensitive attributes, such as race and gender, rebranded as fairness.

Sensitive attributes are a key ingredient to measure inclusion and increase the social relevance of a dataset. Although often impractical, it is typically preferable for sensitive attributes to be self-reported by data subjects. Externally assigned labels and taxonomies can harm individuals by erasing their needs and points of view. Sensitive attribute labelling is thus a shortcut whose advantages and disadvantages should be carefully weighted and, if chosen, it should be properly documented. Possible approaches based on human labour include expert and non-expert annotation, while automated approaches range from simple rule-based systems to complex and opaque algorithms. To label is to classify, hence measuring and reporting per-group accuracy is in order. Some labeling endeavours are more sensible than others: while skin tone can arguably be retrieved from pictures, annotations of race from an image actually capture *perceived race* from the perspective of the annotator. Rigorous nomenclature favours better understanding and clarifies the subjectivity of certain labels.

Reliable documentation shines a light on inevitable choices made by dataset creators and on the context surrounding the data. This provides dataset users with information they can leverage to select appropriate datasets for their tasks and avoid unintentional misuse. Datasets for which some curation choices are poorly documented may appear more objective at first sight. However, it should be clear that objective data and turbid data are very different things. Proper documentation increases transparency, trust, and understanding. At a minimum, it should include the purpose of a data artifact, a description of the sample, the features and related annotation procedures, along with an explicit discussion of the associated task, if any. It should also clarify who was involved in the different stages of the data development procedure, with special attention to annotation. Data documentation also supports reviewers and readers of academic research in assessing whether a dataset was selected with good reason and utilized in compliance with creators' guidelines.

Understanding and budgeting for all these aspects during early design phases, rather than after collection or release, can be invaluable for data subjects, data users, and society. While possible remedies exist, data is an extremely fluid asset allowing for easy reproduction and derivatives of all sorts; remedies applied to a dataset do not necessarily benefit its derivates. In this work, we have targeted the collective documentation debt of the algorithmic fairness community, resulting from the opacity surrounding certain resources and the sparsity of existing documentation. We have mainly targeted sparsity in a centralized documentation effort; as a result, we have found and described a range of weaknesses and best practices that can be adopted to reduce opacity and mitigate concerns of privacy and inclusion. Similarly to other types of data interventions, useful documentation can be produced after release, but, as shown in this work, the documentation debt may propagate nonetheless. In a mature research community, curators, users, and reviewers can all contribute to cultivating a data documentation culture and keep the overall documentation debt in check.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10618-022-00854-z.

Acknowledgements The authors would like to thank the following researchers and dataset creators for the useful feedback on the data briefs: Alain Barrat, Luc Behaghel, Asia Biega, Marko Bohanec, Chris Burgess, Robin Burke, Alejandro Noriega Campero, Margarida Carvalho, Abhijnan Chakraborty, Robert Cheetham, Won Ik Cho, Paulo Cortez, Thomas Davidson, Maria De-Arteaga, Lucas Dixon, Danijela Djordjević, Michele Donini, Marco Duarte, Natalie Ebner, Elaine Fehrman, H. Altay Guvenir, Moritz Hardt, Irina Higgins, Yu Hen Hu, Rachel Huddart, Lalana Kagal, Dean Karlan, Vijay Keswani, Been Kim, Hyunjik Kim, Jiwon Kim, Svetlana Kiritchenko, Pang Wei Koh, Joseph A. Konstan, Varun Kumar, Jeremy Andrew Irvin, Jamie N. Larson, Jure Leskovec, Jonathan Levy, Andrea Lodi, Oisin Mac Aodha, Loic Matthey, Julian McAuley, Brendan McMahan, Sergio Moro, Luca Oneto, Orestis Papakyriakopoulos, Stephen Robert Pfohl, Christopher G. Potts, Mike Redmond, Kit Rodolfa, Ben Roshan, Veronica Rotemberg, Rachel Rudinger, Sivan Sabato, Kate Saenko, Mark D. Shermis, Daniel Slunge, David Solans, Luca Soldaini, Efstathios Stamatatos, Ryan Steed, Rachael Tatman, Schrasing Tong, Alan Tsang, Sathishkumar V E, Andreas van Cranenburgh, Lucy Vasserman, Roland Vollgraf, Alex Wang, Zeerak Waseem, Kellie Webster, Bryan Wilder, Nick Wilson, I-Cheng Yeh, Elad Yom-Tov, Neil Yorke-Smith, Michal Zabovsky, Yukun Zhu.

Funding Open access funding provided by Università degli Studi di Padova within the CRUI-CARE Agreement.

References

Abbasi M, Bhaskara A, Venkatasubramanian S (2021) Fair clustering via equitable group representations. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, association for computing machinery, New York, FAccT '21, pp 504–514. https://doi.org/10.1145/3442188.3445913

Adragna R, Creager E, Madras D, Zemel R (2020) Fairness and robustness in invariant learning: a case study in toxicity classification. NeurIPS 2020 workshop: "algorithmic fairness through the lens of causality and interpretability (AFCI)". arXiv:2011.06485

Agarwal A, Beygelzimer A, Dudik M, Langford J, Wallach H (2018a) A reductions approach to fair classification. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of machine learning research, vol 80, pp 60–69. http://proceedings.mlr.press/v80/agarwal18a.html

Agrawal M, Zitnik M, Leskovec J, et al. (2018b) Large-scale analysis of disease pathways in the human interactome. In: PSB, World Scientific, pp 111–122

Agarwal A, Dudik M, Wu ZS (2019) Fair regression: quantitative definitions and reduction-based algorithms. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, California, USA, Proceedings of machine learning research, vol 97, pp 120–129. http://proceedings.mlr.press/v97/agarwal19d.html

Ahmadian S, Epasto A, Knittel M, Kumar R, Mahdian M, Moseley B, Pham P, Vassilvitskii S, Wang Y (2020) Fair hierarchical clustering. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, NeurIPS 2020, December 6–12, 2020, virtual. https://proceedings.neurips.cc/paper/2020/hash/f10f2da9a238b746d2bac55759915f0d-Abstract.html

Aka O, Burke K, Bauerle A, Greer C, Mitchell M (2021) Measuring model biases in the absence of ground truth. Association for Computing Machinery, New York, pp 327–335. https://doi.org/10.1145/3461702.3462557

Albanese G, Calbimonte JP, Schumacher M, Calvaresi D (2020) Dynamic consent management for clinical trials via private blockchain technology. J Amb Intell Human Comput 1–18

Ali J, Babaei M, Chakraborty A, Mirzasoleiman B, Gummadi KP, Singla A (2019a) On the fairness of time-critical influence maximization in social networks. NeurIPS 2019 workshop: "Human-Centric Machine Learning". arXiv:1905.06618

Ali J, Zafar MB, Singla A, Gummadi KP (2019b) Loss-aversively fair classification. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, AIES '19, pp 211–218. https://doi.org/10.1145/3306618.3314266,

Ali J, Lahoti P, Gummadi KP (2021) Accounting for model uncertainty in algorithmic discrimination. Association for Computing Machinery, New York, pp 336–345. https://doi.org/10.1145/3461702.3462630

Amini A, Soleimany AP, Schwarting W, Bhatia SN, Rus D (2019) Uncovering and mitigating algorithmic bias through learned latent structure. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, AIES '19, pp 289–295. https://doi.org/10.1145/3306618.3314243,

Anderson E (1936) The species problem in iris. Ann Mo Bot Gard 23(3):457–509

Andrus M, Spitzer E, Brown J, Xiang A (2021) What we can't measure, we can't understand: challenges to demographic data procurement in the pursuit of fairness. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAccT '21, pp 249–260. https://doi.org/10.1145/3442188.3445888

Andrzejak RG, Lehnertz K, Mormann F, Rieke C, David P, Elger CE (2001) Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: dependence on recording region and brain state. Phys Rev E 64(6):061907

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D (2020) Invariant risk minimization. arXiv:1907.02893

Atwood J, Srinivasan H, Halpern Y, Sculley D (2019) Fair treatment allocations in social networks. NeurIPS 2019 workshop: "Fair ML for Health". arXiv:1911.05489

Awasthi P, Beutel A, Kleindessner M, Morgenstern J, Wang X (2021) Evaluating fairness of machine learning models under uncertain and incomplete information. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAccT '21, pp 206–214. https://doi.org/10.1145/3442188.3445884

Babaeianjelodar M, Lorenz S, Gordon J, Matthews J, Freitag E (2020) Quantifying gender bias in different corpora. In: Companion proceedings of the web conference 2020. Association for Computing Machinery, New York, WWW '20, pp 752–759. https://doi.org/10.1145/3366424.3383559

Babaioff M, Nisan N, Talgam-Cohen I (2019) Fair allocation through competitive equilibrium from generic incomes. In: Proceedings of the conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '19, pp 180. https://doi.org/10.1145/3287560.3287582

Backurs A, Indyk P, Onak K, Schieber B, Vakilian A, Wagner T (2019) Scalable fair clustering. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, California, proceedings of machine learning research, vol 97, pp 405–413. http://proceedings.mlr.press/v97/backurs19a.html

Bagdasaryan E, Poursaeed O, Shmatikov V (2019) Differential privacy has disparate impact on model accuracy. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems. Curran Associates, Inc., vol 32. https://proceedings.neurips.cc/paper/2019/file/fc0de4e0396fff257ea362983c2dda5a-Paper.pdf

Baharlouei S, Nouiehed M, Beirami A, Razaviyayn M (2020) Rényi fair inference. In: International conference on learning representations. https://openreview.net/forum?id=HkgsUJrtDB

Bakker MA, Tu DP, Valdés HR, Gummadi KP, Varshney KR, Weller A, Pentland A (2019) Dadi: Dynamic discovery of fair information with adversarial reinforcement learning. NeurIPS 2019 workshop: "Human-centric machine learning". arXiv:1910.13983

Bakker MA, Tu DP, Gummadi KP, Pentland AS, Varshney KR, Weller A (2021) Beyond reasonable doubt: improving fairness in budget-constrained decision making using confidence thresholds. Association for Computing Machinery, New York, pp 346–356. https://doi.org/10.1145/3461702.3462575

Ball-Burack A, Lee MSA, Cobbe J, Singh J (2021) Differential tweetment: mitigating racial dialect bias in harmful tweet detection. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAccT '21, pp 116–128. https://doi.org/10.1145/3442188.3445875

Bandy J, Vincent N (2021) Addressing" documentation debt" in machine learning research: A retrospective datasheet for bookcorpus. arXiv:2105.05241

Bao M, Zhou A, Zottola S, Brubach B, Desmarais S, Horowitz A, Lum K, Venkatasubramanian S (2021) It's compaslicated: the messy relationship between rai datasets and algorithmic fairness benchmarks. arXiv:2106.05498

Barabas C, Dinakar K, Doyle C (2019) The problems with risk assessment tools. https://www.nytimes.com/2019/07/17/opinion/pretrial-ai.html

Barbaro M (2007) In apparel, all tariffs aren't created equal. https://www.nytimes.com/2007/04/28/business/28gender.html

Barenstein M (2019) Propublica's compas data revisited. arXiv:1906.04711

Barman-Adhikari A, Begun S, Rice E, Yoshioka-Maxwell A, Perez-Portillo A (2016) Sociometric network structure and its association with methamphetamine use norms among homeless youth. Soc Sci Res 58:292–308

Barocas S, Hardt M, Narayanan A (2019) Fairness and machine learning. fairmlbook.org. http://www.fairmlbook.org

Baudry JP, Cardoso M, Celeux G, Amorim MJ, Ferreira AS (2015) Enhancing the selection of a model-based clustering with external categorical variables. Adv Data Anal Classif 9(2):177–196

Behaghel L, Crépon B, Gurgand M (2014) Private and public provision of counseling to job seekers: evidence from a large controlled experiment. Am Econ J Appl Econ 6(4):142–74. https://doi.org/10.1257/app.6.4.142

Belitz C, Jiang L, Bosch N (2021) Automating procedurally fair feature selection in machine learning. Association for Computing Machinery, New York, pp 379–389. https://doi.org/10.1145/3461702.3462585

Bender EM, Friedman B (2018) Data statements for natural language processing: toward mitigating system bias and enabling better science. Trans Assoc Comput Linguist 6:587–604. https://doi.org/10.1162/tacl_a_00041, https://www.aclweb.org/anthology/Q18-1041

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: can language models be too big? Association for Computing Machinery, New York, FAccT '21, pp 610–623. https://doi.org/10.1145/3442188.3445922,

Benenson R, Popov S, Ferrari V (2019) Large-scale interactive object segmentation with human annotators. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11700–11709

Bera S, Chakrabarty D, Flores N, Negahbani M (2019) Fair algorithms for clustering. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems. Curran Associates, Inc., vol 32, pp 4954–4965. https://proceedings.neurips.cc/paper/2019/file/fc192b0c0d270dbf41870a63a8c76c2f-Paper.pdf

Beretta E, Vetrò A, Lepri B, Martin JCD (2021) Detecting discriminatory risk through data annotation based on Bayesian inferences. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAccT '21, pp 794-804. https://doi.org/10.1145/3442188.3445940

Berk R, Heidari H, Jabbari S, Joseph M, Kearns M, Morgenstern J, Neel S, Roth A (2017) A convex framework for fair regression. In: KDD 2017 workshop: fairness, accountability, and transparency in machine learning (FAT/ML). arXiv:1706.02409

Bertin-Mahieux T, Ellis DPW, Whitman B, Lamere P (2011) The million song dataset. In: Proceedings of the 12th international society for music information retrieval conference, ISMIR, Miami, pp 591–596. https://doi.org/10.5281/zenodo.1415820

Bertrand M, Mullainathan S (2004) Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. Am Econ Rev 94(4):991–1013

Beutel A, Chen J, Zhao Z, Chi EH (2017) Data decisions and theoretical implications when adversarially learning fair representations. In: KDD 2017 workshop: "fairness, accountability, and transparency in machine learning (FAT/ML)". arXiv:1707.00075

Biega AJ, Diaz F, Ekstrand MD, Kohlmeier S (2019) Overview of the trec 2019 fair ranking track. In: The twenty-eighth text REtrieval conference (TREC 2019) proceedings

Biswas A, Mukherjee S (2021) Ensuring fairness under prior probability shifts. Association for Computing Machinery, New York, pp 414–424. https://doi.org/10.1145/3461702.3462596

Black E, Fredrikson M (2021) Leave-one-out unfairness. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAccT '21, pp 285–295. https://doi.org/10.1145/3442188.3445894

Black E, Yeom S, Fredrikson M (2020) Fliptest: fairness testing via optimal transport. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '20, pp 111–121. https://doi.org/10.1145/3351095.3372845

Blodget SL, O'Connor B (2017) Racial disparity in natural language processing: a case study of social media african-american english. KDD 2017 workshop: "fairness, accountability, and transparency in machine learning (FAT/ML)". arXiv:1707.00061

Blodgett SL, Green L, O'Connor B (2016) Demographic dialectal variation in social media: a case study of African-American English. In: Proceedings of the 2016 conference on empirical methods in natural language processing. Association for Computational Linguistics, Austin, pp 1119–1130. https://doi.org/10.18653/v1/D16-1120, https://www.aclweb.org/anthology/D16-1120

Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT (2016) Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R (eds) Advances in neural information processing systems.

Curran Associates, Inc., vol 29, pp 4349–4357. https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O (2013) Translating embeddings for modeling multi-relational data. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds) Advances in neural information processing systems. Curran Associates, Inc., vol 26. https://proceedings.neurips.cc/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf

Bordia S, Bowman SR (2019) Identifying and reducing gender bias in word-level language models. In: Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: student research workshop. Association for Computational Linguistics, Minneapolis, pp 7–15. https://doi.org/10.18653/v1/N19-3002, https://aclanthology.org/N19-3002

Borkan D, Dixon L, Sorensen J, Thain N, Vasserman L (2019) Nuanced metrics for measuring unintended bias with real data for text classification. In: Companion proceedings of The 2019 world wide web conference. Association for Computing Machinery, New York, WWW '19, pp 491–500. https://doi.org/10.1145/3308560.3317593

Bose A, Hamilton W (2019) Compositional fairness constraints for graph embeddings. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, proceedings of machine learning research, vol 97, pp 715–724. http://proceedings.mlr.press/v97/bose19a.html

Bower A, Niss L, Sun Y, Vargo A (2018) Debiasing representations by removing unwanted variation due to protected attributes. In: ICML 2018 workshop: "fairness, accountability, and transparency in machine learning (FAT/ML)". arXiv:1807.00461

Bower A, Eftekhari H, Yurochkin M, Sun Y (2021) Individually fair rankings. In: International conference on learning representations. https://openreview.net/forum?id=71zCSP_HuBN

Brennan T, Dieterich W, Ehret B (2009) Evaluating the predictive validity of the compas risk and needs assessment system. Crim Justice Behav 36(1):21–40. https://doi.org/10.1177/0093854808326545

Brockman G, Cheung V, Pettersson L, Schneider J, Schulman J, Tang J, Zaremba W (2016) Openai gym. arXiv:1606.01540

Brooks-Gunn J, Fr L, Klebanov PK (1992) Effects of early intervention on cognitive function of low birth weight preterm infants. J Pediatr 120(3):350–359

Brožovský L (2006) Recommender system for a dating service. Master's thesis, Charles University in Prague, Prague. http://colfi.wz.cz/colfi.pdf

Brozovsky L, Petricek V (2007) Recommender system for online dating service. arXiv:0703042 [cs]

Brubach B, Chakrabarti D, Dickerson J, Khuller S, Srinivasan A, Tsepenekas L (2020) A pairwise fair and community-preserving approach to k-center clustering. In: III HD, Singh A (eds) Proceedings of the 37th international conference on machine learning, PMLR, virtual, proceedings of machine learning research, vol 119, pp 1178–1189. http://proceedings.mlr.press/v119/brubach20a.html

Brunet ME, Alkalay-Houlihan C, Anderson A, Zemel R (2019) Understanding the origins of bias in word embeddings. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, Proceedings of machine learning research, vol 97, pp 803–811. http://proceedings.mlr.press/v97/brunet19a.html

Buolamwini J, Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Friedler SA, Wilson C (eds) Proceedings of the 1st conference on fairness, accountability and transparency, PMLR, New York, proceedings of machine learning research, vol 81, pp 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

Burke R, Kontny J, Sonboli N (2018a) Synthetic attribute data for evaluating consumer-side fairness. RecSys 2018 workshop: "workshop on responsible recommendation (FAT/Rec)". arXiv:1809.04199

Burke R, Sonboli N, Ordonez-Gauger A (2018b) Balanced neighborhoods for multi-sided fairness in recommendation. In: Friedler SA, Wilson C (eds) Proceedings of the 1st conference on fairness, accountability and transparency, PMLR, New York, proceedings of machine learning research, vol 81, pp 202–214. http://proceedings.mlr.press/v81/burke18a.html

Buyl M, De Bie T (2020) DeBayes: a Bayesian method for debiasing network embeddings. In: III HD, Singh A (eds) Proceedings of the 37th international conference on machine learning, PMLR, virtual, proceedings of machine learning research, vol 119, pp 1220–1229. http://proceedings.mlr.press/v119/buyl20a.html

Cai W, Gaebler J, Garg N, Goel S (2020) Fair allocation through selective information acquisition. In: Proceedings of the AAAI/ACM conference on AI, ethics, and society, association for computing machinery, New York, AIES '20, pp 22–28. https://doi.org/10.1145/3375627.3375823, Caldas S, Duddu SMK, Wu P, Li T, Konečnỳ J, McMahan HB, Smith V, Talwalkar A (2018) Leaf: a benchmark for federated settings. arXiv:1812.01097

Calders T, Verwer S (2010) Three Naive Bayes approaches for discrimination-free classification. Data Min Knowl Discov 21(2):277–292. https://doi.org/10.1007/s10618-010-0190-x

Calders T, Kamiran F, Pechenizkiy M (2009) Building classifiers with independency constraints. In: 2009 IEEE international conference on data mining workshops, pp 13–18. https://doi.org/10.1109/ICDMW.2009.83

Caliskan A, Bryson J, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. Science 356(6334):183–186. https://doi.org/10.1126/science.aal4230

Calmon F, Wei D, Vinzamuri B, Natesan Ramamurthy K, Varshney KR (2017) Optimized pre-processing for discrimination prevention. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems. Curran Associates, Inc., vol 30, pp 3992–4001. https://proceedings.neurips.cc/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf

Canetti R, Cohen A, Dikkala N, Ramnarayan G, Scheffler S, Smith A (2019) From soft classifiers to hard decisions: How fair can we be? In: Proceedings of the conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* ’19, pp 309–318. https://doi.org/10.1145/3287560.3287561

Caragiannis I, Kurokawa D, Moulin H, Procaccia AD, Shah N, Wang J (2016) The unreasonable fairness of maximum nash welfare. In: Proceedings of the 2016 ACM conference on economics and computation. Association for Computing Machinery, New York, EC ’16, pp 305–322. https://doi.org/10.1145/2940716.2940726

Cardoso RL, Meira Jr W, Almeida V, Zaki MJ (2019) A framework for benchmarking discrimination-aware models in machine learning. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, AIES ’19, pp 437–444. https://doi.org/10.1145/3306618.3314262

Carvalho M, Lodi A (2019) Game theoretical analysis of kidney exchange programs. arXiv:1911.09207

Caton S, Haas C (2020) Fairness in machine learning: a survey. arXiv:2010.04053

Celis LE, Keswani V (2020) Implicit diversity in image summarization. Proc ACM Hum Comput Interact 4(CSCW2):1–28. https://doi.org/10.1145/3415210

Celis LE, Deshpande A, Kathuria T, Vishnoi NK (2016) How to be fair and diverse? DTL 2016 workshop: “fairness, accountability, and transparency in machine learning (FAT/ML)”. arXiv:1610.07183

Celis E, Keswani V, Straszak D, Deshpande A, Kathuria T, Vishnoi N (2018) Fair and diverse DPP-based data summarization. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Stockholm Sweden, proceedings of machine learning research, vol 80, pp 716–725. http://proceedings.mlr.press/v80/celis18a.html

Celis E, Mehrotra A, Vishnoi N (2019a) Toward controlling discrimination in online ad auctions. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, proceedings of machine learning research, vol 97, pp 4456–4465. http://proceedings.mlr.press/v97/mehrotra19a.html

Celis LE, Huang L, Keswani V, Vishnoi NK (2019b) Classification with fairness constraints: a meta-algorithm with provable guarantees. In: Proceedings of the conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* ’19, pp 319–328. https://doi.org/10.1145/3287560.3287586

Celis LE, Keswani V, Vishnoi N (2020a) Data preprocessing to mitigate bias: A maximum entropy based approach. In: III HD, Singh A (eds) Proceedings of the 37th international conference on machine learning, PMLR, virtual, proceedings of machine learning research, vol 119, pp 1349–1359. http://proceedings.mlr.press/v119/celis20a.html

Celis LE, Mehrotra A, Vishnoi NK (2020b) Interventions for ranking in the presence of implicit bias. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* ’20, pp 369–380. https://doi.org/10.1145/3351095.3372858

Celma O (2010) Music recommendation and discovery in the long tail. Springer, Berlin

Chaibub Neto E (2020) A causal look at statistical definitions of discrimination. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, KDD ’20, pp 873–881. https://doi.org/10.1145/3394486.3403130

Chakraborty A, Patro GK, Ganguly N, Gummadi KP, Loiseau P (2019) Equality of voice: towards fair representation in crowdsourced top-k recommendations. In: Proceedings of the conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '19, pp 129-138. https://doi.org/10.1145/3287560.3287570

Chapelle O, Chang Y (2010) Yahoo! learning to rank challenge overview. In: Proceedings of the 2010 international conference on Yahoo! Learning to rank challenge-Volume 14, JMLR.org, YLRC'10, pp 1–24

Chaudhari HA, Lin S, Linda O (2020) A general framework for fairness in multistakeholder recommendations. RecSys 2020 workshop: "3rd FAccTRec workshop on responsible recommendation". arXiv:2009.02423

Chelba C, Mikolov T, Schuster M, Ge Q, Brants T, Koehn P, Robinson T (2014) One billion word benchmark for measuring progress in statistical language modeling. In: INTERSPEECH-2014

Chen B, Deng W, Shen H (2018a) Virtual class enhanced discriminative embedding learning. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems. Curran Associates, Inc., vol 31. https://proceedings.neurips.cc/paper/2018/file/d79aac075930c83c2f1e369a511148fe-Paper.pdf

Chen CW, Lamere P, Schedl M, Zamani H (2018b) Recsys challenge 2018: Automatic music playlist continuation. Association for Computing Machinery, New York, RecSys '18, pp 527–528. https://doi.org/10.1145/3240323.3240342

Chen I, Johansson FD, Sontag D (2018c) Why is my classifier discriminatory? In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems. Curran Associates, Inc., vol 31. https://proceedings.neurips.cc/paper/2018/file/1f1baa5b8edac74eb4eaa329f14a0361-Paper.pdf

Chen J, Kallus N, Mao X, Svacha G, Udell M (2019a) Fairness under unawareness: assessing disparity when protected class is unobserved. In: Proceedings of the conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '19, pp 339–348. https://doi.org/10.1145/3287560.3287594

Chen X, Fain B, Lyu L, Munagala K (2019b) Proportionally fair clustering. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, proceedings of machine learning research, vol 97, pp 1032–1041. http://proceedings.mlr.press/v97/chen19d.html

Chen Y, Mahoney C, Grasso I, Wali E, Matthews A, Middleton T, Njie M, Matthews J (2021) Gender bias and under-representation in natural language processing across human languages. Association for Computing Machinery, New York, pp 24–34. https://doi.org/10.1145/3461702.3462530

Cheng P, Hao W, Yuan S, Si S, Carin L (2021a) Fairfil: Contrastive neural debiasing method for pretrained text encoders. In: International conference on learning representations. https://openreview.net/forum?id=N6JECD-PI5w

Cheng V, Suriyakumar VM, Dullerud N, Joshi S, Ghassemi M (2021b) Can you fake it until you make it? impacts of differentially private synthetic data on downstream classification fairness. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAccT '21, pp 149–160. https://doi.org/10.1145/3442188.3445879

Chierichetti F, Kumar R, Lattanzi S, Vassilvitskii S (2017) Fair clustering through fairlets. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems. Curran Associates, Inc., vol 30, pp 5029–5037. https://proceedings.neurips.cc/paper/2017/file/978fce5bcc4eccc88ad48ce3914124a2-Paper.pdf

Chiplunkar A, Kale S, Ramamoorthy SN (2020) How to solve fair k-center in massive data models. In: III HD, Singh A (eds) Proceedings of the 37th international conference on machine learning, PMLR, virtual, proceedings of machine learning research, vol 119, pp 1877–1886. http://proceedings.mlr.press/v119/chiplunkar20a.html

Cho J, Hwang G, Suh C (2020) A fair classifier using kernel density estimation. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in neural information processing systems. Curran Associates, Inc., vol 33, pp 15088–15099. https://proceedings.neurips.cc/paper/2020/file/ac3870fcad1cfc367825cda0101eee62-Paper.pdf

Cho WI, Kim J, Yang J, Kim NS (2021) Towards cross-lingual generalization of translation gender bias. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAccT '21, pp 449–457. https://doi.org/10.1145/3442188.3445907

Choi K, Grover A, Singh T, Shu R, Ermon S (2020a) Fair generative modeling via weak supervision. In: III HD, Singh A (eds) Proceedings of the 37th international conference on machine learning, PMLR, virtual, proceedings of machine learning research, vol 119, pp 1887–1898. http://proceedings.mlr.press/v119/choi20a.html

Choi Y, Dang M, den Broeck GV (2020b) Group fairness by probabilistic modeling with latent fair decisions. NeurIPS 2020 workshop: "algorithmic fairness through the lens of causality and interpretability (AFCI)". arXiv:2009.09031

Chouldechova A (2017) Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big data 5(2):153–163

Chouldechova A, G'Sell M (2017) Fairer and more accurate, but for whom? KDD 2017 workshop: "fairness, accountability, and transparency in machine learning (FAT/ML)". arXiv:1707.00046

Chouldechova A, Roth A (2020) A snapshot of the frontiers of fairness in machine learning. Commun ACM 63(5):82–89. https://doi.org/10.1145/3376898

Chuang CY, Mroueh Y (2021) Fair mixup: fairness via interpolation. In: International conference on learning representations. https://openreview.net/forum?id=DNl5s5BXeBn

Chzhen E, Denis C, Hebiri M, Oneto L, Pontil M (2019) Leveraging labeled and unlabeled data for consistent fair binary classification. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems. Curran Associates, Inc., vol 32, pp 12760–12770. https://proceedings.neurips.cc/paper/2019/file/ba51e6158bcaf80fd0d834950251e693-Paper.pdf

Chzhen E, Denis C, Hebiri M, Oneto L, Pontil M (2020a) Fair regression via plug-in estimator and recalibration with statistical guarantees. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, NeurIPS 2020, December 6–12, 2020, virtual. https://proceedings.neurips.cc/paper/2020/hash/ddd808772c035aed16d42ad3559be5f-Abstract.html

Chzhen E, Denis C, Hebiri M, Oneto L, Pontil M (2020b) Fair regression with wasserstein barycenters. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in neural information processing systems. Curran Associates, Inc., vol 33, pp 7321–7331. https://proceedings.neurips.cc/paper/2020/file/51cdbd2611e844ece5d80878eb770436-Paper.pdf

Cohany SR, Polivka AE, Rothgeb JM (1994) Revisions in the current population survey effective january 1994. Emp Earn 41:13

Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A (2017) Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, KDD '17, pp 797–806. https://doi.org/10.1145/3097983.3098095

Cortez P, Silva AMG (2008) Using data mining to predict secondary school student performance. In: Proceedings of 5th FUture BUsiness TEChnology conference

Coston A, Ramamurthy KN, Wei D, Varshney KR, Speakman S, Mustahsan Z, Chakraborty S (2019) Fair transfer learning with missing protected attributes. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, AIES '19, pp 91–98. https://doi.org/10.1145/3306618.3314236

Coston A, Mishler A, Kennedy EH, Chouldechova A (2020) Counterfactual risk assessments, evaluation, and fairness. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '20, pp 582–593. https://doi.org/10.1145/3351095.3372851

Coston A, Guha N, Ouyang D, Lu L, Chouldechova A, Ho DE (2021) Leveraging administrative data for bias audits: assessing disparate coverage with mobility data for covid-19 policy. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAccT '21, pp 173–184. https://doi.org/10.1145/3442188.3445881

Cotter A, Gupta M, Jiang H, Srebro N, Sridharan K, Wang S, Woodworth B, You S (2018) Training fairness-constrained classifiers to generalize. ICML 2018 workshop: "fairness, accountability, and transparency in machine learning (FAT/ML)"

Cotter A, Gupta M, Jiang H, Srebro N, Sridharan K, Wang S, Woodworth B, You S (2019) Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, proceedings of machine learning research, vol 97, pp 1397–1405. http://proceedings.mlr.press/v97/cotter19b.html

Crawford K, Paglen T (2021) Excavating ai: the politics of images in machine learning training sets. https://excavating.ai/

Creager E, Madras D, Jacobsen JH, Weis M, Swersky K, Pitassi T, Zemel R (2019) Flexibly fair representation learning by disentanglement. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, proceedings of machine learning research, vol 97, pp 1436–1445. http://proceedings.mlr.press/v97/creager19a.html

Creager E, Madras D, Pitassi T, Zemel R (2020) Causal modeling for fairness in dynamical systems. In: III HD, Singh A (eds) Proceedings of the 37th international conference on machine learning, PMLR, virtual, proceedings of machine learning research, vol 119, pp 2185–2195. http://proceedings.mlr.press/v119/creager20a.html

Creager E, Jacobsen JH, Zemel R (2021) Exchanging lessons between algorithmic fairness and domain generalization. https://openreview.net/forum?id=DC1Im3MkGG, neurIPS 2020 workshop: "algorithmic fairness through the lens of causality and interpretability (AFCI)"

D'Amour A, Srinivasan H, Atwood J, Baljekar P, Sculley D, Halpern Y (2020) Fairness is not static: deeper understanding of long term fairness via simulation studies. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '20, pp 525–534. https://doi.org/10.1145/3351095.3372878

Dash A, Chakraborty A, Ghosh S, Mukherjee A, Gummadi KP (2021) When the umpire is also a player: bias in private label product recommendations on e-commerce marketplaces. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAccT '21, pp 873–884. https://doi.org/10.1145/3442188.3445944

Datta S, Posada J, Olson G, Li W, O'Reilly C, Balraj D, Mesterhazy J, Pallas J, Desai P, Shah N (2020) A new paradigm for accelerating clinical data science at stanford medicine. arXiv:2003.10534

David KE, Liu Q, Fong R (2020) Debiasing convolutional neural networks via meta orthogonalization. NeurIPS 2020 workshop: "algorithmic fairness through the lens of causality and interpretability (AFCI)". arXiv:2011.07453

Davidson I, Ravi SS (2020) A framework for determining the fairness of outlier detection. In: ECAI 2020. IOS Press, pp 2465–2472

Davidson T, Warmsley D, Macy MW, Weber I (2017) Automated hate speech detection and the problem of offensive language. In: Proceedings of the eleventh international conference on web and social media, ICWSM 2017, Montréal, May 15–18, 2017, AAAI Press, pp 512–515. https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665

De-Arteaga M, Romanov A, Wallach H, Chayes J, Borgs C, Chouldechova A, Geyik S, Kenthapadi K, Kalai AT (2019) Bias in bios: A case study of semantic representation bias in a high-stakes setting. In: Proceedings of the conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '19, pp 120–128. https://doi.org/10.1145/3287560.3287572

Delobelle P, Temple P, Perrouin G, Frénay B, Heymans P, Berendt B (2020) Ethical adversaries: towards mitigating unfairness with adversarial machine learning. ECMLPKDD 2020 workshop: "BIAS 2020: bias and fairness in AI". arXiv:2005.06852

Deng J, Dong W, Socher R, Li L, Kai Li, Li Fei-Fei (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Denton E, Hanna A, Amironesei R, Smart A, Nicole H, Scheuerman MK (2020) Bringing the people back in: contesting benchmark machine learning datasets. arXiv:2007.07399

Deshpande KV, Pan S, Foulds JR (2020) Mitigating demographic bias in ai-based resume filtering. In: Adjunct publication of the 28th ACM conference on user modeling, adaptation and personalization. Association for Computing Machinery, New York, UMAP '20 Adjunct, pp 268–275. https://doi.org/10.1145/3386392.3399569,

Detrano R, Janosi A, Steinbrunn W, Pfisterer M, Schmid JJ, Sandhu S, Guppy KH, Lee S, Froelicher V (1989) International application of a new probability algorithm for the diagnosis of coronary artery disease. Am J Cardiol 64(5):304–310

Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp 4171–4186

Dhamala J, Sun T, Kumar V, Krishna S, Pruksachatkun Y, Chang KW, Gupta R (2021) Bold: dataset and metrics for measuring biases in open-ended language generation. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, association for computing machinery, New York, FAccT '21, pp 862–872. https://doi.org/10.1145/3442188.3445924

Diana E, Gill W, Kearns M, Kenthapadi K, Roth A (2021) Minimax group fairness: algorithms and experiments, association for computing machinery, New York, pp 66–76. https://doi.org/10.1145/3461702.3462523

DiCiccio C, Vasudevan S, Basu K, Kenthapadi K, Agarwal D (2020) evaluating fairness using permutation tests. Association for Computing Machinery, New York, pp 1467–1477. https://doi.org/10.1145/3394486.3403199

Dickens C, Singh R, Getoor L (2020) Hyperfair: A soft approach to integrating fairness criteria. RecSys 2020 workshop: "3rd FAccTRec workshop on responsible recommendation". arXiv:2009.08952

Dieterich W, Mendoza C, Brennan T (2016) Compas risk scales: demonstrating accuracy equity and predictive parity

Ding F, Hardt M, Miller J, Schmidt L (2021) Retiring adult: new datasets for fair machine learning. In: Advances in neural information processing systems 34

Dixon L, Li J, Sorensen J, Thain N, Vasserman L (2018) Measuring and mitigating unintended bias in text classification. In: Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, AIES '18, pp 67–73. https://doi.org/10.1145/3278721.3278729

Donini M, Oneto L, Ben-David S, Shawe-Taylor JS, Pontil M (2018) Empirical risk minimization under fairness constraints. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems. Curran Associates, Inc., vol 31, pp 2791–2801. https://proceedings.neurips.cc/paper/2018/file/83cdcec08fbf90370fcf53bdd56604ff-Paper.pdf

Dressel J, Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. Sci Adv 4(1):eaao5580

Duarte MF, Hu YH (2004) Vehicle classification in distributed sensor networks. J Parallel Distrib Comput 64(7):826–838. https://doi.org/10.1016/j.jpdc.2004.03.020

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. Association for Computing Machinery, New York, ITCS '12, pp 214–226. https://doi.org/10.1145/2090236.2090255

Dwork C, Immorlica N, Kalai AT, Leiserson M (2017) Decoupled classifiers for fair and efficient machine learning. KDD 2017 workshop: "fairness, accountability, and transparency in machine learning (FAT/ML). arXiv:1707.06613

Dwork C, Immorlica N, Kalai AT, Leiserson M (2018) Decoupled classifiers for group-fair and efficient machine learning. In: Friedler SA, Wilson C (eds) Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, New York, NY, USA, Proceedings of Machine Learning Research, vol 81, pp 119–133, http://proceedings.mlr.press/v81/dwork18a.html

Ebner NC, Riediger M, Lindenberger U (2010) Faces-a database of facial expressions in young, middle-aged, and older women and men: development and validation. Behav Res Methods 42(1):351–362

Eidinger E, Enbar R, Hassner T (2014) Age and gender estimation of unfiltered faces. IEEE Trans Inf Forensics Secur 9(12):2170–2179. https://doi.org/10.1109/TIFS.2014.2359646

Ekstrand MD, Tian M, Azpiazu IM, Ekstrand JD, Anuyah O, McNeill D, Pera MS (2018) All the cool kids, how do they fit in?: popularity and demographic biases in recommender evaluation and effectiveness. In: Friedler SA, Wilson C (eds) Proceedings of the 1st conference on fairness, accountability and transparency, PMLR, New York, proceedings of machine learning research, vol 81, pp 172–186. http://proceedings.mlr.press/v81/ekstrand18b.html

El Emam K, Arbuckle L, Koru G, Eze B, Gaudette L, Neri E, Rose S, Howard J, Gluck J (2012) De-identification methods for open health data: the case of the heritage health prize claims dataset. J Med Internet Res 14(1):e33

El Halabi M, Mitrović S, Norouzi-Fard A, Tardos J, Tarnawski JM (2020) Fairness in streaming submodular maximization: algorithms and hardness. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in neural information processing systems. Curran Associates, Inc., vol 33, pp 13609–13622. https://proceedings.neurips.cc/paper/2020/file/9d752cb08ef466fc480fba981cfa44a1-Paper.pdf

Elzayn H, Jabbari S, Jung C, Kearns M, Neel S, Roth A, Schutzman Z (2019) Fair algorithms for learning in allocation problems. In: Proceedings of the conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '19, pp 170–179. https://doi.org/10.1145/3287560.3287571

Epstein L, Landes W, Posner R (2013) the behavior of federal judges: a theoretical and empirical study of rational choice. Harvard University Press. https://books.google.it/books?id=RcQEBeic3ecC

Equivant (2019) Practitioner's guide to compas core. https://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf

Esmaeili S, Brubach B, Tsepenekas L, Dickerson J (2020) Probabilistic fair clustering. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in neural information processing systems. Curran Associates, Inc., vol 33, pp 12743–12755. https://proceedings.neurips.cc/paper/2020/file/95f2b84de5660ddf45c8a34933a2e66f-Paper.pdf

European Union (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/2016/679/2016-05-04

Fabbrizzi S, Papadopoulos S, Ntoutsi E, Kompatsiaris I (2021) A survey on bias in visual datasets. arXiv:2107.07919

Fabris A, Mishler A, Gottardi S, Carletti M, Daicampi M, Susto GA, Silvello G (2021) Algorithmic audit of Italian car insurance: evidence of unfairness in access and pricing. Association for Computing Machinery, New York, pp 458–468. https://doi.org/10.1145/3461702.3462569

Fabris A, Messina S, Silvello G, Susto GA (2022) Tackling documentation debt: a survey on algorithmic fairness datasets. In: Equity and access in algorithms, mechanisms, and optimization. Association for Computing Machinery, New York, NY. https://doi.org/10.1145/3551624.3555286

Farnad G, Babaki B, Gendreau M (2020) A unifying framework for fairness-aware influence maximization. In: Companion proceedings of the web conference 2020. Association for Computing Machinery, New York, WWW '20, pp 714–722. https://doi.org/10.1145/3366424.3383555

Farnadi G, Kouki P, Thompson SK, Srinivasan S, Getoor L (2018) A fairness-aware hybrid recommender system. RecSys 2018 workshop: "workshop on responsible recommendation (FAT/Rec)". arXiv:1809.09030

Farnadi G, Babaki B, Carvalho M (2019) Enhancing fairness in kidney exchange program by ranking solutions. NeurIPS 2019 workshop: "fair ML for Health". arXiv:1911.05489

Fehrman E, Muhammad AK, Mirkes EM, Egan V, Gorban AN (2017) The five factor model of personality and evaluation of drug consumption risk. In: Palumbo F, Montanari A, Vichi M (eds) Data Science. Springer, Cham, pp 231–242

Fehrman E, Egan V, Gorban AN, Levesley J, Mirkes EM, Muhammad AK (2019) Personality traits and drug consumption: a story told by data. Springer, Berlin

Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S (2015) Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, KDD '15, pp 259–268. https://doi.org/10.1145/2783258.2783311

Ferraro A, Bogdanov D, Serra X, Yoon J (2019) Artist and style exposure bias in collaborative filtering based music recommendations. ISMIR 2019 workshop: "workshop on designing human-centric MIR systems". arXiv:1911.04827

Fish B, Kun J, Lelkes Á (2015) Fair boosting: a case study. In: ICML 2015 workshop: "fairness, accountability, and transparency in machine learning (FAT/ML)"

Fisher RA (1936) The use of multiple measurements in taxonomic problems. Ann Eugen 7(2):179–188

Fisher J, Palfrey D, Christodoulopoulos C, Mittal A (2020) Measuring social bias in knowledge graph embeddings. AKBC 2020 workshop: "bias in automatic knowledge graph construction". arXiv:1912.02761

Fisman R, Iyengar S, Kamenica E, Simonson I (2006) Gender differences in mate selection: evidence from a speed dating experiment. Q J Econ 121:673–697. https://doi.org/10.1162/qjec.2006.121.2.673

Fitzpatrick TB (1988) The validity and practicality of sun-reactive skin types i through vi. Arch Dermatol 124(6):869–871

Flanigan B, Gölz P, Gupta A, Procaccia AD (2020) Neutralizing self-selection bias in sampling for sortition. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, NeurIPS 2020, December 6–12, 2020, virtual. https://proceedings.neurips.cc/paper/2020/hash/48237d9f2dea8c74c2a72126cf63d933-Abstract.html

Florez OU (2019) On the unintended social bias of training language generation models with data from local media. NeurIPS 2019 workshop: "human-centric machine learning". arXiv:1911.00461

Fogliato R, Xiang A, Lipton Z, Nagin D, Chouldechova A (2021) On the validity of arrest as a proxy for offense: race and the likelihood of arrest for violent crimes. In: Proceedings of the 4th AAAI/ACM conference on AI, ethics, and society (AIES 2021), Virtual Event, pp 100–111. https://doi.org/10.1145/3461702.3462538

Founta A, Djouvas C, Chatzakou D, Leontiadis I, Blackburn J, Stringhini G, Vakali A, Sirivianos M, Kourtellis N (2018) Large scale crowdsourcing and characterization of twitter abusive behavior. In: Proceedings of the twelfth international conference on web and social media, ICWSM 2018, Stanford, June 25–28, 2018, AAAI Press, pp 491–500. https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17909

Framingham Heart Study (2021) Framingham heart study. offspring exam 10, omni 1 exam 5. research consent form. https://framinghamheartstudy.org/files/2021/01/FHS-Offspring-Exam-10-Omni-1-Exam-5-Informed-Consent-English-Language-v21.pdf

Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton EP, Roth D (2019) A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '19, pp 329–338. https://doi.org/10.1145/3287560.3287589

Friedler SA, Scheidegger C, Venkatasubramanian S (2021) The (im)possibility of fairness: different value systems require different mechanisms for fair decision making. Commun ACM 64(4):136–143. https://doi.org/10.1145/3433949

Galhotra S, Saisubramanian S, Zilberstein S (2021) Learning to generate fair clusters from demonstrations. Association for Computing Machinery, New York, pp 491–501. https://doi.org/10.1145/3461702.3462558

Garbin C, Rajpurkar P, Irvin J, Lungren MP, Marques O (2021) Structured dataset documentation: a datasheet for chexpert. arXiv:2105.03020

Garg S, Perot V, Limtiaco N, Taly A, Chi EH, Beutel A (2019) Counterfactual fairness in text classification through robustness. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, AIES '19, pp 219–226. https://doi.org/10.1145/3306618.3317950

Gastwirth JL, Miao W (2009) Formal statistical analysis of the data in disparate impact cases provides sounder inferences than the us government's 'four-fifths' rule: an examination of the statistical evidence in ricci v. destefano. Law Probab Risk 8(2):171–191

Ge H, Caverlee J, Lu H (2016) Taper: A contextual tensor-based approach for personalized expert recommendation. In: Proceedings of the 10th ACM conference on recommender systems. Association for Computing Machinery, New York, RecSys '16, pp 261–268. https://doi.org/10.1145/2959100.2959151

Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé III H, Crawford K (2018) Datasheets for datasets. arXiv:1803.09010

Geiger RS, Yu K, Yang Y, Dai M, Qiu J, Tang R, Huang J (2020) Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In: Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '20, pp 325–336. https://doi.org/10.1145/3351095.3372862

Gelman A, Fagan J, Kiss A (2007) An analysis of the new york city police department's "stop-and-frisk" policy in the context of claims of racial bias. J Am Stat Assoc 102(479):813–823

Gerritse EJ, de Vries AP (2020) Effect of debiasing on information retrieval. In: Boratto L, Faralli S, Marras M, Stilo G (eds) Bias and social aspects in search and recommendation. Springer, Cham, pp 35–42

Ghadiri M, Samadi S, Vempala S (2021) Socially fair k-means clustering. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAccT '21, pp 438–448. https://doi.org/10.1145/3442188.3445906

Ginart A, Guan M, Valiant G, Zou JY (2019) Making ai forget you: data deletion in machine learning. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems, vol 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/cb79f8fa58b91d3af6c9c991f63962d3-Paper.pdf

Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. Processing 150

Goel N, Faltings B (2019) Crowdsourcing with fairness, diversity and budget constraints. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, AIES '19, pp 297–304. https://doi.org/10.1145/3306618.3314282

Goel S, Rao JM, Shroff R et al (2016) Precinct or prejudice? understanding racial disparities in New York city's stop-and-frisk policy. Ann Appl Stat 10(1):365–394

Goel S, Perelman M, Shroff R, Sklansky D (2017) Combatting police discrimination in the age of big data. New Crim Law Rev 20(2):181–232. https://doi.org/10.1525/nclr.2017.20.2.181

Goel N, Yaghini M, Faltings B (2018) Non-discriminatory machine learning through convex fairness criteria. In: Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, AIES '18, pp 116. https://doi.org/10.1145/3278721.3278722

Goel N, Amayuelas A, Deshpande A, Sharma A (2020) The importance of modeling data missingness in algorithmic fairness: a causal perspective. NeurIPS 2020 workshop: "algorithmic fairness through the lens of causality and interpretability (AFCI)". arXiv:2012.11448

Goelz P, Kahng A, Procaccia AD (2019) Paradoxes in fair machine learning. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems. Curran Associates, Inc., vol 32, pp 8342–8352. https://proceedings.neurips.cc/paper/2019/file/bbc92a647199b832ec90d7cf57074e9e-Paper.pdf

Golbeck J, Ashktorab Z, Banjo RO, Berlinger A, Bhagwan S, Buntain C, Cheakalos P, Geller AA, Gergory Q, Gnanasekaran RK, Gunasekaran RR, Hoffman KM, Hottle J, Jienjitlert V, Khare S, Lau R, Martindale MJ, Naik S, Nixon HL, Ramachandran, Rogers KM, Rogers L, Sarin MS, Shahane G, Thanki J, Vengataraman P, Wan Z, Wu DM (2017) A large labeled corpus for online harassment research. In: Proceedings of the 2017 ACM on web science conference. Association for Computing Machinery, New York, WebSci '17, pp 229–233. https://doi.org/10.1145/3091478.3091509

Goldstein H (1991) Multilevel modelling of survey data. J R Stat Soc Ser D (Statist) 40(2):235–244, http://www.jstor.org/stable/2348496

Gong S, Liu X, Jain AK (2021) Mitigating face recognition bias via group adaptive classifier. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 3414–3424

Gordaliza P, Barrio ED, Fabrice G, Loubes JM (2019) Obtaining fairness using optimal transport theory. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, California, proceedings of machine learning research, vol 97, pp 2357–2365. http://proceedings.mlr.press/v97/gordaliza19a.html

Gordon J, Babaeianjelodar M, Matthews J (2020) Studying political bias via word embeddings. In: Companion proceedings of the web conference 2020. Association for Computing Machinery, New York, WWW '20, pp 760-764. https://doi.org/10.1145/3366424.3383560

Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D (2017) Making the v in vqa matter: elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6904–6913

Graffam J, Shinkfield AJ, Hardcastle L (2008) The perceived employability of ex-prisoners and offenders. Int J Offender Ther Comp Criminol 52(6):673–685. https://doi.org/10.1177/0306624X07307783

Green B, Chen Y (2019) Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In: Proceedings of the conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, NY, FAT* '19, pp 90–99. https://doi.org/10.1145/3287560.3287563

Greenwald AG, McGhee DE, Schwartz JL (1998) Measuring individual differences in implicit cognition: the implicit association test. J Pers Soc Psychol 74(6):1464

Grgic-Hlaca N, Zafar M, Gummadi K, Weller A (2016) The case for process fairness in learning: feature selection for fair decision making. NeurIPS 2016 workshop: "machine learning and the law"

Grömping U (2019) South German credit data: correcting a widely used data set. Report. Tech. rep., Beuth University of Applied Sciences Berlin. http://www1.beuth-hochschule.de/FB_II/reports/Report-2019-004.pdf

Gulla JA, Zhang L, Liu P, Özgöbek O, Su X (2017) The adressa dataset for news recommendation. In: Proceedings of the international conference on web intelligence. Association for Computing Machinery, New York, WI '17, pp 1042–1048. https://doi.org/10.1145/3106426.3109436

Gungor A (2018) Benchmarking authorship attribution techniques using over a thousand books by fifty victorian era novelists. Master's thesis, Purdue University

Guo W, Caliskan A (2021) Detecting emergent intersectional biases: contextualized word embeddings contain a distribution of human-like biases. Association for Computing Machinery, New York, pp 122–133. https://doi.org/10.1145/3461702.3462536

Guo G, Zhang J, Yorke-Smith N (2016a) A novel evidence-based Bayesian similarity measure for recommender systems. ACM Trans Web. https://doi.org/10.1145/2856037

Guo Y, Zhang L, Hu Y, He X, Gao J (2016b) Ms-celeb-1m: a dataset and benchmark for large-scale face recognition. In: Leibe B, Matas J, Sebe N, Welling M (eds) Computer vision—ECCV 2016. Springer, Cham, pp 87–102

Guvenir HA, Acar B, Demiroz G, Cekin A (1997) A supervised machine learning algorithm for arrhythmia analysis. In: Computers in cardiology 1997. IEEE, pp 433–436

Han H, Jain AK (2014) Age, gender and race estimation from unconstrained face images. http://biometrics.cse.msu.edu/Publications/Face/HanJain_UnconstrainedAgeGenderRaceEstimation_MSUTechReport2014.pdf

Hannák A, Wagner C, Garcia D, Mislove A, Strohmaier M, Wilson C (2017) Bias in online freelance marketplaces: evidence from taskrabbit and fiverr. In: Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing. Association for Computing Machinery, New York, CSCW '17, pp 1914–1933. https://doi.org/10.1145/2998181.2998327

Har-Peled S, Mahabadi S (2019) Near neighbor: Who is the fairest of them all? In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems, vol 32. Curran Associates, Inc., pp 13176–13187. https://proceedings.neurips.cc/paper/2019/file/742141ceda6b8f6786609d31c8ef129f-Paper.pdf

Harb E, Lam HS (2020) Kfc: A scalable approximation algorithm for k-center fair clustering. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in neural information processing systems, vol 33. Curran Associates, Inc., pp 14509–14519. https://proceedings.neurips.cc/paper/2020/file/a6d259bfbfa2062843ef543e21d7ec8e-Paper.pdf

Hardt M, Price E, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R (eds) Advances in neural information processing systems, vol 29. Curran Associates, Inc., pp 3315–3323. https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

Harper FM, Konstan JA (2015) The movielens datasets: history and context. ACM Trans Interact Intell Syst. https://doi.org/10.1145/2827872

Hashimoto T, Srivastava M, Namkoong H, Liang P (2018) Fairness without demographics in repeated loss minimization. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Stockholm Sweden, proceedings of machine learning research, vol 80, pp 1929–1938. http://proceedings.mlr.press/v80/hashimoto18a.html

He R, McAuley J (2016) Ups and downs. Proceedings of the 25th international conference on World Wide Web https://doi.org/10.1145/2872427.2883037

He R, Kang WC, McAuley J (2017) Translation-based recommendation. In: Proceedings of the eleventh ACM conference on recommender systems. Association for Computing Machinery, New York, RecSys '17, pp 161–169. https://doi.org/10.1145/3109859.3109882

He Y, Burghardt K, Guo S, Lerman K (2020a) Inherent trade-offs in the fair allocation of treatments. NeurIPS 2020 workshop: "algorithmic fairness through the lens of causality and interpretability (AFCI)". arXiv:2010.16409

He Y, Burghardt K, Lerman K (2020b) A geometric solution to fair representations. In: Proceedings of the AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, AIES '20, pp 279–285. https://doi.org/10.1145/3375627.3375864

Heidari H, Ferrari C, Gummadi K, Krause A (2018) Fairness behind a veil of ignorance: a welfare analysis for automated decision making. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems, vol 31. Curran Associates, Inc., pp 1265–1276. https://proceedings.neurips.cc/paper/2018/file/be3159ad04564bfb90db9e32851ebf9c-Paper.pdf

Heidari H, Loi M, Gummadi KP, Krause A (2019a) A moral framework for understanding fair ml through economic models of equality of opportunity. In: Proceedings of the conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '19, pp 181–190. https://doi.org/10.1145/3287560.3287584

Heidari H, Nanda V, Gummadi K (2019b) On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, proceedings of machine learning research, vol 97, pp 2692–2701. http://proceedings.mlr.press/v97/heidari19a.html

Hendricks LA, Burns K, Saenko K, Darrell T, Rohrbach A (2018) Women also snowboard: overcoming bias in captioning models. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer vision-ECCV 2018. Springer, Cham, pp 793–811

Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, Mohamed S, Lerchner A (2017) beta-vae: learning basic visual concepts with a constrained variational framework. In: ICLR

Holland S, Hosny A, Newman S, Joseph J, Chmielinski K (2018) The dataset nutrition label: a framework to drive higher data quality standards. arXiv:1805.03677

Hollywood J, McKay K, Woods D, Agniel D (2019) Real time crime centers in chicago. https://www.rand.org/content/dam/rand/pubs/research_reports/RR3200/RR3242/RAND_RR3242.pdf

Holmes MD, Smith BW, Freng AB, Muñoz EA (2008) Minority threat, crime control, and police resource allocation in the southwestern united states. Crime Delinq 54(1):128–152. https://doi.org/10.1177/0011128707309718

Holstein K, Wortman Vaughan J, Daumé III H, Dudik M, Wallach H (2019) Improving fairness in machine learning systems: What do industry practitioners need? In: Proceedings of the ACM conference on human factors in computing systems (CHI 2019), Glasgow, pp 1–16

Houvardas J, Stamatatos E (2006) N-gram feature selection for authorship identification. In: International conference on artificial intelligence: methodology, systems, and applications. Springer, pp 77–86

Hu L, Chen Y (2020) Fair classification and social welfare. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '20, pp 535–545. https://doi.org/10.1145/3351095.3372857

Hu Y, Wu Y, Zhang L, Wu X (2020) Fair multiple decision making through soft interventions. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, NeurIPS 2020, December 6–12, 2020, virtual. https://proceedings.neurips.cc/paper/2020/hash/d0921d442ee91b89ad95059d13df618-Abstract.html

Huan W, Wu Y, Zhang L, Wu X (2020) Fairness through equality of effort. In: Companion proceedings of the web conference 2020. Association for Computing Machinery, New York, WWW '20, pp 743–751. https://doi.org/10.1145/3366424.3383558,

Huang L, Vishnoi N (2019) Stable and fair classification. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, proceedings of machine learning research, vol 97, pp 2879–2890. http://proceedings.mlr.press/v97/huang19e.html

Huang GB, Ramesh M, Berg T, Learned-Miller E (2007) Labeled faces in the wild: a database for studying face recognition in unconstrained environments

Huang L, Jiang S, Vishnoi N (2019) Coresets for clustering with fairness constraints. In: Advances in neural information processing systems, pp 7589–7600

Huang L, Wei J, Celis E (2020) Towards just, fair and interpretable methods for judicial subset selection. In: Proceedings of the AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, AIES '20, pp 293–299. https://doi.org/10.1145/3375627.3375848

Hull J (1994) A database for handwritten text recognition research. IEEE Trans Pattern Anal Mach Intell 16(5):550–554. https://doi.org/10.1109/34.291440

Hussain S, Dahan NA, Ba-Alwib FM, Ribata N (2018) Educational data mining and analysis of students' academic performance using weka. Indones J Electr Eng Comput Sci 9(2):447–459

Hutchinson B, Prabhakaran V, Denton E, Webster K, Zhong Y, Denuyl S (2020) Unintended machine learning biases as social barriers for persons with disabilities. In: SIGACCESS Access Comput, vol 125https://doi.org/10.1145/3386296.3386305

Häußler M Walter (1979) Empirische ergebnisse zu diskriminationsverfahren bei kreditscoringsystemen. https://link.springer.com/article/10.1007/BF01917956

International Warfarin Pharmacogenetics Consortium (2009) Estimation of the warfarin dose with clinical and pharmacogenetic data. N Engl J Med 360(8):753–764

Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball R, Shpanskaya K, Seekins J, Mong D, Halabi S, Sandberg J, Jones R, Larson D, Langlotz C, Patel B, Lungren M, Ng A (2019) Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: 33rd AAAI conference on artificial intelligence, AAAI 2019, 31st innovative applications of artificial intelligence conference, IAAI 2019 and the 9th AAAI symposium on educational advances in artificial intelligence, EAAI 2019, AAAI Press, 33rd AAAI conference on artificial intelligence, AAAI 2019, 31st innovative applications of artificial intelligence conference, IAAI 2019 and the 9th AAAI symposium on educational advances in artificial intelligence, EAAI 2019, pp 590–597, publisher Copyright: 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.; 33rd AAAI conference on artificial intelligence, AAAI 2019, 31st annual conference on innovative applications of artificial intelligence, IAAI 2019 and the 9th AAAI symposium on educational advances in artificial intelligence, EAAI 2019 ; conference date: 27-01-2019 through 01-02-2019

Islam R, Pan S, Foulds JR (2021) Can we obtain fairness for free?. Association for Computing Machinery, New York, pp 586–596. https://doi.org/10.1145/3461702.3462614

Jabbari S, Ou HC, Lakkaraju H, Tambe M (2020) An empirical study of the trade-offs between interpretability and fairness. In: ICML 2020 workshop on human interpretability in machine learning, preliminary version, iCML 2020 workshop: "Workshop on human interpretability in machine learning (WHI)"

Jagielski M, Kearns M, Mao J, Oprea A, Roth A, Malvajerdi SS, Ullman J (2019) Differentially private fair learning. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, proceedings of machine learning research, vol 97, pp 3000–3008. http://proceedings.mlr.press/v97/jagielski19a.html

Ji D, Smyth P, Steyvers M (2020) Can I trust my fairness metric? assessing fairness with unlabeled data and bayesian inference. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, NeurIPS 2020, December 6–12, 2020, virtual. https://proceedings.neurips.cc/paper/2020/hash/d83de59e10227072a9c034ce10029c39-Abstract.html

Jiang W, Pardos ZA (2021) Towards equity and algorithmic fairness in student grade prediction. Association for Computing Machinery, New York, pp 608–617. https://doi.org/10.1145/3461702.3462623

Jo ES, Gebru T (2020) Lessons from archives: Strategies for collecting sociocultural data in machine learning. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '20, pp 306–316. https://doi.org/10.1145/3351095.3372829

Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG (2016) Mimic-iii, a freely accessible critical care database. Sci Data 3:160035

Johnson AE, Pollard TJ, Greenbaum NR, Lungren MP, Deng Cy, Peng Y, Lu Z, Mark RG, Berkowitz SJ, Horng S (2019) Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv:1901.07042

Jones M, Nguyen H, Nguyen T (2020) Fair k-centers via maximum matching. In: III HD, Singh A (eds) Proceedings of the 37th international conference on machine learning, PMLR, virtual, proceedings of machine learning research, vol 119, pp 4940–4949. http://proceedings.mlr.press/v119/jones20a.html

Jones E, Sagawa S, Koh PW, Kumar A, Liang P (2021) Selective classification can magnify disparities across groups. In: International conference on learning representations. https://openreview.net/forum?id=N0M_4BkQ05i

Jung S, Lee D, Park T, Moon T (2021) Fair feature distillation for visual recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 12115–12124

Kallus N, Zhou A (2018) Residual unfairness in fair machine learning from prejudiced data. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Stockholm Sweden, proceedings of machine learning research, vol 80, pp 2439–2448. http://proceedings.mlr.press/v80/kallus18a.html

Kallus N, Zhou A (2019a) Assessing disparate impact of personalized interventions: identifiability and bounds. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems, vol 32. Curran Associates, Inc., pp 3426–3437. https://proceedings.neurips.cc/paper/2019/file/d54e99a6c03704e95e6965532dec148b-Paper.pdf

Kallus N, Zhou A (2019b) The fairness of risk scores beyond classification: bipartite ranking and the xauc metric. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems, vol 32. Curran Associates, Inc., pp 3438–3448. https://proceedings.neurips.cc/paper/2019/file/73e0f7487b8e5297182c5a711d20bf26-Paper.pdf

Kallus N, Zhou A (2021) Fairness, welfare, and equity in personalized pricing. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAccT '21, pp 296–314. https://doi.org/10.1145/3442188.3445895

Kallus N, Mao X, Zhou A (2020) Assessing algorithmic fairness with unobserved protected class using data combination. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '20, pp 110. https://doi.org/10.1145/3351095.3373154

Kamishima T (2003) Nantonac collaborative filtering: Recommendation based on order responses. In: Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, KDD '03, pp 583–588. https://doi.org/10.1145/956750.956823

Kang J, He J, Maciejewski R, Tong H (2020) InFoRM: individual fairness on graph mining. Association for Computing Machinery, New York, pp 379–389. https://doi.org/10.1145/3394486.3403080

Kannel WB, McGee DL (1979) Diabetes and cardiovascular disease: the framingham study. JAMA 241(19):2035–2038

Karako C, Manggala P (2018) Using image fairness representations in diversity-based re-ranking for recommendations. UMAP 2018 workshop: "fairness in user modeling, adaptation and personalization (FairUMAP)". arXiv:1809.03577

Karkkainen K, Joo J (2021) Fairface: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 1548–1558

Karlan DS, Zinman J (2008) Credit elasticities in less-developed economies: implications for microfinance. Am Econ Rev 98(3):1040–68

Kasy M, Abebe R (2021) Fairness, equality, and power in algorithmic decision-making. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAccT '21, pp 576–586. https://doi.org/10.1145/3442188.3445919

Kato M, Teshima T, Honda J (2019) Learning from positive and unlabeled data with a selection bias. In: International conference on learning representations. https://openreview.net/forum?id=rJzLciCqKm

Kearns M, Neel S, Roth A, Wu ZS (2018) Preventing fairness gerrymandering: auditing and learning for subgroup fairness. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Stockholm Sweden, proceedings of machine learning research, vol 80, pp 2564–2572. http://proceedings.mlr.press/v80/kearns18a.html

Kearns M, Neel S, Roth A, Wu ZS (2019) An empirical study of rich subgroup fairness for machine learning. In: Proceedings of the conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '19, pp 100–109. https://doi.org/10.1145/3287560.3287592

Keswani V, Lease M, Kenthapadi K (2021) Towards unbiased and accurate deferral to multiple experts. Association for Computing Machinery, New York, pp 154–165. https://doi.org/10.1145/3461702.3462516

Keyes O, Stevens N, Wernimont J (2019) The government is using the most vulnerable people to test facial recognition software

Khan Z, Fu Y (2021) One label, one billion faces: Usage and consistency of racial categories in computer vision. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAccT '21, pp 587–597. https://doi.org/10.1145/3442188.3445920

Kilbertus N, Gascon A, Kusner M, Veale M, Gummadi K, Weller A (2018) Blind justice: fairness with encrypted sensitive attributes. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Stockholm Sweden, proceedings of machine learning research, vol 80, pp 2630–2639. http://proceedings.mlr.press/v80/kilbertus18a.html

Kim H, Mnih A (2018) Disentangling by factorising. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, proceedings of machine learning research, vol 80, pp 2649–2658. http://proceedings.mlr.press/v80/kim18b.html

Kim MP, Ghorbani A, Zou J (2019) Multiaccuracy: Black-box post-processing for fairness in classification. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, AIES '19, pp 247–254. https://doi.org/10.1145/3306618.3314287

Kim JS, Chen J, Talwalkar A (2020) FACT: A diagnostic for group fairness trade-offs. In: III HD, Singh A (eds) Proceedings of the 37th international conference on machine learning, PMLR, virtual, proceedings of machine learning research, vol 119, pp 5264–5274. http://proceedings.mlr.press/v119/kim20a.html

Kim E, Bryant D, Srikanth D, Howard A (2021) Age bias in emotion detection: an analysis of facial emotion recognition performance on young, middle-aged, and older adults. Association for Computing Machinery, New York, pp 638–644. https://doi.org/10.1145/3461702.3462609

Kiritchenko S, Mohammad S (2018) Examining gender and race bias in two hundred sentiment analysis systems. In: Proceedings of the seventh joint conference on lexical and computational semantics.

Association for Computational Linguistics, New Orleans, pp 43–53. https://doi.org/10.18653/v1/S18-2005, https://aclanthology.org/S18-2005

Kizhner I, Terras M, Rumyantsev M, Khokhlova V, Demeshkova E, Rudov I, Afanasieva J (2020) Digital cultural colonialism: measuring bias in aggregated digitized content held in Google Arts and Culture. Digital Scholarship in the Humanities 36(3):607–640. https://doi.org/10.1093/llc/fqaa055, https://academic.oup.com/dsh/article-pdf/36/3/607/40873280/fqaa055.pdf

Klare BF, Klein B, Taborsky E, Blanton A, Cheney J, Allen K, Grother P, Mah A, Jain AK (2015) Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)

Kleindessner M, Awasthi P, Morgenstern J (2019a) Fair k-center clustering for data summarization. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, California, USA, proceedings of machine learning research, vol 97, pp 3448–3457. http://proceedings.mlr.press/v97/kleindessner19a.html

Kleindessner M, Samadi S, Awasthi P, Morgenstern J (2019b) Guarantees for spectral clustering with fairness constraints. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, proceedings of machine learning research, vol 97, pp 3458–3467. http://proceedings.mlr.press/v97/kleindessner19b.html

Knees P, Hübler M (2019) Towards uncovering dataset biases: investigating record label diversity in music playlists. ISMIR 2019 workshop: “workshop on designing human-centric MIR systems”

Kobren A, Saha B, McCallum A (2019) Paper matching with local fairness constraints. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, KDD ’19, pp 1247–1257. https://doi.org/10.1145/3292500.3330899

Kocijan V, Camburu OM, Lukasiewicz T (2020) The gap on gap: tackling the problem of differing data distributions in bias-measuring datasets. NeurIPS 2020 workshop: “algorithmic fairness through the lens of causality and interpretability (AFCI)”. arXiv:2011.01837

Kohavi R (1996) 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. AAAI Press, KDD’96, pp 202–207

Komiyama J, Takeda A, Honda J, Shimao H (2018) Nonconvex optimization for regression with fairness constraints. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Stockholm Sweden, proceedings of machine learning research, vol 80, pp 2737–2746. http://proceedings.mlr.press/v80/komiyama18a.html

Konstantakis G, Promponas G, Dretakis M, Papadakos P (2020) Bias goggles: exploring the bias of web domains through the eyes of users. In: Boratto L, Faralli S, Marras M, Stilo G (eds) Bias and social aspects in search and recommendation. Springer, Cham, pp 66–71

Koolen C (2018) Reading beyond the female the relationship between perception of author gender and literary quality. PhD thesis, University of Amsterdam

Koolen C, van Cranenburgh A (2017) These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. In: Proceedings of the First ACL workshop on ethics in natural language processing. Association for Computational Linguistics, Valencia, pp 12–22. https://doi.org/10.18653/v1/W17-1602, https://www.aclweb.org/anthology/W17-1602

Krizhevsky A (2009) Learning multiple layers of features from tiny images

Kröger JL, Miceli M, Müller F (2021) How data can be used against people: a classification of personal data misuses. Available at SSRN 3887097

Kuhlman C, Rundensteiner E (2020) Rank aggregation algorithms for fair consensus. Proc VLDB Endow 13(12):2706–2719. https://doi.org/10.14778/3407790.3407855

Kuhlman C, Gerych W, Rundensteiner E (2021) Measuring group advantage: a comparative study of fair ranking metrics. In: Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, AIES ’21, pp 674–682. https://doi.org/10.1145/3461702.3462588

Kulshrestha J, Eslami M, Messias J, Zafar MB, Ghosh S, Gummadi KP, Karahalios K (2017) Quantifying search bias: investigating sources of bias for political searches in social media. In: Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing. Association for Computing Machinery, New York, CSCW ’17, pp 417–432. https://doi.org/10.1145/2998181.2998321

Kushmerick N (1999) Learning to remove internet advertisements. In: Proceedings of the third annual conference on autonomous agents. Association for Computing Machinery, New York, AGENTS '99, pp 175–181. https://doi.org/10.1145/301136.301186,

Kusner MJ, Loftus J, Russell C, Silva R (2017) Counterfactual fairness. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems, vol 30. Curran Associates, Inc., pp 4066–4076. https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

Kuznetsova A, Rom H, Alldrin N, Uijlings J, Krasin I, Pont-Tuset J, Kamali S, Popov S, Malloci M, Kolesnikov A et al (2020) The open images dataset v4. Int J Comput Vis 128(7):1956–1981

Kügelgen JV, Karimi AH, Bhatt U, Valera I, Weller A, Schölkopf B (2021) On the fairness of causal algorithmic recourse. NeurIPS 2020 workshop: "algorithmic fairness through the lens of causality and interpretability (AFCI)". arXiv:2010.06529

Lahoti P, Beutel A, Chen J, Lee K, Prost F, Thain N, Wang X, Chi E (2020) Fairness without demographics through adversarially reweighted learning. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in neural information processing systems, vol 33. Curran Associates, Inc., pp 728–740. https://proceedings.neurips.cc/paper/2020/file/07fc15c9d169ee48573edd749d25945d-Paper.pdf

Lake BM, Salakhutdinov R, Tenenbaum JB (2015) Human-level concept learning through probabilistic program induction. Science 350(6266):1332–1338. https://doi.org/10.1126/science.aab3050

Lamy A, Zhong Z, Menon AK, Verma N (2019) Noise-tolerant fair classification. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems, vol 32. Curran Associates, Inc., pp 294–306. https://proceedings.neurips.cc/paper/2019/file/8d5e957f297893487bd98fa830fa6413-Paper.pdf

Lan C, Huan J (2017) Discriminatory transfer. KDD 2017 workshop: "fairness, accountability, and transparency in machine learning (FAT/ML)". arXiv:1707.00780

Larson J, Mattu S, Kirchner L, Angwin J (2016) How we analyzed the compas recidivism algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Le Quy T, Roy A, Iosifidis V, Zhang W, Ntoutsi E (2022) A survey on datasets for fairness-aware machine learning. WIREs Data Mining and Knowledge Discovery n/a(n/a):e1452, https://doi.org/10.1002/widm.1452, https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1452

Leavy S, Meaney G, Wade K, Greene D (2019) Curatr: a platform for semantic analysis and curation of historical literary texts, pp 354–366. https://doi.org/10.1007/978-3-030-36599-8_31

Leavy S, Meaney G, Wade K, Greene D (2020) Mitigating gender bias in machine learning data sets. In: Boratto L, Faralli S, Marras M, Stilo G (eds) Bias and social aspects in search and recommendation. Springer, Cham, pp 12–26

Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324. https://doi.org/10.1109/5.726791

LeCun Y, Fu Jie Huang, Bottou L (2004) Learning methods for generic object recognition with invariance to pose and lighting. In: Proceedings of the 2004 IEEE computer society conference on computer vision and pattern recognition, 2004. CVPR 2004., vol 2, pp II–104. https://doi.org/10.1109/CVPR.2004.1315150

Lee H, Kizilcec RF (2020) Evaluation of fairness trade-offs in predicting student success. In: International conference on educational data mining workshop: "fairness, accountability, and transparency, in educational data (mining)". arXiv:2007.00088

Leonelli S, Tempini N (2020) Data journeys in the sciences. Springer, Berlin

Leskovec J, Mcauley J (2012) Learning to discover social circles in ego networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in neural information processing systems, vol 25. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2012/file/7a614fd06c325499f1680b9896beedeb-Paper.pdf

Leskovec J, Kleinberg J, Faloutsos C (2007) Graph evolution: densification and shrinking diameters. ACM Trans Knowl Discov From Data (TKDD) 1(1):2

Lesmana NS, Zhang X, Bei X (2019) Balancing efficiency and fairness in on-demand ridesourcing. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems, , vol 32. Curran Associates, Inc., pp 5309–5319. https://proceedings.neurips.cc/paper/2019/file/3070e6addcd702cb58de5d7897bfdae1-Paper.pdf

Levy D, Splansky GL, Strand NK, Atwood LD, Benjamin EJ, Blease S, Cupples LA, D'Agostino RB Sr, Fox CS, Kelly-Hayes M et al (2010) Consent for genetic research in the framingham heart study. Am J Med Genet A 152(5):1250–1256

Li Z, Zhao H, Liu Q, Huang Z, Mei T, Chen E (2018) Learning from history and present: next-item recommendation via discriminatively exploiting user behaviors. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, KDD '18, pp 1734–1743. https://doi.org/10.1145/3219819.3220014

Li P, Zhao H, Liu H (2020a) Deep fair clustering for visual learning. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR)

Li T, Sanjabi M, Beirami A, Smith V (2020b) Fair resource allocation in federated learning. In: International conference on learning representations. https://openreview.net/forum?id=ByexElSYDr

Li Y, Ning Y, Liu R, Wu Y, Hui Wang W (2020c) Fairness of classification using users' social relationships in online peer-to-peer lending. In: Companion proceedings of the web conference 2020. Association for Computing Machinery, New York, WWW '20, pp 733–742. https://doi.org/10.1145/3366424.3383557

Li Y, Sun H, Wang WH (2020d) Towards fair truth discovery from biased crowdsourced answers. Association for Computing Machinery, New York, pp 599–607. https://doi.org/10.1145/3394486.3403102

Li P, Wang Y, Zhao H, Hong P, Liu H (2021) On dyadic fairness: exploring and mitigating bias in graph connections. In: International conference on learning representations. https://openreview.net/forum?id=xgGS6PmzNq6

Liang L, Acuna DE (2020) Artificial mental phenomena: Psychophysics as a framework to detect perception biases in ai models. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '20, pp 403–412. https://doi.org/10.1145/3351095.3375623

Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: common objects in context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) Computer vision-ECCV 2014. Springer, Cham, pp 740–755

Lipton Z, McAuley J, Chouldechova A (2018) Does mitigating ml' s impact disparity require treatment disparity? In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems, vol 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/8e0384779e58ce2af40eb365b318cc32-Paper.pdf

Liu W, Burke R (2018) Personalizing fairness-aware re-ranking. RecSys 2018 workshop: "workshop on responsible recommendation (FAT/Rec)". arXiv:1809.02921

Liu Z, Luo P, Wang X, Tang X (2015) Deep learning face attributes in the wild. arXiv:1411.7766

Liu LT, Dean S, Rolf E, Simchowitz M, Hardt M (2018) Delayed impact of fair machine learning. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Stockholm Sweden, proceedings of machine learning research, vol 80, pp 3150–3158. http://proceedings.mlr.press/v80/liu18c.html

Liu LT, Simchowitz M, Hardt M (2019) The implicit fairness criterion of unconstrained learning. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, proceedings of machine learning research, vol 97, pp 4051–4060. http://proceedings.mlr.press/v97/liu19f.html

Liu LT, Wilson A, Haghtalab N, Kalai AT, Borgs C, Chayes J (2020) The disparate equilibria of algorithmic decision making when individuals invest rationally. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '20, pp 381–391. https://doi.org/10.1145/3351095.3372861

Liu D, Shafi Z, Fleisher W, Eliassi-Rad T, Alfeld S (2021) RAWLSNET: altering bayesian networks to encode rawlsian fair equality of opportunity. Association for Computing Machinery, New York, pp 745–755. https://doi.org/10.1145/3461702.3462618

Locatello F, Abbati G, Rainforth T, Bauer S, Schölkopf B, Bachem O (2019) On the fairness of disentangled representations. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems, vol 32. Curran Associates, Inc., pp 14611–14624. https://proceedings.neurips.cc/paper/2019/file/1b486d7a5189ebe8d8c46afc64b0d1b4-Paper.pdf

Lohaus M, Perrot M, Luxburg UV (2020) Too relaxed to be fair. In: III HD, Singh A (eds) Proceedings of the 37th international conference on machine learning, PMLR, Virtual, proceedings of machine learning research, vol 119, pp 6360–6369. http://proceedings.mlr.press/v119/lohaus20a.html

Louizos C, Swersky K, Li Y, Welling M, Zemel RS (2016) The variational fair autoencoder. In: Bengio Y, LeCun Y (eds) 4th international conference on learning representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, conference track proceedings. arXiv:1511.00830

Lowe H, Ferris TA, Hernandez P, Weber S (2009) Stride—an integrated standards-based translational research informatics platform. AMIA Ann Sympos Proc AMIA Sympos 2009:391–5

Lu Q, Getoor L (2003) Link-based classification. In: Proceedings of the twentieth international conference on international conference on machine learning. AAAI Press, ICML'03, pp 496–503

Lum K, Johndrow J (2016) A statistical framework for fair predictive algorithms. DTL 2016 workshop: "fairness, accountability, and transparency in machine learning (FAT/ML)". arXiv:1610.08077

Lum K, Boudin C, Price M (2020) The impact of overbooking on a pre-trial risk assessment tool. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '20, pp 482–491. https://doi.org/10.1145/3351095.3372846,

Luong BT, Ruggieri S, Turini F (2016) Classification rule mining supported by ontology for discrimination discovery. In: 2016 IEEE 16th international conference on data mining workshops (ICDMW), pp 868–875. https://doi.org/10.1109/ICDMW.2016.0128

Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C (2011) Learning word vectors for sentiment analysis. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. Association for Computational Linguistics, Portland, pp 142–150. https://www.aclweb.org/anthology/P11-1015

Madnani N, Loukina A, von Davier A, Burstein J, Cahill A (2017) Building better open-source tools to support fairness in automated scoring. In: Proceedings of the first ACL workshop on ethics in natural language processing. Association for Computational Linguistics, Valencia, pp 41–52. https://doi.org/10.18653/v1/W17-1605, https://www.aclweb.org/anthology/W17-1605

Madras D, Creager E, Pitassi T, Zemel R (2018a) Learning adversarially fair and transferable representations. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Proceedings of machine learning research, vol 80, pp 3384–3393. http://proceedings.mlr.press/v80/madras18a.html

Madras D, Pitassi T, Zemel R (2018b) Predict responsibly: Improving fairness and accuracy by learning to defer. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems, vol 31, Curran Associates, Inc., pp 6147–6157. https://proceedings.neurips.cc/paper/2018/file/09d37c08f7b129e96277388757530c72-Paper.pdf

Madras D, Creager E, Pitassi T, Zemel R (2019) Fairness through causal awareness: learning causal latent-variable models for biased data. In: Proceedings of the conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '19, pp 349–358. https://doi.org/10.1145/3287560.3287564

Mahabadi S, Vakilian A (2020) Individual fairness for k-clustering. In: III HD, Singh A (eds) Proceedings of the 37th international conference on machine learning, PMLR, virtual, proceedings of machine learning research, vol 119, pp 6586–6596. http://proceedings.mlr.press/v119/mahabadi20a.html

Maity S, Xue S, Yurochkin M, Sun Y (2021) Statistical inference for individual fairness. In: International conference on learning representations. https://openreview.net/forum?id=z9k8BWL-_2u

Mandal D, Deng S, Jana S, Wing JM, Hsu DJ (2020) Ensuring fairness beyond the training data. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, NeurIPS 2020, December 6–12, 2020, virtual. https://proceedings.neurips.cc/paper/2020/hash/d6539d3b57159bab6a72e106beb45bd-Abstract.html

Manjunatha V, Saini N, Davis LS (2019) Explicit bias discovery in visual question answering models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)

Martinez N, Bertran M, Sapiro G (2020) Minimax pareto fairness: a multi objective perspective. In: III HD, Singh A (eds) Proceedings of the 37th international conference on machine learning, PMLR, virtual, proceedings of machine learning research, vol 119, pp 6755–6764. http://proceedings.mlr.press/v119/martinez20a.html

Mary J, Calauzènes C, Karoui NE (2019) Fairness-aware learning for continuous attributes and treatments. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, proceedings of machine learning research, vol 97, pp 4382–4391. http://proceedings.mlr.press/v97/mary19a.html

Mastrandrea R, Fournet J, Barrat A (2015) Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. PLoS ONE 10(9):e0136497. https://doi.org/10.1371/journal.pone.0136497

Mattei N, Saffidine A, Walsh T (2018a) An axiomatic and empirical analysis of mechanisms for online organ matching. In: Proceedings of the 7th international workshop on computational social choice (COMSOC)

Mattei N, Saffidine A, Walsh T (2018b) Fairness in deceased organ matching. In: Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, AIES '18, pp 236–242. https://doi.org/10.1145/3278721.3278749

Mayson SG (2018) Bias in, bias out. YAle lJ 128:2218

McAuley J, Targett C, Shi Q, van den Hengel A (2015) Image-based recommendations on styles and substitutes. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval. Association for Computing Machinery, New York, SIGIR '15, pp 43–52. https://doi.org/10.1145/2766462.2767755,

McCallum AK, Nigam K, Rennie J, Seymore K (2000) Automating the construction of internet portals with machine learning. Inf Retr 3(2):127–163

McDuff D, Ma S, Song Y, Kapoor A (2019) Characterizing bias in classifiers using generative models. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems, vol 32. Curran Associates, Inc., pp 5403–5414. https://proceedings.neurips.cc/paper/2019/file/7f018eb7b301a66658931cb8a93fd6e8-Paper.pdf

McFee B, Bertin-Mahieux T, Ellis DP, Lanckriet GR (2012) The million song dataset challenge. In: Proceedings of the 21st international conference on world wide web. Association for Computing Machinery, New York, WWW '12 Companion, pp 909–916. https://doi.org/10.1145/2187980.2188222

McKenna L (2019a) A history of the current population survey and disclosure avoidance. https://www2.census.gov/adrm/CED/Papers/FY20/2019-04-McKenna-cps%20and%20da.pdf

McKenna L (2019b) A history of the us census bureau's disclosure review board. https://www2.census.gov/adrm/CED/Papers/FY20/2019-04-McKenna-DRB.pdf

McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA (2017) Communication-efficient learning of deep networks from decentralized data. In: Singh A, Zhu J (eds) Proceedings of the 20th international conference on artificial intelligence and statistics, PMLR, Fort Lauderdale, proceedings of machine learning research, vol 54, pp 1273–1282. http://proceedings.mlr.press/v54/mcmahan17a.html

McNamara D (2019) Equalized odds implies partially equalized outcomes under realistic assumptions. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, AIES '19, pp 313–320. https://doi.org/10.1145/3306618.3314290

Meek C, Thiesson B, Heckerman D (2002) The learning-curve sampling method applied to model-based clustering. J Mach Learn Res 2(Feb):397–418

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. ACM Comput Surv. https://doi.org/10.1145/3457607

Mehrotra A, Celis LE (2021) Mitigating bias in set selection with noisy protected attributes. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAccT '21, pp 237–248. https://doi.org/10.1145/3442188.3445887

Mehrotra R, Anderson A, Diaz F, Sharma A, Wallach H, Yilmaz E (2017) Auditing search engines for differential satisfaction across demographics. In: Proceedings of the 26th international conference on world wide web companion, international world wide web conferences steering committee, Republic and Canton of Geneva, CHE, WWW '17 Companion, pp 626–633. https://doi.org/10.1145/3041021.3054197

Merkley R (2019) Use and fair use: statement on shared images in facial recognition ai

Merler M, Ratha N, Feris RS, Smith JR (2019) Diversity in faces. arXiv:1901.10436

Metevier B, Giguere S, Brockman S, Kobren A, Brun Y, Brunskill E, Thomas PS (2019) Offline contextual bandits with high probability fairness guarantees. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems, vol 32. Curran Associates, Inc., pp 14922–14933. https://proceedings.neurips.cc/paper/2019/file/d69768b3da745b77e82cdbddcc8bac98-Paper.pdf

Mhasawade V, Chunara R (2021) Causal multi-level fairness. Association for Computing Machinery, New York, pp 784–794. https://doi.org/10.1145/3461702.3462587

Miao W (2010) Did the results of promotion exams have a disparate impact on minorities? Using statistical evidence in ricci v. destefano. J Stat Educ 18(3)

Miceli M, Yang T, Naudts L, Schuessler M, Serbanescu D, Hanna A (2021) Documenting computer vision datasets: an invitation to reflexive data practices. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAccT '21, pp 161–172

Miller E (1998) An introduction to the resource description framework. D-lib Magazine

Mirkin S, Nowson S, Brun C, Perez J (2015) Motivating personality-aware machine translation. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Lisbon, pp 1102–1108. https://doi.org/10.18653/v1/D15-1130, https://www.aclweb.org/anthology/D15-1130

Mishler A, Kennedy EH, Chouldechova A (2021) Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAccT '21, pp 386–400. https://doi.org/10.1145/3442188.3445902

Mishra S, He S, Belli L (2020) Assessing demographic bias in named entity recognition. AKBC 2020 workshop: "bias in automatic knowledge graph construction". arXiv:2008.03415

Mislove A, Viswanath B, Gummadi KP, Druschel P (2010) You are who you know: inferring user profiles in online social networks. In: Proceedings of the third ACM international conference on web search and data mining. Association for Computing Machinery, New York, WSDM '10, pp 251–260. https://doi.org/10.1145/1718487.1718519

Moore JC, Stinson LL, Welniak EJ (2000) Income measurement error in surveys: a review. J Off Stat 16(4):331–362

Moreland A, Herlihy C, Tynan MA, Sunshine G, McCord RF, Hilton C, Poovey J, Werner AK, Jones CD, Fulmer EB et al (2020) Timing of state and territorial Covid-19 stay-at-home orders and changes in population movement-United States, March 1–May 31, 2020. Morb Mortal Wkly Rep 69(35):1198

Moro S, Cortez P, Rita P (2014) A data-driven approach to predict the success of bank telemarketing. Decis Support Syst 62:22–31

Mozannar H, Ohannessian M, Srebro N (2020) Fair learning with private demographic data. In: III HD, Singh A (eds) Proceedings of the 37th international conference on machine learning, PMLR, virtual, proceedings of machine learning research, vol 119, pp 7066–7075. http://proceedings.mlr.press/v119/mozannar20a.html

Mukherjee D, Yurochkin M, Banerjee M, Sun Y (2020) Two simple ways to learn individual fairness metrics from data. In: III HD, Singh A (eds) Proceedings of the 37th international conference on machine learning, PMLR, virtual, proceedings of machine learning research, vol 119, pp 7097–7107. http://proceedings.mlr.press/v119/mukherjee20a.html

Muller M, Lange I, Wang D, Piorkowski D, Tsay J, Liao QV, Dugan C, Erickson T (2019) How data science workers work with data: discovery, capture, curation, design, creation. Association for Computing Machinery, New York, pp 1–15. https://doi.org/10.1145/3290605.3300356

Murgia M (2019) Microsoft quietly deletes largest public face recognition data set. https://www.ft.com/content/7d3e0d6a-87a0-11e9-a028-86cea8523dc2

Nabi R, Malinsky D, Shpitser I (2019) Learning optimal fair policies. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, proceedings of machine learning research, vol 97, pp 4674–4682. http://proceedings.mlr.press/v97/nabi19a.html

Namata G, London B, Getoor L, Huang B, EDU U (2012) Query-driven active surveying for collective classification. In: 10th international workshop on mining and learning with graphs, vol 8

Nanda V, Xu P, Sankararaman KA, Dickerson JP, Srinivasan A (2020) Balancing the tradeoff between profit and fairness in rideshare platforms during high-demand hours. In: Proceedings of the AAAI/ACM conference on AI, ethics, and society, association for computing machinery, New York, AIES '20, pp 131. https://doi.org/10.1145/3375627.3375818

Nanda V, Dooley S, Singla S, Feizi S, Dickerson JP (2021) Fairness through robustness: investigating robustness disparity in deep learning. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAccT '21, pp 466–477. https://doi.org/10.1145/3442188.3445910

Narayanan A, Shmatikov V (2008) Robust de-anonymization of large sparse datasets. In: 2008 IEEE symposium on security and privacy (sp 2008). IEEE, pp 111–125

Nasr M, Tschantz MC (2020) Bidding strategies with gender nondiscrimination constraints for online ad auctions. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '20, pp 337–347. https://doi.org/10.1145/3351095.3375783

Ngong IC, Maughan K, Near JP (2020) Towards auditability for fairness in deep learning. NeurIPS 2020 workshop: "algorithmic fairness through the lens of causality and interpretability (AFCI)". arXiv:2012.00106

NLST Trial Research Team (2011) The national lung screening trial: overview and study design. Radiology 258(1):243–253

Noriega-Campero A, Bakker MA, Garcia-Bulle B, Pentland AS (2019) Active fairness in algorithmic decision making. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, AIES '19, pp 77–83. https://doi.org/10.1145/3306618.3314277

Noriega-Campero A, Garcia-Bulle B, Cantu LF, Bakker MA, Tejerina L, Pentland A (2020) Algorithmic targeting of social policies: fairness, accuracy, and distributed governance. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '20, pp 241–251. https://doi.org/10.1145/3351095.3375784

Nuttall DL, Goldstein H, Prosser R, Rasbash J (1989) Differential school effectiveness. Int J Educ Res 13(7):769–776. https://doi.org/10.1016/0883-0355(89)90027-X

Obermeyer Z, Mullainathan S (2019) Dissecting racial bias in an algorithm that guides health decisions for 70 million people. In: Proceedings of the conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '19, p 89. https://doi.org/10.1145/3287560.3287593

Ogura H, Takeda A (2020) Convex fairness constrained model using causal effect estimators. In: Companion proceedings of the web conference 2020. Association for Computing Machinery, New York, WWW '20, pp 723–732. https://doi.org/10.1145/3366424.3383556

Olave M, Rajkovic V, Bohanec M (1989) An application for admission in public school systems. Expert Syst Pub Admin 1:145–160

Oneto L, Siri A, Luria G, Anguita D (2017) Dropout prediction at university of genoa: a privacy preserving data driven approach. In: ESANN

Oneto L, Donini M, Elders A, Pontil M (2019a) Taking advantage of multitask learning for fair classification. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, AIES '19, pp 227–237. https://doi.org/10.1145/3306618.3314255

Oneto L, Donini M, Maurer A, Pontil M (2019b) Learning fair and transferable representations. NeurIPS 2019 workshop: "human-centric machine learning"

Oneto L, Donini M, Luise G, Ciliberto C, Maurer A, Pontil M (2020) Exploiting MMD and sinkhorn divergences for fair and transferable representation learning. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, NeurIPS 2020, December 6–12, 2020, virtual. https://proceedings.neurips.cc/paper/2020/hash/af9c0e0c1dee63e5cad8b7ed1a5be96-Abstract.html

Pandey A, Caliskan A (2021) Disparate impact of artificial intelligence bias in ridehailing economy's price discrimination algorithms. Association for Computing Machinery, New York, pp 822–833. https://doi.org/10.1145/3461702.3462561

Papakyriakopoulos O, Hegelich S, Serrano JCM, Marco F (2020) Bias in word embeddings. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '20, pp 446–457. https://doi.org/10.1145/3351095.3372843

Paraschakis D, Nilsson B (2020) Matchmaking under fairness constraints: a speed dating case study. ECIR 2020 workshop: "international workshop on algorithmic bias in search and recommendation (BIAS 2020)"

Patro GK, Chakraborty A, Ganguly N, Gummadi KP (2019) Incremental fairness in two-sided market platforms: on smoothly updating recommendations. NeurIPS 2019 workshop: "human-centric machine learning". arXiv:1909.10005

Paullada A, Raji ID, Bender EM, Denton E, Hanna A (2020) Data and its (dis) contents: A survey of dataset development and use in machine learning research. arXiv:2012.05345

Pedreshi D, Ruggieri S, Turini F (2008) Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, KDD '08, pp 560–568. https://doi.org/10.1145/1401890.1401959

Peng K, Mathur A, Narayanan A (2021) Mitigating dataset harms requires stewardship: lessons from 1000 papers. arXiv:2108.02922

Perrone V, Donini M, Zafar MB, Schmucker R, Kenthapadi K, Archambeau C (2021) Fair Bayesian optimization, association for computing machinery, New York, pp 854–863. https://doi.org/10.1145/3461702.3462629

Pessach D, Shmueli E (2020) Algorithmic fairness. arXiv:2001.09784

Peters ME, Lecocq D (2013) Content extraction using diverse feature sets. In: Proceedings of the 22nd international conference on world wide web, pp 89–90

Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, vol 1 (long papers). Association for Computational Linguistics, New Orleans, pp 2227–2237

Pfohl S, Marafino B, Coulet A, Rodriguez F, Palaniappan L, Shah NH (2019) Creating fair models of atherosclerotic cardiovascular disease risk. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, AIES '19, pp 271–278. https://doi.org/10.1145/3306618.3314278

Pinard M (2010) Collateral consequences of criminal convictions: confronting issues of race and dignity. NYUL Rev 85:457

Pitoura E, Stefanidis K, Koutrika G (2021) Fairness in rankings and recommendations: an overview. VLDB J 1–28

Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger KQ (2017) On fairness and calibration. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems, vol 30. Curran Associates, Inc., pp 5680–5689. https://proceedings.neurips.cc/paper/2017/file/b8b9c74ac526fffbeb2d39ab038d1cd7-Paper.pdf

Pont-Tuset J, Uijlings J, Changpinyo S, Soricut R, Ferrari V (2020) Connecting vision and language with localized narratives. In: European conference on computer vision. Springer, pp 647–664

Prabhu VU, Birhane A (2020) Large image datasets: a pyrrhic win for computer vision? arXiv:2006.16923

Preoţiuc-Pietro D, Ungar L (2018) User-level race and ethnicity predictors from Twitter text. In: Proceedings of the 27th international conference on computational linguistics. Association for Computational Linguistics, Santa Fe, pp 1534–1545. https://www.aclweb.org/anthology/C18-1130

ProPublica (2016) Compas analysis github repository. https://github.com/propublica/compas-analysis

ProPublica (2021) Propublica data store terms. https://www.propublica.org/datastore/terms

Pujol D, McKenna R, Kuppam S, Hay M, Machanavajjhala A, Miklau G (2020) Fair decision making using privacy-protected data. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '20, pp 189–199. https://doi.org/10.1145/3351095.3372872

Qian S, Cao J, Mouël FL, Sahel I, Li M (2015) Scram: A sharing considered route assignment mechanism for fair taxi route recommendations. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, KDD '15, pp 955–964. https://doi.org/10.1145/2783258.2783261

Qin T, Liu TY (2013) Introducing letor 4.0 datasets. arXiv:1306.2597

Quadrianto N, Sharmanska V (2017) Recycling privileged learning and distribution matching for fairness. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems, vol 30. Curran Associates, Inc., pp 677–688. https://proceedings.neurips.cc/paper/2017/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf

Quadrianto N, Sharmanska V, Thomas O (2019) Discovering fair representations in the data domain. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)

Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners

Radin J (2017) "Digital natives": How medical and indigenous histories matter for big data. Osiris 32(1):43–64

Raff E, Sylvester J (2018) Gradient reversal against discrimination. ICML 2018 workshop: "fairness, accountability, and transparency in machine learning (FAT/ML)". arXiv:1807.00392

Raff E, Sylvester J, Mills S (2018) Fair forests: Regularized tree induction to minimize model bias. In: Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, AIES '18, pp 243–250. https://doi.org/10.1145/3278721.3278742, Rahmattalabi A, Vayanos P, Fulginiti A, Rice E, Wilder B, Yadav A, Tambe M (2019) Exploring algorithmic fairness in robust graph covering problems. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems, vol 32. Curran Associates, Inc., pp 15776–15787. https://proceedings.neurips.cc/paper/2019/file/1d7c2aae840867027b7edd17b6aaa0e9-Paper.pdf

Raj A, Wood C, Montoly A, Ekstrand MD (2020) Comparing fair ranking metrics. RecSys 2020 workshop: "3rd FAccTRec workshop on responsible recommendation". arXiv:2009.01311

Raji ID, Buolamwini J (2019) Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society, association for computing machinery, New York, AIES '19, pp 429–435. https://doi.org/10.1145/3306618.3314244

Ramachandran GS, Brugere I, Varshney LR, Xiong C (2021) GAEA: graph augmentation for equitable access via reinforcement learning. Association for Computing Machinery, New York, pp 884–894. https://doi.org/10.1145/3461702.3462615

Ramaswamy VV, Kim SSY, Russakovsky O (2021) Fair attribute classification through latent space debiasing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 9301–9310

Red V, Kelsic ED, Mucha PJ, Porter MA (2011) Comparing community structure to characteristics in online collegiate social networks. SIAM Rev 53(3):526–543

Redmond M, Baveja A (2002) A data-driven software tool for enabling cooperative information sharing among police departments. Eur J Oper Res 141:660–678. https://doi.org/10.1016/S0377-2217(01)00264-8

Redmond U, Cunningham P (2013) A temporal network analysis reveals the unprofitability of arbitrage in the prosper marketplace. Expert Syst Appl 40(9):3715–3721. https://doi.org/10.1016/j.eswa.2012.12.077

Reed SE, Zhang Y, Zhang Y, Lee H (2015) Deep visual analogy-making. In: Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R (eds) Advances in neural information processing systems, vol 28. Curran Associates, Inc., pp 1252–1260. https://proceedings.neurips.cc/paper/2015/file/e07413354875be01a996dc560274708e-Paper.pdf

Rezaei A, Liu A, Memarrast O, Ziebart B (2021) Robust fairness under covariate shift. NeurIPS 2020 workshop: "algorithmic fairness through the lens of causality and interpretability (AFCI)". arXiv:2010.05166

Riederer C, Chaintreau A (2017) The price of fairness in location based advertising. https://doi.org/10.18122/B2MD8C, recSys 2017 workshop: "workshop on responsible recommendation (FAT/Rec)"

Rocher L, Hendrickx JM, De Montjoye YA (2019) Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun 10(1):1–9

Rodolfa KT, Salomon E, Haynes L, Mendieta IH, Larson J, Ghani R (2020) Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '20, pp 142–153. https://doi.org/10.1145/3351095.3372863

Roh Y, Lee K, Whang S, Suh C (2020) FR-train: a mutual information-based approach to fair and robust training. In: III HD, Singh A (eds) Proceedings of the 37th international conference on machine learning, PMLR, virtual, proceedings of machine learning research, vol 119, pp 8147–8157. http://proceedings.mlr.press/v119/roh20a.html

Roh Y, Lee K, Whang SE, Suh C (2021) Fairbatch: batch selection for model fairness. In: International conference on learning representations. https://openreview.net/forum?id=YNnpaAKeCfx

Romano Y, Bates S, Candes E (2020) Achieving equalized odds by resampling sensitive attributes. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in neural information processing systems, vol 33. Curran Associates, Inc., pp 361–371. https://proceedings.neurips.cc/paper/2020/file/03593ce517feac573fdaafa6dcedef61-Paper.pdf

Romei A, Ruggieri S (2014) A multidisciplinary survey on discrimination analysis. Knowl Eng Rev 29(5):582–638. https://doi.org/10.1017/S0269888913000039

Rotemberg V, Kurtansky N, Betz-Stablein B, Caffery L, Chousakos E, Codella N, Combalia M, Dusza S, Guitera P, Gutman D et al (2021) A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Sci Data 8(1):1–8

Rozemberczki B, Allen C, Sarkar R (2021) Multi-scale attributed node embedding. J Complex Netw 9(2):cnab014

Rudinger R, May C, Van Durme B (2017) Social bias in elicited natural language inferences. In: Proceedings of the first ACL workshop on ethics in natural language processing. Association for Computational Linguistics, pp 74–79. https://doi.org/10.18653/v1/W17-1609, https://www.aclweb.org/anthology/W17-1609

Rudinger R, Naradowsky J, Leonard B, Van Durme B (2018) Gender bias in coreference resolution. In: Proceedings of the 2018 conference of the North American Chapter of the association for computational linguistics: human language technologies, volume 2 (short papers). Association for Computational Linguistics, New Orleans, pp 8–14. https://doi.org/10.18653/v1/N18-2002, https://www.aclweb.org/anthology/N18-2002

Ruoss A, Balunovic M, Fischer M, Vechev M (2020) Learning certified individually fair representations. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in neural information processing systems, vol 33. Curran Associates, Inc., pp 7584–7596. https://proceedings.neurips.cc/paper/2020/file/55d491cf951b1b920900684d71419282-Paper.pdf

Russell C, Kusner MJ, Loftus J, Silva R (2017) When worlds collide: integrating different counterfactual assumptions in fairness. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems, vol 30. Curran Associates, Inc., pp 6414–6423. https://proceedings.neurips.cc/paper/2017/file/1271a7029c9df08643b631b02cf9e116-Paper.pdf

Sabato S, Yom-Tov E (2020) Bounding the fairness and accuracy of classifiers from population statistics. In: III HD, Singh A (eds) Proceedings of the 37th international conference on machine learning, PMLR, virtual, proceedings of machine learning research, vol 119, pp 8316–8325. http://proceedings.mlr.press/v119/sabato20a.html

Saenko K, Kulis B, Fritz M, Darrell T (2010) Adapting visual category models to new domains. In: Proceedings of the 11th European conference on computer vision: part IV. Springer, Berlin, Heidelberg, ECCV'10, pp 213–226

Sagawa S, Koh PW, Hashimoto TB, Liang P (2020) Distributionally robust neural networks. In: International conference on learning representations. https://openreview.net/forum?id=ryxGuJrFvS

Samadi S, Tantipongpipat U, Morgenstern JH, Singh M, Vempala S (2018) The price of fair pca: one extra dimension. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems, vol 31. Curran Associates, Inc., pp 10976–10987. https://proceedings.neurips.cc/paper/2018/file/cc4af25fa9d2d5c953496579b75f6f6c-Paper.pdf

Savani Y, White C, Govindarajulu NS (2020) Intra-processing methods for debiasing neural networks. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, NeurIPS 2020, December 6–12, 2020, virtual. https://proceedings.neurips.cc/paper/2020/hash/1d8d70dddf147d2d92a634817f01b239-Abstract.html

Scheuerman MK, Wade K, Lustig C, Brubaker JR (2020) How we've taught algorithms to see identity: constructing race and gender in image databases for facial analysis. In: Proceedings of the ACM human–computer interaction 4(CSCW1). https://doi.org/10.1145/3392866

Schumann C, Ricco S, Prabhu U, Ferrari V, Pantofaru C (2021) A step toward more inclusive people annotations for fairness. Association for Computing Machinery, New York, pp 916-925. https://doi.org/10.1145/3461702.3462594

Schutzman Z (2020) Trade-offs in fair redistricting. In: Proceedings of the AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, AIES '20, pp 159–165. https://doi.org/10.1145/3375627.3375802

Segal S, Adi Y, Pinkas B, Baum C, Ganesh C, Keshet J (2021) Fairness in the eyes of the data: certifying machine-learning models. Association for Computing Machinery, New York, pp 926–935. https://doi.org/10.1145/3461702.3462554

Sen P, Namata G, Bilgic M, Getoor L, Galligher B, Eliassi-Rad T (2008) Collective classification in network data. AI Mag 29(3):93–93

Shah K, Gupta P, Deshpande A, Bhattacharyya C (2021) Rawlsian fair adaptation of deep learning classifiers. Association for Computing Machinery, New York, pp 936–945. https://doi.org/10.1145/3461702.3462592

Shang J, Sun M, Lam NS (2020) List-wise fairness criterion for point processes. Association for Computing Machinery, New York, pp 1948–1958. https://doi.org/10.1145/3394486.3403246

Sharifi-Malvajerdi S, Kearns M, Roth A (2019) Average individual fairness: algorithms, generalization and experiments. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds)

Advances in neural information processing systems, vol 32. Curran Associates, Inc., pp 8242–8251. https://proceedings.neurips.cc/paper/2019/file/0e1feae55e360ff05fef58199b3fa521-Paper.pdf

Sharma S, Henderson J, Ghosh J (2020a) Certifai: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In: Proceedings of the AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, AIES '20, pp 166–172. https://doi.org/10.1145/3375627.3375812

Sharma S, Zhang Y, Ríos Aliaga JM, Bouneffouf D, Muthusamy V, Varshney KR (2020b) Data augmentation for discrimination prevention and bias disambiguation. In: Proceedings of the AAAI/ACM conference on AI, ethics, and society, Association for Computing Machinery, New York, AIES '20, pp 358–364. https://doi.org/10.1145/3375627.3375865

Sharma S, Gee AH, Paydarfar D, Ghosh J (2021) FaiR-N: fair and robust neural networks for structured data. Association for Computing Machinery, New York, pp 946–955. https://doi.org/10.1145/3461702.3462559

Shekhar S, Shah N, Akoglu L (2021) Fairod: fairness-aware outlier detection. In: Proceedings of the 2021 AAAI/ACM conference on ai, ethics, and society. Association for Computing Machinery, New York, AIES '21, pp 210–220. https://doi.org/10.1145/3461702.3462517

Shen JH, Fratamico L, Rahwan I, Rush AM (2018) Darling or babygirl? investigating stylistic bias in sentiment analysis. KDD 2018 workshop: "fairness, accountability, and transparency in machine learning (FAT/ML)"

Shermis MD (2014) State-of-the-art automated essay scoring: competition, results, and future directions from a united states demonstration. Assess Writ 20:53–76. https://doi.org/10.1016/j.asw.2013.04.001

Singh A, Joachims T (2018) Fairness of exposure in rankings. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, KDD '18, pp 2219–2228. https://doi.org/10.1145/3219819.3220088

Singh A, Joachims T (2019) Policy learning for fairness in ranking. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems, vol 32. Curran Associates, Inc., pp 5426–5436. https://proceedings.neurips.cc/paper/2019/file/9e82757e9a1c12cb710ad680db11f6f1-Paper.pdf

Singh M, Ramamurthy KN (2019) Understanding racial bias in health using the medical expenditure panel survey data. NeurIPS 2019 workshop: "Fair ML for health". arXiv:1911.01509

Singh H, Singh R, Mhasawade V, Chunara R (2021) Fairness violations and mitigation under covariate shift. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAccT '21, pp 3–13. https://doi.org/10.1145/3442188.3445865

Slack D, Friedler S, Givental E (2019a) Fair meta-learning: learning how to learn fairly. https://drive.google.com/file/d/1F5YF1Ar1hJ7l2H7zIsC35SzXOWqUylVW/view, neurIPS 2019 workshop: "human-centric machine learning"

Slack D, Friedler S, Givental E (2019b) Fairness warnings. https://drive.google.com/file/d/1eeu703ulWkehk0WEepYDwXg2KXSwOzc2/view. neurIPS 2019 workshop: "human-centric machine learning"

Slack D, Friedler SA, Givental E (2020) Fairness warnings and fair-maml: Learning fairly with minimal data. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '20, pp 200–209. https://doi.org/10.1145/3351095.3372839

Slunge D (2015) The willingness to pay for vaccination against tick-borne encephalitis and implications for public health policy: evidence from Sweden. PLoS ONE 10(12):e0143875

Smith JW, Everhart J, Dickson W, Knowler W, Johannes R (1988) Using the adap learning algorithm to forecast the onset of diabetes mellitus. In: Proceedings symposium on computer applications in medical care, pp 261–265. https://europepmc.org/articles/PMC2245318

Solans D, Fabbri F, Calsamiglia C, Castillo C, Bonchi F (2021) Comparing equity and effectiveness of different algorithms in an application for the room rental market. Association for Computing Machinery, New York, pp 978–988. https://doi.org/10.1145/3461702.3462600

Sonboli N, Burke R (2019) Localized fairness in recommender systems. In: Adjunct publication of the 27th conference on user modeling, adaptation and personalization. Association for Computing Machinery, New York, UMAP'19 Adjunct, pp 295–300. https://doi.org/10.1145/3314183.3323845

Sonboli N, Burke R, Mattei N, Eskandanian F, Gao T (2020) "And the winner is...": dynamic lotteries for multi-group fairness-aware recommendation. RecSys 2020 workshop: "3rd FAccTRec workshop on responsible recommendation". arXiv:2009.02590

Speakman S, Sridharan S, Markus I (2018) Three population covariate shift for mobile phone-based credit scoring. In: Proceedings of the 1st ACM SIGCAS conference on computing and sustainable societies. Association for Computing Machinery, New York, NY, USA, COMPASS '18

Speicher T, Ali M, Venkatadri G, Ribeiro FN, Arvanitakis G, Benevenuto F, Gummadi KP, Loiseau P, Mislove A (2018a) Potential for discrimination in online targeted advertising. In: Friedler SA, Wilson C (eds) Proceedings of the 1st conference on fairness, accountability and transparency, PMLR, New York, proceedings of machine learning research, vol 81, pp 5–19. http://proceedings.mlr.press/v81/speicher18a.html

Speicher T, Heidari H, Grgic-Hlaca N, Gummadi KP, Singla A, Weller A, Zafar MB (2018b) A unified approach to quantifying algorithmic unfairness: measuring individual and group unfairness via inequality indices. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery anddata mining. Association for Computing Machinery, New York, KDD '18, pp 2239–2248. https://doi.org/10.1145/3219819.3220046

Squire RF (2019) Measuring and correcting sampling bias in safegraph patterns for more accurate demographic analysis. https://www.safegraph.com/blog/measuring-and-correcting-sampling-bias-for-accurate-demographic-analysis/?utm_source=content&utm_medium=referral&utm_campaign=colabnotebook&utm_content=panel_bias

Stanovsky G, Smith NA, Zettlemoyer L (2019) Evaluating gender bias in machine translation. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Florence, pp 1679–1684. https://doi.org/10.18653/v1/P19-1164, https://aclanthology.org/P19-1164

Steed R, Caliskan A (2021) Image representations learned with unsupervised pre-training contain human-like biases. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAccT '21, pp 701–713. https://doi.org/10.1145/3442188.3445932

Strack B, Deshazo J, Gennings C, Olmo Ortiz JL, Ventura S, Cios K, Clore J (2014) Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. Biomed Res Int 2014:781670. https://doi.org/10.1155/2014/781670

Sühr T, Biega AJ, Zehlike M, Gummadi KP, Chakraborty A (2019) Two-sided fairness for repeated matchings in two-sided markets: a case study of a ride-hailing platform. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, KDD '19, pp 3082–3092. https://doi.org/10.1145/3292500.3330793

Sühr T, Hilgard S, Lakkaraju H (2021) Does fair ranking improve minority outcomes? Understanding the interplay of human and algorithmic biases in online hiring. Association for Computing Machinery, New York, pp 989–999. https://doi.org/10.1145/3461702.3462602

Sun Y, Han J, Gao J, Yu Y (2009) itopicmodel: Information network-integrated topic modeling. In: 2009 Ninth IEEE international conference on data mining, pp 493–502. https://doi.org/10.1109/ICDM.2009.43

Sun T, Gaut A, Tang S, Huang Y, ElSherief M, Zhao J, Mirza D, Belding E, Chang KW, Wang WY (2019) Mitigating gender bias in natural language processing: literature review. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Florence, pp 1630–1640. https://doi.org/10.18653/v1/P19-1159, https://aclanthology.org/P19-1159

Swinger N, De-Arteaga M, Heffernan IV NT, Leiserson MD, Kalai AT (2019) What are the biases in my word embedding? In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, AIES '19, pp 305–311. https://doi.org/10.1145/3306618.3314270

Takac L, Zabovsky M (2012) Data analysis in public social networks. In: International scientific conference and international workshop present day trends of innovations, vol 1

Tan YC, Celis LE (2019) Assessing social and intersectional biases in contextualized word representations. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems, vol 32. Curran Associates, Inc., pp 13230–13241. https://proceedings.neurips.cc/paper/2019/file/201d546992726352471cfea6b0df0a48-Paper.pdf

Tang J, Zhang J, Yao L, Li J, Zhang l, Su Z (2008) Arnetminer: extraction and mining of academic social networks. pp 990–998. https://doi.org/10.1145/1401890.1402008

Tantipongpipat U, Samadi S, Singh M, Morgenstern JH, Vempala S (2019) Multi-criteria dimensionality reduction with applications to fairness. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems, vol 32. Curran Associates, Inc., pp 15161–15171. https://proceedings.neurips.cc/paper/2019/file/2201611d7a08ffda97e3e8c6b667a1bc-Paper.pdf

Taskesen B, Blanchet J, Kuhn D, Nguyen VA (2021) A statistical test for probabilistic fairness. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAccT '21, pp 648–665. https://doi.org/10.1145/3442188.3445927

Tatman R (2017) Gender and dialect bias in YouTube's automatic captions. In: Proceedings of the First ACL workshop on ethics in natural language processing. Association for Computational Linguistics, Valencia, pp 53–59. https://doi.org/10.18653/v1/W17-1606, https://www.aclweb.org/anthology/W17-1606

Tavallaee M, Bagheri E, Lu W, Ghorbani AA (2009) A detailed analysis of the kdd cup 99 data set. In: 2009 IEEE symposium on computational intelligence for security and defense applications, pp 1–6. https://doi.org/10.1109/CISDA.2009.5356528

Team Conduent Public Safety Solutions (2018) Real time crime forecasting challenge: post-mortem analysis challenge performance

Tjong Kim Sang EF, De Meulder F (2003) Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In: Proceedings of the seventh conference on natural language learning at HLT-NAACL 2003, pp 142–147. https://www.aclweb.org/anthology/W03-0419

Tong S, Kagal L (2020) Investigating bias in image classification using model explanations. ICML 2020 workshop: "workshop on human interpretability in machine learning (WHI)". arXiv:2012.05463

Toutanova K, Chen D, Pantel P, Poon H, Choudhury P, Gamon M (2015) Representing text for joint embedding of text and knowledge bases. https://doi.org/10.18653/v1/D15-1174

Tsang A, Wilder B, Rice E, Tambe M, Zick Y (2019) Group-fairness in influence maximization. In: International joint conference on artificial intelligence

Tsao CW, Vasan RS (2015) Cohort profile: the framingham heart study (fhs): overview of milestones in cardiovascular epidemiology. Int J Epidemiol 44(6):1800–1813

Tschandl P, Rosendahl C, Kittler H (2018) The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 5(1):1–9

Tziavelis N, Giannakopoulos I, Doka K, Koziris N, Karras P (2019) Equitable stable matchings in quadratic time. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems, , vol 32. Curran Associates, Inc., pp 457–467. https://proceedings.neurips.cc/paper/2019/file/cb70ab375662576bd1ac5aaf16b3fca4-Paper.pdf

UCI machine learning repository (1994) Statlog (german credit data) data set. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

UCI Machine Learning Repository (1996) Adult data set. https://archive.ics.uci.edu/ml/datasets/adult

UCI Machine Learning Repository (2019) South German credit data set. https://archive.ics.uci.edu/ml/datasets/South+German+Credit

US Dept of Commerce Bureau of the Census (1978) The current population survey: design and methodology

US Dept of Commerce Bureau of the Census (1995) Current population survey: Annual demographic file, 1994

US Federal Reserve (2007) Report to the congress on credit scoring and its effects on the availability and affordability of credi

Ustun B, Westover MB, Rudin C, Bianchi MT (2016) Clinical prediction models for sleep apnea: the importance of medical history over symptoms. J Clin Sleep Med 12(02):161–168. https://doi.org/10.5664/jcsm.5476

Ustun B, Liu Y, Parkes D (2019) Fairness without harm: decoupled classifiers with preference guarantees. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, proceedings of machine learning research, vol 97, pp 6373–6382. http://proceedings.mlr.press/v97/ustun19a.html

VE S, Cho Y (2020) A rule-based model for seoul bike sharing demand prediction using weather data. Eur J Remote Sens 53(sup1):166–183. https://doi.org/10.1080/22797254.2020.1725789

V E S, Park J, Cho Y, (2020) Using data mining techniques for bike sharing demand prediction in metropolitan city. Comput Commun 153:353–366. https://doi.org/10.1016/j.comcom.2020.02.007

Vaithianathan R, Putnam-Hornstein E, Jiang N, Nand P, Maloney T (2017) Developing predictive models to support child maltreatment hotline screening decisions: allegheny county methodology and implementation. https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-2.pdf

Valera I, Singla A, Gomez Rodriguez M (2018) Enhancing the accuracy and fairness of human decision making. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems, Curran Associates, Inc., vol 31, pp 1769–1778. https://proceedings.neurips.cc/paper/2018/file/0a113ef6b61820daa5611c870ed8d5ee-Paper.pdf

Van Horn G, Mac Aodha O, Song Y, Cui Y, Sun C, Shepard A, Adam H, Perona P, Belongie S (2018) The inaturalist species classification and detection dataset. arXiv:1707.06642

Van Horn G, Cole E, Beery S, Wilber K, Belongie S, Mac Aodha O (2021) Benchmarking representation learning for natural world image collections. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12884–12893

Vargo A, Zhang F, Yurochkin M, Sun Y (2021) Individually fair gradient boosting. In: international conference on learning representations. https://openreview.net/forum?id=JBAa9we1AL

Vig J, Gehrmann S, Belinkov Y, Qian S, Nevo D, Singer Y, Shieber SM (2020) Investigating gender bias in language models using causal mediation analysis. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, NeurIPS 2020, December 6–12, 2020, virtual. https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html

Vijayaraghavan P, Vosoughi S, Roy D (2017) Twitter demographic classification using deep multi-modal multi-task learning. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: short papers). Association for Computational Linguistics, Vancouver, pp 478–483. https://doi.org/10.18653/v1/P17-2076, https://www.aclweb.org/anthology/P17-2076

Voigt R, Jurgens D, Prabhakaran V, Jurafsky D, Tsvetkov Y (2018) RtGender: a corpus for studying differential responses to gender. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki. https://www.aclweb.org/anthology/L18-1445

Voorhees E (2005) Overview of the trec 2005 robust retrieval track. https://trec.nist.gov/pubs/trec13/papers/ROBUST.OVERVIEW.pdf

Wadsworth C, Vera F, Piech C (2018) Achieving fairness through adversarial learning: an application to recidivism prediction. ICML 2018 workshop: "fairness, accountability, and transparency in machine learning (FAT/ML)", arXiv:1807.00199

Wah C, Branson S, Welinder P, Perona P, Belongie S (2011) The caltech-ucsd birds200-2011 dataset. Advances in Water Resources-ADV WATER RESOUR

Wan M, McAuley J (2018) Item recommendation on monotonic behavior chains. In: Proceedings of the 12th ACM conference on recommender systems. Association for Computing Machinery, New York, RecSys '18, pp 86–94. https://doi.org/10.1145/3240323.3240369

Wang M, Deng W (2020) Mitigating bias in face recognition using skewness-aware reinforcement learning. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR)

Wang T, Saar-Tsechansky M (2020) Augmented fairness: an interpretable model augmenting decision-makers' fairness. NeurIPS 2020 workshop: "algorithmic fairness through the lens of causality and interpretability (AFCI)". arXiv:2011.08398

Wang A, Pruksachatkun Y, Nangia N, Singh A, Michael J, Hill F, Levy O, Bowman S (2019a) Superglue: a stickier benchmark for general-purpose language understanding systems. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems, vol 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf

Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR (2019b) GLUE: A multi-task benchmark and analysis platform for natural language understanding. In: International conference on learning representations. https://openreview.net/forum?id=rJ4km2R5t7

Wang H, Grgic-Hlaca N, Lahoti P, Gummadi KP, Weller A (2019c) An empirical study on learning fairness metrics for compas data with human supervision. NeurIPS 2019 workshop: "human-centric machine learning", arXiv:1910.10255

Wang H, Ustun B, Calmon F (2019d) Repairing without retraining: avoiding disparate impact with counterfactual distributions. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, proceedings of machine learning research, vol 97, pp 6618–6627. http://proceedings.mlr.press/v97/wang19l.html

Wang M, Deng W, Hu J, Tao X, Huang Y (2019e) Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 692–702

Wang S, Guo W, Narasimhan H, Cotter A, Gupta M, Jordan M (2020a) Robust optimization for fairness with noisy protected groups. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in neural information processing systems, vol 33. Curran Associates, Inc., pp 5190–5203. https://proceedings.neurips.cc/paper/2020/file/37d097caf1299d9aa79c2c2b843d2d78-Paper.pdf

Wang Z, Qinami K, Karakozis IC, Genova K, Nair P, Hata K, Russakovsky O (2020b) Towards fairness in visual recognition: effective strategies for bias mitigation. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR)

Wang J, Liu Y, Levy C (2021) Fair classification with group-dependent label noise. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAccT '21, pp 526–536. https://doi.org/10.1145/3442188.3445915,

Waseem Z, Hovy D (2016) Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In: Proceedings of the NAACL student research workshop. Association for Computational Linguistics, San Diego, pp 88–93. https://doi.org/10.18653/v1/N16-2013, https://www.aclweb.org/anthology/N16-2013

Webster K, Recasens M, Axelrod V, Baldridge J (2018) Mind the gap: a balanced corpus of gendered ambiguous pronouns. arXiv:1810.05201

Weeks M, Clair S, Borgatti S, Radda K, Schensul J (2002) Social networks of drug users in high-risk sites: finding the connections. AIDS Behav 6:193–206. https://doi.org/10.1023/A:1015457400897

Wick M, panda s, Tristan JB (2019) Unlocking fairness: a trade-off revisited. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems, Curran Associates, Inc., vol 32, pp 8783–8792. https://proceedings.neurips.cc/paper/2019/file/373e4c5d8edfa8b74fd4b6791d0cf6dc-Paper.pdf

Wieringa J, Kannan P, Ma X, Reutterer T, Risselada H, Skiera B (2021) Data analytics in a privacy-concerned world. J Bus Res 122:915–925. https://doi.org/10.1016/j.jbusres.2019.05.005

Wightman L, Ramsey H, Council LSA (1998) LSAC National Longitudinal Bar Passage Study. LSAC research report series, Law School Admission Council. https://books.google.it/books?id=WdA7AQAAIAAJ

Wilder B, Ou HC, de la Haye K, Tambe M (2018) Optimizing network structure for preventative health. In: Proceedings of the 17th international conference on autonomous agents and multiagent systems. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '18, pp 841–849

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE et al (2016) The fair guiding principles for scientific data management and stewardship. Sci Data 3(1):1–9

Williams JV, Razavian N (2019) Quantification of bias in machine learning for healthcare: a case study of renal failure prediction. https://drive.google.com/file/d/1dvJfvVLIQVeeKaLrMlXfX6lcVTzhkDQ0/view, neurIPS 2019 workshop: "Fair ML for Health"

Williamson R, Menon A (2019) Fairness risk measures. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, proceedings of machine learning research, vol 97, pp 6786–6797. http://proceedings.mlr.press/v97/williamson19a.html

Wilson C, Ghosh A, Jiang S, Mislove A, Baker L, Szary J, Trindel K, Polli F (2021) tbuilding and auditing fair algorithms: a case study in candidate screening. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAccT '21, pp 666–677. https://doi.org/10.1145/3442188.3445928

Wondracek G, Holz T, Kirda E, Kruegel C (2010) A practical attack to de-anonymize social network users. In: 2010 IEEE symposium on security and privacy, pp 223–238. https://doi.org/10.1109/SP.2010.21

Wu Y, Zhang L, Wu X (2018) On discrimination discovery and removal in ranked data using causal graph. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, KDD '18, pp 2536–2544. https://doi.org/10.1145/3219819.3220087

Wu Y, Zhang L, Wu X, Tong H (2019) Pc-fairness: A unified framework for measuring causality-based fairness. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems, vol 32. Curran Associates, Inc., pp 3404–3414. https://proceedings.neurips.cc/paper/2019/file/44a2e0804995faf8d2e3b084a1e2db1d-Paper.pdf

Xiao H, Rasul K, Vollgraf R (2017) Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747

Xiao W, Zhao H, Pan H, Song Y, Zheng VW, Yang Q (2019) Beyond personalization: Social content recommendation for creator equality and consumer satisfaction. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, KDD '19, pp 235–245. https://doi.org/10.1145/3292500.3330965,

Xie M, Lauritsen JL (2012) Racial context and crime reporting: a test of Black's stratification hypothesis. J Quant Criminol 28(2):265–293

Xu D, Yuan S, Zhang L, Wu X (2018) Fairgan: Fairness-aware generative adversarial networks. In: 2018 IEEE international conference on big data (big data). IEEE, pp 570–575

Xu R, Cui P, Kuang K, Li B, Zhou L, Shen Z, Cui W (2020) Algorithmic decision making with conditional fairness. Association for Computing Machinery, New York, pp 2125–2135. https://doi.org/10.1145/3394486.3403263

Xu X, Huang Y, Shen P, Li S, Li J, Huang F, Li Y, Cui Z (2021) Consistent instance false positive improves fairness in face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 578–586

Yang K, Stoyanovich J (2017) Measuring fairness in ranked outputs. In: Proceedings of the 29th international conference on scientific and statistical database management. Association for Computing Machinery, New York, SSDBM '17. https://doi.org/10.1145/3085504.3085526

Yang M, Kim B (2019) Benchmarking attribution methods with relative feature importance. arXiv:1907.09701

Yang F, Cisse M, Koyejo S (2020a) Fairness with overlapping groups; a probabilistic perspective. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in neural information processing systems, vol 33. Curran Associates, Inc., pp 4067–4078. https://proceedings.neurips.cc/paper/2020/file/29c0605a3bab4229e46723f89cf59d83-Paper.pdf

Yang K, Qinami K, Fei-Fei L, Deng J, Russakovsky O (2020b) Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAT* '20, pp 547–558. https://doi.org/10.1145/3351095.3375709

Yao S, Huang B (2017a) Beyond parity: fairness objectives for collaborative filtering. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems, vol 30. Curran Associates, Inc., pp 2921–2930. https://proceedings.neurips.cc/paper/2017/file/e6384711491713d29bc63fc5eeb5ba4f-Paper.pdf

Yao S, Huang B (2017b) New fairness metrics for recommendation that embrace differences. KDD 2017 workshop: "fairness, accountability, and transparency in machine learning (FAT/ML)". arXiv:1706.09838

Yeh IC, Hui Lien C (2009) The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Syst Appl 36(2, Part 1):2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020

Yi S, Xiaogang W, Xiaoou T (2013) Deep convolutional network cascade for facial point detection. In: 2013 IEEE conference on computer vision and pattern recognition, pp 3476–3483. https://doi.org/10.1109/CVPR.2013.446

Yi S, Wang S, Joshi S, Ghassemi M (2019) Fair and robust treatment effect estimates: estimation under treatment and outcome disparity with deep neural models. https://drive.google.com/file/d/1hUHRovnfzxnPaselTczzuQfvGU9jbTI1/view, neurIPS 2019 workshop: "Fair ML for Health"

Yurochkin M, Sun Y (2021) Sensei: sensitive set invariance for enforcing individual fairness. In: International conference on learning representations. https://openreview.net/forum?id=DktZb97_Fx

Yurochkin M, Bower A, Sun Y (2020) Training individually fair ml models with sensitive subspace robustness. In: International conference on learning representations. https://openreview.net/forum?id=B1gdkxHFDH

Zafar MB, Valera I, Gomez Rodriguez M, Gummadi KP (2017a) Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th international conference on world wide web, international world wide web conferences steering committee, Republic and Canton of Geneva, CHE, WWW '17, pp 1171–1180. https://doi.org/10.1145/3038912.3052660

Zafar MB, Valera I, Rodriguez M, Gummadi K, Weller A (2017b) From parity to preference-based notions of fairness in classification. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems, vol 30. Curran Associates, Inc., pp 229–239. https://proceedings.neurips.cc/paper/2017/file/82161242827b703e6acf9c726942a1e4-Paper.pdf

Zafar MB, Valera I, Rogriguez MG, Gummadi KP (2017c) Fairness constraints: mechanisms for fair classification. In: Artificial intelligence and statistics, PMLR, pp 962–970

Zehlike M, Yang K, Stoyanovich J (2021) Fairness in ranking: a survey. arXiv:2103.14000

Zhang Y (2005) Bayesian graphical model for adaptive information filtering. PhD thesis, Carnegie Mellon University

Zhang J, Bareinboim E (2018) Equality of opportunity in classification: a causal approach. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems, vol 31. Curran Associates, Inc., pp 3671–3681. https://proceedings.neurips.cc/paper/2018/file/ff1418e8cc993fe8abcfe3ce2003e5c5-Paper.pdf

Zhang H, Davidson I (2021) Towards fair deep anomaly detection. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, FAccT '21, pp 138–148. https://doi.org/10.1145/3442188.3445878

Zhang Z, Neill DB (2017) Identifying significant predictive bias in classifiers. KDD 2017 workshop: "fairness, accountability, and transparency in machine learning (FAT/ML)". arXiv:1611.08292

Zhang Z, Luo P, Loy CC, Tang X (2014). Facial landmark detection by deep multi-task learning. https://doi.org/10.1007/978-3-319-10599-4_7

Zhang Z, Luo P, Loy CC, Tang X (2015) Learning deep representation for face alignment with auxiliary attributes. IEEE Trans Pattern Anal Mach Intell 38(5):918–930

Zhang L, Wu Y, Wu X (2017a) Achieving non-discrimination in data release. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, KDD '17, pp 13350–1344. https://doi.org/10.1145/3097983.3098167

Zhang Z, Song Y, Qi H (2017b) Age progression/regression by conditional adversarial autoencoder. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)

Zhang BH, Lemoine B, Mitchell M (2018) Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, AIES '18, pp 335–340. https://doi.org/10.1145/3278721.3278779

Zhang X, Khaliligarekani M, Tekin C, liu m (2019) Group retention when using machine learning in sequential decision making: the interplay between user dynamics and fairness. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems, vol 32. Curran Associates, Inc., pp 15269–15278. https://proceedings.neurips.cc/paper/2019/file/7690dd4db7a92524c684e3191919eb6b-Paper.pdf

Zhang H, Lu AX, Abdalla M, McDermott M, Ghassemi M (2020a) Hurtful words: quantifying biases in clinical contextual word embeddings. In: Proceedings of the ACM conference on health, inference, and learning, association for computing machinery, New York, CHIL '20, pp 110–120. https://doi.org/10.1145/3368555.3384448

Zhang X, Tu R, Liu Y, Liu M, Kjellström H, Zhang K, Zhang C (2020b) How do fair decisions fare in long-term qualification? In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, NeurIPS 2020, December 6–12, 2020, virtual, https://proceedings.neurips.cc/paper/2020/hash/d6d231705f96d5a3aeb3a76402e49a3-Abstract.html

Zhang Y, Bellamy R, Varshney K (2020c) Joint optimization of ai fairness and utility: a human-centered approach. In: Proceedings of the AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, AIES '20, pp 400–406. https://doi.org/10.1145/3375627.3375862

Zhao H, Gordon G (2019) Inherent tradeoffs in learning fair representations. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems, vol 32. Curran Associates, Inc., pp 15675–15685. https://proceedings.neurips.cc/paper/2019/file/b4189d9de0fb2b9cce090bd1a15e3420-Paper.pdf

Zhao J, Wang T, Yatskar M, Ordonez V, Chang KW (2017) Men also like shopping: reducing gender bias amplification using corpus-level constraints. In: Proceedings of the 2017 conference on empirical methods in natural language processing, association for computational linguistics, Copenhagen, Denmark, pp 2979–2989. https://doi.org/10.18653/v1/D17-1323, https://www.aclweb.org/anthology/D17-1323

Zhao J, Wang T, Yatskar M, Ordonez V, Chang KW (2018) Gender bias in coreference resolution: Evaluation and debiasing methods. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 2 (short papers). Association for Computational Linguistics, New Orleans, pp 15–20. https://doi.org/10.18653/v1/N18-2003, https://www.aclweb.org/anthology/N18-2003

Zhao B, Xiao X, Gan G, Zhang B, Xia ST (2020a) Maintaining discrimination and fairness in class incremental learning. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR)

Zhao C, Li C, Li J, Chen F (2020b) Fair meta-learning for few-shot classification. In: 2020 IEEE international conference on knowledge graph (ICKG), pp 275–282. https://doi.org/10.1109/ICBK50248.2020.00047

Zhao H, Coston A, Adel T, Gordon GJ (2020c) Conditional learning of fair representations. In: International conference on learning representations. https://openreview.net/forum?id=Hkekl0NFPr

Zhao Y, Kong S, Fowlkes C (2021) Camera pose matters: Improving depth prediction by mitigating pose distribution bias. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 15759–15768

Zheng Y, Dave T, Mishra N, Kumar H (2018) Fairness in reciprocal recommendations: a speed-dating study. In: Adjunct publication of the 26th conference on user modeling, adaptation and personalization. Association for Computing Machinery, New York, UMAP '18, pp 29–34. https://doi.org/10.1145/3213586.3226207

Zhong Y, Deng W, Wang M, Hu J, Peng J, Tao X, Huang Y (2019) Unequal-training for deep face recognition with long-tailed noisy data. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition (CVPR)

Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A (2018) Places: a 10 million image database for scene recognition. IEEE Trans Pattern Anal Mach Intell 40(6):1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009

Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, Fidler S (2015) Aligning books and movies: towards story-like visual explanations by watching movies and reading books. arXiv:1506.06724

Zhu Z, Wang J, Zhang Y, Caverlee J (2018) Fairness-aware recommendation of information curators. RecSys 2018 workshop: "workshop on responsible recommendation (FAT/Rec)", arXiv:1809.03040

Žliobaité I (2015) On the relation between accuracy and fairness in binary classification. ICML 2015 workshop: "fairness, accountability, and transparency in machine learning (FAT/ML)". arXiv:1505.05723

Žliobaité I, Kamiran F, Calders T (2011) Handling conditional discrimination. In: 2011 IEEE 11th international conference on data mining, pp 992–1001. https://doi.org/10.1109/ICDM.2011.72

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.