Fares on Fairness: Using a Total Error Framework to Examine the Role of Measurement and Representation in Training Data on Model Fairness and Bias

PATRICK OLIVER SCHENK, Department of Statistics, LMU Munich, and Munich Center for Machine Learning (MCML), Germany

CHRISTOPH KERN, Department of Statistics, LMU Munich, and MCML, Germany

TRENT D. BUSKIRK, Old Dominion University, United States

Data-driven decisions, often based on predictions from machine learning (ML) models are becoming ubiquitous. For these decisions to be just, the underlying ML models must be fair, i.e., work equally well for all parts of the population such as groups defined by gender or age. What are the logical next steps if, however, a trained model is accurate but not fair? How can we guide the whole data pipeline such that we avoid training unfair models based on inadequate data, recognizing possible sources of unfairness early on? How can the concepts of data-based sources of unfairness that exist in the fair ML literature be organized, perhaps in a way to gain new insight? In this paper, we explore two total error frameworks from the social sciences, *Total Survey Error* and its generalization *Total Data Quality*, to help elucidate issues related to fairness and trace its antecedents. The goal of this thought piece is to acquaint the fair ML community with these two frameworks, discussing errors of measurement and errors of representation through their organized structure. We illustrate how they may be useful, both practically and conceptually.

CCS Concepts: • **General and reference** → **Surveys and overviews**; **Empirical studies**; **Evaluation**; **Measurement**; **Metrics**; Reliability; • **Information systems** → **Data provenance**; **Incomplete data**; **Inconsistent data**; **Uncertainty**; • **Mathematics of computing** → *Probability and statistics*; *Statistical paradigms*; • **Computing methodologies** → *Machine learning*; • **Applied computing**; • **Social and professional topics**;

Keywords: Algorithmic Fairness, Representation, Measurement, Training Data Quality, Data Quality Frameworks, Data-centric Machine Learning, Fair Machine Learning, Survey Methodology, Total Survey Error, Total Data Quality

Reference Format:
Patrick Oliver Schenk, Christoph Kern, and Trent D. Buskirk. 2025. Fares on Fairness: Using a Total Error Framework to Examine the Role of Measurement and Representation in Training Data on Model Fairness and Bias. In *Proceedings of Fourth European Workshop on Algorithmic Fairness (EWAF'25).* Proceedings of Machine Learning Research, 25 pages.

1 Introduction

Data and Fairness, Fairness and Data. Data-driven decisions are becoming ubiquitous in both the private and the public sector [100, 115]: many areas, such as lending, hiring, and medicine, as well as criminal justice and

Authors' Contact Information: Patrick Oliver Schenk, p.o.s.on.stats@gmail.com, Department of Statistics, LMU Munich, and Munich Center for Machine Learning (MCML), Munich, Germany; Christoph Kern, christoph.kern@stat.uni-muenchen.de, Department of Statistics, LMU Munich, and MCML, Munich, Germany; Trent D. Buskirk, Old Dominion University, Norfolk, VA, United States, tbuskirk@odu.edu.

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

policing, public health, welfare and unemployment services, rely more and more on predictions from supervised machine learning (ML) models. Legal requirements, human rights, and moral obligations require these decisions to be just – which means that the trained ML models on which these decisions are based must be fair [85, 94, 136]. That is, these ML models must work equally well for different parts of the population, which are often defined by protected attributes such as gender or age. Algorithmic fairness becomes even more important in a world with (semi-)automated decision-making (ADM). First, while human decision-makers can reasonably be expected to incorporate justice and fairness into their decisions, a trained prediction model does not come with a moral compass [140]. Second, algorithms scale [29]: the fairness of one algorithm may affect many more people than any single street-level human decision-maker ever could. Third, perception, trust, and acceptance of an algorithm are influenced by its (perceived) fairness [76, 136], making private and public algorithm owners care about fairness. Beyond the domain of explicit decision-making, ML and AI are increasingly used for many tasks, including in data production [124]. Individuals and organizations employ pre-trained Large Language Models (LLMs) which can exhibit biases in gender, race, profession, or political ideology that reflect issues within the training data [43, 102, 123]. Instructional fine-tuning and human-in-the-loop reinforcement learning have not solved the inherent biases and in some cases made them more pronounced [43, 67, 123].

ML models can only be as good and as applicable as the data on which they were trained [11, 57]. Thus, unsurprisingly, many biases identified in the fair ML literature can be traced back to data shortcomings (see Section 3). For instance, how and which criminological data are captured differs across years and regions even within a country [91]. Also, performance differences across sex and race in the crime prediction algorithm COMPAS have been linked to historical bias in the training data [5, 8, 58]. Predictions of health needs were worse for Black than White patients because of how the outcome was operationalized and measured [105]. The Gender Shades study traced drastic accuracy differences across gender and skin type in facial recognition software back to issues related to dataset composition and representation issues in the benchmark datasets [19]. Such data deficits may persist or even self-amplify, as the biases in the models trained on them lead to decisions that cause the same or even more errors in future data collections [109]. Hence, while doing the "data work" may be less popular than engaging in the modeling work and may depend on context and require expertise [71, 122], data are crucial.

TSE and TDQ: The Two Total Error Frameworks. Statisticians, social scientists, and specifically survey researchers have long grappled with the many different sources of error in data [54, ch. 2]. In response, the *Total Survey Error (TSE) framework* was developed to provide a structure to assist anyone involved in designing, planning, collecting, processing, or analyzing (survey) data [16, 55]. Its long-standing success spawned methodological extensions, e.g., for big data and for platform data [2, 129]. [151] recently generalized the TSE beyond the survey sphere to the *Total Data Quality (TDQ) framework*, providing a scaffolding for the study of errors in *any* type of data. In this context, 'errors' encompass all deviations: uncertainty arising from random variation (e.g., sampling variance) and bias from systematic causes. These two conceptual frameworks are simply structures to categorize, organize, and understand the many sources of error in the data production pipeline (see Figure 1): They distinguish between errors of *measurement* and errors of *representation*. Each of these two sides contains multiple error mechanisms, explicated in Sections 4 and 5 respectively, that are arranged according to the sequential logic of data production. First, errors of measurement (left-hand sides within Figure 1) relate to questions about whether the values of a specific variable in the data accurately and precisely reflect what the data analyst wants the variable to measure. This view pertains to the outcome variable (do we actually predict what we want to predict?) and to fairness-relevant attributes (can we accurately evaluate model fairness for the exact groups we care about?). For the features in prediction tasks, however, we are more interested in whether the way they are measured during deployment is consistent with how they were measured in the training data. If the features are defined, measured, or processed differently in the deployment phase, the quality of the predictions will decline overall and possibly more so for some social groups than for others [57]. Second, errors in representation (right-hand sides within Figure 1) relate to whether or not the composition of the training data matches that of the target population. Misrepresentation *can* refer to whether or not a group is represented at all in the data, but in general we care about the degree of over-/under-representation. Representation is not only pertinent to fairness-relevant attributes, but is also relevant with respect to the outcome variable (e.g., women in gender classification, [19], and healthy images in COVID prediction, [116]). This paper focuses more on representation, as this central concept often remains ambiguous and underexplored in the ML literature [26, 131]. In his keynote at EWAF'23 [1], Arvind Narayanan, perhaps deliberately exaggerating a bit, even suggested that the importance of representation had not yet been conclusively established in the fair ML literature.

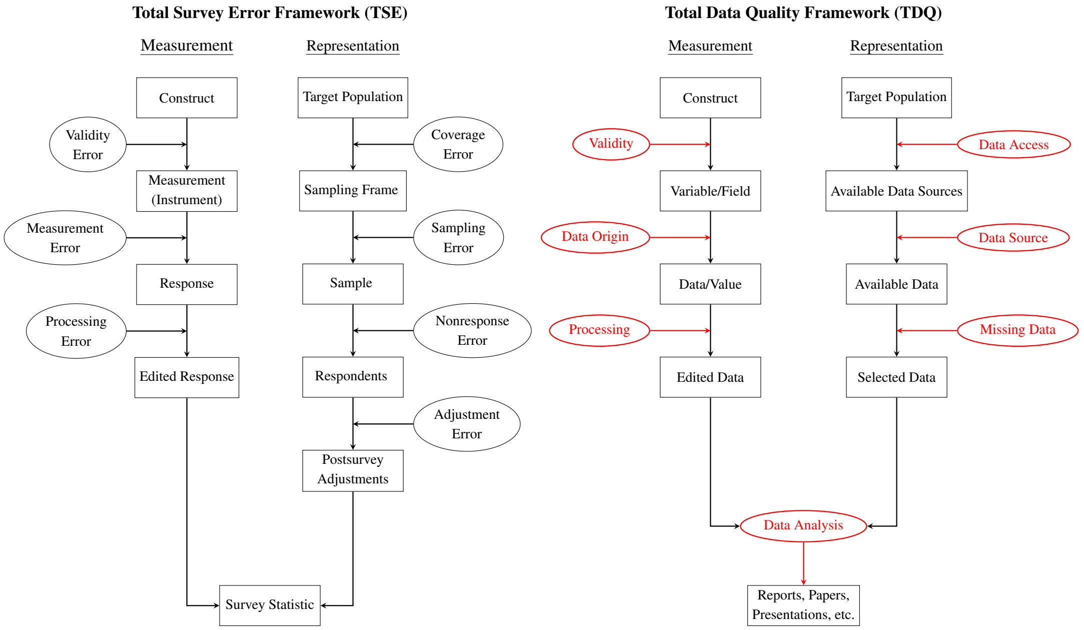

Fig. 1. The Two Total Error Frameworks. Left: The Total Survey Error Framework, inspired by similar figures in, e.g., [54, 55]. Right: The Total Data Quality Framework (Dimensions of TDQ by [151] is licensed under CC BY-NC 4.0).
As TDQ is a generalization of TSE, they exhibit a very similar structure. Errors of Measurement are on the left-hand side of TSE and TDQ, respectively, and Errors of Representation are on the right-hand side of TSE and TDQ, respectively. The frameworks as well as Measurement and Representation are explained briefly in Section 1, while the specific concepts are elaborated upon in Sections 4 and 5.

Tackling Data-based Sources of Unfairness. We believe that adapting frameworks such as the TSE/TDQ to the fairness paradigm can provide researchers and practitioners with a road map for tracing back and metrics and tools for subsequently evaluating antecedents of unfairness related to measurement or representation issues within the data used to train an ML model. On the practical side, adapting such frameworks to model fairness can: (1) minimize loss related to time, energy, and money from training models unlikely to be deployed due to their fairness inadequacies (see [116] for a stark example), (2) aid in identifying, preventing, and repairing errors directly at their source, (3) possibly improve the design of the data collection up front to produce higher-quality data, and (4) provide insights into which data sets may be unsuitable for specific data analysis goals. The frameworks also promote the perspective that data are (or can be) designed [121],[1] and that therefore errors can be prevented, mitigated, or remedied. We thus have several audiences and several goals with this paper. First, we would like to encourage commercial and noncommercial data collectors and providers [92, 124] to use comprehensive data error considerations based on the TSE/TDQ frameworks that also explicitly incorporate fairness.[2] The same also applies to the selection of benchmarks, and the study of errors of measurement and representation within them, both for the fair ML community specifically [41, 87, 108, 134] and the ML community as a whole for which, e.g., (differential) errors in labels and of geographic representation have been noted [15, 104, 112]. Second, the data analyst or modeler is often uninvolved in the data collection – which is all the more a reason to have a structured way of learning properties of the data (collection):[3] What are the error mechanisms that are the root causes of the unfairness that a modeler then tries to repair? Have one's own data pre-processing steps introduced sources of unfairness? For a given task, which datasets are likely to yield unfair models and thus should be avoided in favor of other datasets? Just as there sometimes is not a technological solution for social problem [10, ch. 2], or at least not one that does not make things worse, a particular dataset may simply not be fit-for-use from a fairness perspective. Recognizing this up front instead of wasting resources on or becoming attached to models that consumed a lot of resources during training but are unfair, is a positive. Finally, the theoretical conceptualization, understanding, and discussion of data-based sources of unfairness in the fair ML field stands to profit from the systematic structure of the frameworks: as mentioned in Section 7, some of the well-known concepts of data-based sources of unfairness are ambiguous, not defined precisely enough, or used with mean different meanings by different authors – often without this being realized. Building towards this insight – constructively and without criticism of the fair ML community – and familiarizing the community with these frameworks are the main goals of our paper. In the social sciences, employing these frameworks to track errors has helped us to better understand the relationship between possible errors and the properties of the output of models (such as estimates in statistical inference or predictions

[1] This also moves us away from the understandable, but ultimately rather limiting view of all data as "data exhaust" [70], as "found" or "organic" [52, 53], "gathered" [151], as "data in the wild" [4] or "naturally occurring data" [49] – i.e., as data without any provenance or information about their genesis. On the contrary, data are virtually always designed – but these decisions may simply have been made by someone else, or without conscious deliberation, or may not be sufficiently documented.

[2] Survey producers may use the frameworks but typically neglect fairness, while (some) other data collectors may consider fairness, but not in a structured way.

[3] While fairness issues may indeed be related to the model or its specification (e.g., inclusion/exclusion of useful features), they can also be related to biases in the labels (i.e., measurement issues) or to the composition of the training data with respect to particular subsets of the population. In the prediction of diseases or the efficacy of medication, certain groups defined by age and/or sex after often all but completely missing; this flaw can not be fixed by smarter modeling or fairness patches applied ex-post – and the fairness evaluation may also be consequently off, too. Only better data can help in such situations and the sooner this is realized the better.

from supervised ML). Building on the research on and experience with data and data quality in which these frameworks play a central role [38], we believe that their use likewise can help researchers identify, understand, and subsequently avoid or rectify data deficiencies as possible sources of unfairness.

In Section 2, we explain the data types from which we picked the examples in Sections 4 and 5. In Section 3, we place representation issues within the broader context of model fairness. Then, we elaborate on errors of measurement (Section 4) and representation (Section 5): after starting each section with a brief overview, we illuminate the error mechanisms as described by the frameworks through myriad real-world data examples. Section 6 provides two examples of how the frameworks can be applied within the data science pipeline. We relate the bias source concepts from the fair ML literature to the error mechanisms in the total data error frameworks and show how the latter may enrich our understanding in Section 7. Section 8 addresses limitations and implications, and gives an outlook on future research directions. We focus a bit more on representation than measurement throughout.

2 Tracing Errors Across Data Types and Selection Mechanisms

Data Types. Our discussion of errors (Sections 4 and 5) draws on examples from three main data types. While not fully exhaustive, these categories capture much of the diversity among real-world data sets. We describe the data types below; in Sections 4 and 5, we provide one example from each data type for every error discussed, but without explicitly naming the respective data type every time. *Standard tabular data* have been used for decades in, e.g., the social sciences, and generated through, e.g., surveys [20, 54]. They are typically variable-rich but case-poor. They tend to be analyzed with traditional statistical inference methods, with the goal of estimating population parameters of interest. Somewhat in contrast, the two other data types – which we describe next – have gained prominence in the modern data world, with regard to both how they are generated and the role of ML: either their nature (e.g., their sheer size or lack of structure) necessitates ML or they are analyzed for the purpose of prediction (where ML outperforms traditional statistical methods). *Modern structured data* share the tabular, structured nature with standard tabular data, but deviate in size: e.g., big administrative data (e.g., electronic health records or data from a public employment agency; [2]) but also ‘platform data’ such as from a social media platform (information about time spent on the platform, sociodemographics, social ties, likes), online marketplace platforms (name, location, and purchases), open knowledge platforms like Wikipedia, etc. They feature many users who are typically observed at many points in time but often contain markedly fewer variables than surveys. *Unstructured data* differ strongly from the other two, tabular types in that thinking of rows as ‘units of observations’ and columns as ‘variables’ is of little relevance. Such databases, containing unstructured text, images, and/or audio, already have a major impact on research and business and are, for example, used to train image classification algorithms and LLMs or other foundation models [18, 63]. Often, data of this type require some sort of labeling done by annotators [14]. These three data categories are not always mutually exclusive,[4] which however does not diminish the utility of discussing errors along these three prototypical categories.

[4] E.g., structured platform data may also contain the unstructured text posts, and the largely unstructured information from a smartphone's sensors may come with some structured information about the device owner. Similarly, data may have been created through the ‘record linkage’ or ‘data fusion’ [25, 64, 113] of data sets of differing types (e.g., images linked with their metadata).

Selection Mechanisms. In general, selection mechanisms[5] that can cause misrepresentation may depend on the outcome variable, regular features (i.e., those without any fairness relevance), or fairness-relevant attributes such as sex or age. When insightful for our discussion, we will consider this distinction. It is worth re-emphasizing that representation is not solely about the composition of the data with respect to fairness-relevant attributes: rather, it pertains to all variables, and the outcome variable in particular.

3 Representation in Data As a Driver of Model Unfairness

The fairness in ML literature has developed a rich set of multi-disciplinary concepts and methods to study social implications and biases of ML algorithms across domains [see 22, 93, 96, 100 for overviews]. Evaluations include assessing disparities in model performance and reliability between groups defined by *protected attributes* to mitigate the algorithmic perpetuation of discrimination based on ascribed characteristics such as gender, ethnicity, or social origin. In a narrow sense, protected attributes may be defined based on anti-discrimination legislation, but can also more broadly be understood as identifying social groups that may be considered salient or vulnerable in a specific application context. Given a set of groups, *fairness notions and metrics* can be used to formalize fairness concerns and to quantify disparities in model outputs and performance. Group-based fairness notions formalize requirements that model outputs should satisfy when compared between sets of groups. Such requirements may include comparisons of errors that ask how rates of overall errors, false positives, false negatives, precision, or recall, differ across social groups. Furthermore, notions have been proposed that focus on subgroups that may be defined by intersections of multiple attributes (multi-group fairness, [62], see, e.g., gender and race in [19]), or that focus on the individual level (individual fairness, [35]) with the aim of providing stronger fairness guarantees.

A critical finding in fair ML is that disparate errors between social groups can be traced back to *bias in data* as one of the root causes of model unfairness. Out of 19 types of biases that can affect model (un)fairness as identified by [96], 14 can be attributed to data deficits. One concern is historical bias, which refers to the observation that training data most commonly reflect past decisions and (traces of) interactions of humans and thus encodes human biases, prejudice, and discrimination [139]. Historical bias, however, may be amplified by other data deficits such as measurement bias, which is commonly discussed in the context of biased proxy variables used for model training [105]. Next to historical and measurement biases, representation bias broadly refers to deficits in the composition of the training data. Concerns about misrepresentation have been raised in various application settings in which insufficient data and naive modeling practices have led to the deployment of models with considerable error disparities between subpopulations [78].

While historical bias and measurement bias are covered extensively in the field, the different causes, nuances, and implications of representation bias have not yet been systematically covered. Studies such as [42, 117] analyze the link between misrepresentation and model fairness, but neglect the different mechanisms that can lead to (different forms of) misrepresentation. A broader, more systematic treatment of this complex issue can not only point to gaps in the methodological literature but can also provide guidance for assessing training data quality in practice. And viewing representation through a total data error lens may also provide a more complete framing of the typology of misrepresentation that can be leveraged to better understand when, where, and if it would ultimately result in fairness issues downstream.

[5] We eschew the term 'selection bias' as non-systematic random variability can also cause misrepresentation in the data and model unfairness.

4 Errors of Measurement Within the Total Error Frameworks

Overview. On the left-hand side (Figure 1), the process does *not* start with data that we already have, i.e., with variables for which other people have already made decisions. Rather, the initial step is the 'construct': "the elements of information that are sought by the researcher" [54, ch. 2.2.1], i.e. "a theoretical trait, or characteristic, that we attempt to quantify" [151] – e.g., health or personality [105, 151]. This principally abstract concept of what it is exactly that we would like to measure serves as the reference point against which actual data values will be compared.[6] This abstract notion is then operationalized into a 'measurement' – which can lead to deviations: this may be obvious when, e.g., a particular attitude construct is translated into a single survey question (which may be interpreted differently by a respondent than what the researcher intended), but is also largely true for seemingly direct measurement instruments of seemingly objective values such as for a person's body weight (with/without clothes, after waking up/at 10 a.m./the daily maximum weight). Next, a 'response' is generated by the observational unit itself or, more generally, by the use of the measuring instrument. Finally, this initially measured response value may be changed, leaving one with the 'edited response': the observational unit itself, a device or person involved in the measurement (e.g., the interviewer in a survey), or a data analyst may make edits based on plausibility checks (e.g., truncation or replacement of extreme or contradicting values) or on other considerations (rounding, categorization, harmonization, etc.). We now elaborate on the respective measurement error mechanisms in the two total error frameworks (Figure 1) using examples from the three data types (Section 2), before discussing in Section 7 how these errors relate to concepts from the fair ML literature about bias sources (see also Table 1).

Construct. What exactly is to be predicted? What are the groups for which model fairness should be evaluated? The journey that is the data pipeline has to be grounded with a clear vision of the to-be-measured construct [68]. Without a well thought-out concept that is coordinated and communicated within the research team, or when a researcher unintentionally bases the analysis only on their own life experience or on a dominant population group such as the "default male" [30], the analysis lacks an adequate reference point. A given concept may have totally different meanings or even be missing across cultures, even before trying to measure said concept: e.g., researchers from the US are often surprised that 'race' is not a frequently-employed concept in Europe, highlighting that fairness-relevant attributes can be constructed in multiple ways [61, 69]. The broader the target population and the less one knows about it, the more likely one may be to adopt an inadequate construct. For instance, one might overlook whether 'work hours' should also include unpaid care work, or whether 'wealth' should, in addition to financial and real estate assets, also consider productive livestock. These choices would tend to impact women and people from rural, self-sufficient, or developing backgrounds more.

Validity Errors. Choosing instruments with high validity typically requires expertise and testing. Furthermore, such errors not being invariant across, e.g., age, gender, or cultural groups has been noted repeatedly: e.g., an initially perfectly fine English survey question may be sub-optimally translated into another language. In health data, diagnostic guidelines or items in diagnostic questionnaires may not work as intended, and differently so for different groups [e.g., 33, 47]. In Google Trends, researchers who are interested in the actual internet search terms see only searches after auto-completion and after auto-correction of supposed spelling mistakes [66, 86]. The

[6] Outside the social sciences, it is not common to think of variables in terms of 'constructs' [but see 86]. Yet, upon closer inspection, surprisingly many variables are based on constructs, as there is often not that single, objectively true concept that has only one possible definition.

auto-completion suggestions also depend on the particular user, including on past searches. Actual search terms are classified into so-called entities, but these supposedly language-independent concepts may work better for some languages and some search terms than for others. Finally, major changes to the search algorithm may also in turn change users' search behavior [23].

In annotation tasks, there is a crucial difference between tasks that seek a single objective truth (i.e., ideally each annotator would provide the identical label for a specific instance) versus subjective, "perspectivist" tasks in which capturing the heterogeneity of views in the population is the goal [45]. These two tasks typically require different measurement instruments. For the perspectivist task, the heterogeneity of interest needs to be captured adequately.

Measurement Error/Data Origin. While the previous step was about discrepancies between the measurement instrument and the construct of interest, the present step considers the application of the measurement instrument to individual observational units or data points. Fields such as survey methodology, human-computer interaction, and human-AI interaction have amassed vast knowledge about such error mechanisms in general [54, 72, 132] and how these may differ across fairness-relevant groups [see also 12, 127]. If data about a human subject are collected, such as in a survey, the presence of other humans (interviewers, family members, co-workers, etc.), the wording and order of questions, the suitability of the response scale, the mode (with/without interviewer, by telephone or web, etc.) and the device used to collect the data, are some of the factors known to influence how humans comprehend questions, retrieve or recall information, and decide how to answer [54]. The incentives and the specific context of the human in the data collection situation also factor into how much effort the person gives. Social desirability, particularly for sensitive items, and proxy interviews (e.g., providing information about other household members) are further sources of errors that may be more prevalent and more systematic for some social groups than others [126].

Humans are not only sometimes the observational units, but also provide annotation work [50, 99]: "the design of the annotation instrument, the instructions given to annotators, the characteristics of the annotators, and their interactions can impact training data" [75; see also 13]. Even conditional on showing the same content (say a person wearing a suit), images of females and males may have received different labels (considering only images with a person wearing a suit: the label "suit" is attached by annotators more often for one sex than for the other; see [128] for more). Aside from instructing and educating properly on the specific annotation task, ensuring the representativeness of the annotator pool – directly or by statistical methods – can also mitigate labeling errors that may otherwise differentially affect fairness-relevant groups [37]. Meanwhile, electronic health data depend on the quality of diagnoses, but even given the same symptoms, the same severity of disease status, and the same diagnostic guidelines for men and women, clinicians tend to diagnose women differently [32, 60].

Data origin errors also arise when data from different sources (in the broadest sense) are combined, such as health data from multiple hospitals, location data from multiple devices (GPS and Wifi), or generally, data from multiple databases that use different measurement units or different representations to format and store information [151]. Differential errors may also be introduced by combining data collected through different modes, device types, connection to the internet, and regions [77, 83, 149]. Paradata, metadata, and proper documentation regarding these different origins may help [103, 126, 147].

Processing Error. Initially recorded values may be edited during (pre-)processing by a data provider before the release of the data or by a researcher before the data analysis begins. Data may be processed for multiple reasons, including feature engineering or to enhance model explainability (e.g., re-coding or transforming variables), increasing sample size for data analysis or informational content (e.g., record linkage and imputation of missing values), or general data cleaning. On the one side, the intention of these procedures may be orthogonal to errors or they even may be employed to reduce errors – but this is not guaranteed, and even if so, this may work better for some cases or groups than for others. On the other side are intentional distortions of the original values, e.g., to preserve privacy, and this may affect some groups more than others [144].

Regardless of the data type, errors can occur during record linkage where entries belonging to different entities can be incorrectly merged, and actual matches can be missed. This can disproportionately affect certain social groups, such as those with a small set of highly common surnames, or those affected by multiple transliterations of non-native names or encoding errors in letters that are not native to a locale [64, 151].

Relatedly, [134] report multiple, sometimes questionable processing strategies for the protected attribute ‘race’ in the COMPAS dataset [5; see also 41], even in fairness research. Combining categories with low frequencies, such as in race or ethnicity, may be understandable. Statistically speaking, this is an averaging process, which only works well when the affected groups are very similar [57].

5 Errors of Representation Within the Total Error Frameworks

Overview. At each stage of the right-hand side, we can ask how well the ‘target population’ is represented. First, from this target population, a list of all units that could theoretically be included in the data is explicitly or implicitly generated: the ‘sampling frame’ – which may or may not cover the target population completely. Second, typically only a subset of the units of the sampling frame is to be included in the data: sampling is preferred to a full census for reasons of cost, timeliness, and feasibility. Third, when, for a unit that we would like to include in our data, we get only incomplete or even no data, this is called ‘nonresponse’ or ‘missingness’. Fourth, post data collection, the data may be ‘adjusted’, e.g., by weighting, post-stratification, or decisions about which units to exclude, before they are finally analyzed. To put it differently: The data on which a model is trained come from an ‘implied population’ – the hypothetical population for which the data *would be* representative. Generally, the best that we can hope for an algorithm to do is to learn the exact patterns in the population underlying its data. But if this implied population differs from the target population – misrepresentation – then even in the best-case scenario, a biased model will be learned. We now elaborate on the respective representation error mechanisms in the two frameworks (Figure 1), again using examples from the three data types (Section 2), before relating them to fair ML concepts in Section 7.

Target Population. Whether for statistical inference about a population parameter of interest or for the eventual deployment of a trained prediction model, the work starts with precisely identifying one’s intended target population, including a time frame. This population may consist of humans (e.g., the adult US population on a particular date), relate to human actions/groups/etc. (e.g., all internet searches/small US businesses/etc. on July 1st, 2024), or have no connection to humans. As at each step the data are compared against the target population, a precise definition and compositional information about it are essential.

Coverage / Data Access. A ‘sampling frame’ is a list of all population units that, at the planning stage, in principle *could* be included in the eventual data. If this set of covered units is equal to the whole population, there is no selection at this stage and thus representation certainly is not harmed here. Otherwise, however, strong arguments need to be made as to why the composition of the sampling frame is certainly equal to that of the target population. However, data may often be provided ‘as is’ and documentation of an actual or implied sampling frame may be sorely lacking or be too technical.[7] In a few countries with excellent registers, such high-quality lists exist. However, in most places quite a bit of legwork may be necessary first: e.g., ‘address listers’ have been employed to walk around a geographical area and record every resident and address [54, ch. 3.3]; residents are then later invited by mail to participate in a web survey. *Coverage error* comes from groups or units with certain properties in one’s target population being (partially) missing, duplicated, clustered, or otherwise over- or undercovered on the sampling frame:[8] As the share in the sampling frame deviates from the share in the target population, simple random sampling from the sampling frame will not yield representative samples.

For platform-type data, there are two questions one can ask. First, are there differences between the target population and the platform population?[9] These are implicated by the researcher’s choice of which platform, database, or corpus to analyze, which data vendor to use, etc. [6, 111, 114, 150]. Second, can the researcher actually access each unit from the whole platform population (or at least a representative sampling frame generated from it)? Even a single platform may offer multiple APIs which vary in their levels of access [79]. For example, X only providing each user’s most recent 3200 tweets [110] introduces differential coverage error among tweets, as some tweets (those more in the past and from more frequent posters) are fundamentally inaccessible to the researcher. More details on this example and others from this section can also be found in [151].

For databases of, say, pictures of humans the concept of a (target) population can be quite tricky: e.g., does one consider the set of all humans or of all pictures showing humans? The set of all humans who currently live, who have ever lived including now, or those who will ever have lived? In either case, there is no full list available. Web-scraped data represent a mixture of past and (near-)present, but with weights that are unknown and that vary over time (more pictures from certain time periods than from others, not representing the number of humans living). Inheriting imbalances on the web, webscraped image benchmarks can lack representation concerning, e.g., race and gender [19]. This problem can also be present – and perhaps amplified – if (before scraping) the images are themselves selected through an algorithm, e.g., of a search engine [98].

Sampling / Data Source. Sampling, i.e., the process of choosing units from the sampling frame, is the next potential error source – that is, unless *all* units are to be sampled, i.e., a census. For traditional tabular data on humans, sampling theory is well developed [145]: from the well-known simple random sampling, over stratified random sampling in order to improve efficiency and other sampling mechanisms that need to be taken into account during data analysis, to model-based approaches and recent attempts to make use of non-probability samples [82]. An example of sampling error is when people from the sampling frame without access to the internet, a computer, or a smartphone cannot participate in a web survey [17, 27, 36]. The respective penetration rates are nowadays

[7] Proper documentation, enabling the assessment of representation in the data, includes, e.g., information about the sampling frame, inclusion and exclusion criteria, sampling methods, (subgroup) response rates, and weighting or other adjustments. This realization from the social sciences [54, ch. 10.8] has also reached the ML community [48]. Only some of these measures can be computed by the analyst from the data themselves and compared to gold-standard information about the population. This is less possible for measurement issues; hence, documentation of, e.g., concepts, instruments, encoding schemes (e.g., missing values), and changes to the original data values may be even more essential.

[8] This includes the respective extremes: ineligible units (i.e., those who are not part of the population) being present or certain groups missing completely (e.g., analphabets).

[9] Precision is key here, too: duplicates, dead accounts, shared/group/business accounts, bots, etc. are reasons why the population of humans on a platform deviates from the population of platform accounts.

generally high in many Western locales [142] – but this can be quite different for other settings. With the logic that if most people are on the web, the potential damage of the ‘digital divide’ is limited for the calculation of population-wide average quantities, those survey organizations with erstwhile programs to outfit the offliners with internet/devices have tended to somewhat abandon these programs [7]. This removal of subpopulations that are not or only rarely online is a clear fairness problem from their perspective. Even on the aggregate level, however, if the study variable is highly correlated with internet access/device usage, then neglecting such subpopulations creates bias [34]. Smaller, fairness-relevant groups may be particularly affected. The specific target population and data analysis goals clearly matter for choices such as the data collection mode (e.g., web/telephone/mail/in-person surveys).

For platforms, the continual development of platform algorithms implies that researchers often do not know the exact sampling mechanisms, exacerbated by the unawareness about whether and when changes to algorithms occur that also change sampling behavior [86, 118]. Beyond the important issues of reproducibility and replicability, such changes can fundamentally bias the attempt of ‘sampling from the past’. Another reason is that accounts may have been removed or temporarily banned, or have changed visibility status. Both systematic bias and random variation are present in, e.g., Google trends data [65, 66]. Even when fixing a past date of interest, the data can differ (1) based on the downloader’s geolocation and (2), for the same downloader from the same location, from one download to the next.

Webscrapers can miss some webpages completely due to errors in the website’s code or architecture, a lack of indexing, because the received information is dynamic and tailored to supposed, individual user characteristics, or a webscraper can miss certain subpopulations on a platform if the underlying website handles them differently but the webscraper is not attuned to these differences [44, 89, 151].

Nonresponse / Missing Data. Information may be missing partially (a unit is observed in principle, but with missing values in some of the variables) or the to-be-sampled observational unit may be missing completely (‘missing record’). These two cases are only the same if there is but one single variable of interest, or, more applicably, when the researcher later opts for a complete-case analysis.

In web surveys, missingness can result from, e.g., a failure to contact a unit (at all or in time), its refusal to participate at all or to answer a specific question (e.g. because the true response is sensitive), or because of survey break-off [20, 24, 31, 97, 152]. Voluntary fields, such as sociodemographics on a platform, may be missing because they were forgotten or simply not filled out. Specific fields may also hidden from the researcher’s API access. GPS geolocation or data from other sensors may be missing, e.g., due to a temporary or continuous lack of internet access [9]. Any database based on record linkage or data fusion can suffer from missingness because some links were not recognized or because some units were present only in one but not both of the original data sources.

Adjustments / Data Analysis. Once collected, data often undergo some adjustments before finally being analyzed. These adjustments are often not directly attributable to the data and their collection; they may also often only be carried out by the eventual data analyst – who may be different from those who design, collect, (pre-)process, or provide the data. Thus, we cover this topic more briefly. Adjustments can be seen as part of the data analysis step in the TDQ framework, and in the TSE, usually only one aspect is highlighted: weighting (or post-stratification), either implemented alongside pre-planned sampling designs (such as deliberate over-sampling of small subpopulations to ensure they are sufficiently reflected in the data) or applied to correct for unintended representation errors that crept in at some point along the pipeline. However, whether pre-specified or estimated, these weights may themselves suffer from bias or random variability.

Also, for data with a very imbalanced outcome variable such as in rare diseases, deliberate under- and/or over-sampling of very rare or very prevalent categories, respectively, has gained some traction in ML circles [39]. However, this disturbs representation with respect to the outcome variable, thereby worsening the calibration of prediction models [146]: unchecked, the model predictions reflect the category frequencies in the perturbed data instead of the actual frequencies in the population. Our ongoing work shows that then fairness problems between groups usually arise, even when the under-/over-sampling mechanism is identical across them. A further potential source of representation errors stems from the usual train-test-evaluation data splitting in the supervised ML pipeline: through random variability, random data splitting – while unbiased in expectation – may still affect small-ish data sets, high-dimensional settings, and relatively rare subpopulations in particular. Stratified data splitting is an improvement [101], especially when relevant representational characteristics are known or in a low-dimensional fairness setting.

Finally, record linkage or data fusion of two more data sets may also create or exacerbate misrepresentation.[10] Depending on the context, individuals' consent for record linkage is necessary. Consent rates are not 100% and can differ between social groups: e.g., according to trust and social status, or depending on the precise wording and placement of the consent request [119, 120].

6 The Total Error Frameworks in Action: A Brief Glimpse

Below, we sketch how to apply the frameworks ***step-by-step*** to two important, fairness-relevant prediction tasks. When discussing the error mechanisms, we mostly omit the repeated reminder that fairness issues can be created when these mechanisms depend on fairness-relevant attributes or the outcome variable.

Measurement Example: Crime Prediction. In the first step, the ***construct*** of what is to be analyzed must be agreed upon. Laypersons may not be aware of the legal nuances, such as what is and what is not a criminal offense, what is prosecuted *ex officio* or only if a formal complaint is lodged, and so on.[11] If certain crimes are unintentionally excluded (e.g., committed vs prosecuted crimes), and those tend to be correlated with protected attributes – such as in white-collar crimes associated with sociodemographic factors [21, 46] – a fairness bias may be introduced when choosing or formulating a ***measurement*** instrument. Criminological data can be very hard to compare across years, but also across regions even within a country because of differences in how and which data are captured [91]. As the population differs across regions in terms of, e.g., protected attributes, such regional data differences can then translate into fairness issues in the algorithm trained on them. Finally, the data can be ***processed*** or ***edited*** in multiple ways. For instance, prosecutors may reduce the initial charges to lesser or fewer offenses that are easier to prove in court or which already carry the maximum penalty. Similarly, the verdict of a judge can be seen as an edit of the actually committed offenses.

[10] It is debatable whether record linkage is best considered under the heading of adjustments/data analysis, or whether it should be considered as its own way of creating data and thus have all of the TSE/TDQ steps applied to it. For ease of presentation, we opted for the former.

[11] Many legal systems distinguish between criminal law (e.g., murder or theft), administrative law (e.g., traffic violations), and private law (e.g., contract or property disputes). Among criminal offenses, many must always be prosecuted *ex officio* (e.g., murder), while others are investigated only after a formal complaint is lodged by an affected party (e.g., minor shoplifting). In contrast, private law matters almost universally require legal action from an affected party before state authorities initiate proceedings.

Representation Example: Cancer Prediction. In this example, we assume that we want to predict the occurrence or progression of cancer based on images, clinical information, or behavioral factors, and that our ***target population*** of interest is the entire population of a country. Such prediction models could be trained on cancer registries, but these do not exist in every country; where they do, their ***coverage*** does not necessarily extend to the entire population or the cases may be fragmented across several state-level registries instead of being consolidated in a single, all-encompassing national registry [51, 88, 133, 153]. Making it mandatory for healthcare professionals to report all cancer cases [see 74, 153] is a boon for representativity, and it is a step up from other health data for which inclusion may depend on, e.g., whether someone has private or public health insurance. Still, not all medical facilities possess the same diagnostic expertise and some cancers might not be discovered while a person is still alive; hence, certain hard-to-diagnose cancer types may be under-represented as may be social groups who predominantly live in areas with lower diagnostic expertise. When merging regional cancer registries into a single database, cases can be duplicated or lost; if, in the absence of unique identifiers, linkage is based on name and similar information, the demographic composition of the merged data can deviate from that of the total population. Finally, non-cancer cases are naturally under-represented in cancer registries.

Sampling from such a registry brings sampling uncertainty. This can be eliminated by analyzing the entire dataset instead of a sample. However, for very large databases or if the individual data points require too much memory to store or process, as potentially in image data, sampling may become unavoidable. In addition, ***data sources*** that are made available to researchers may only consist of a subset of all cases as a measure to protect privacy. In any case, stratifying the sampling by the outcome, directly fairness-relevant attributes, regions, and so on can drastically reduce the sampling-induced uncertainty [101], thereby particularly benefiting the analysis of rarer cancer types and for small population groups.

Nonresponse from the side of the healthcare professionals is largely eliminated by mandatory cancer reporting, but patients cannot be forced to continually contribute data. Attrition is a general problem in longitudinal data. As patients move, especially across state lines in a country with regionally separate cancer registries and without a unique record linkage identifier, cases or information may be lost – and differentially so, if mobility is associated with fairness-relevant attributes. While reporting of cancer cases may be mandatory, some information such as the stage of cancer at diagnosis may still be optional for the healthcare professional to record [51] and thus be ***missing***. Similarly, the availability of treatment information varies, for instance by cancer type, which in turn can be strongly connected to, e.g., sex [51]. Hence, a complete-case analysis of the data can create distortions concerning both, the outcome variable and fairness-relevant attributes.

In the last step, ***post-collection adjustments*** of the data such as weighting may be undertaken. Whether this indeed makes the final data match the target population closely must be checked, as weighting comes with increased variability. Finally, the researcher may be tempted to re-balance the data, deliberately oversampling, e.g., rare cancers, for the training data, with the negative consequences described in Section 5.

7 Data-Based Sources of Bias Through the Lens of the Total Error Frameworks

In Table 1, we propose a mapping of the data-based sources of bias common in the fair ML literature [96; see also 107, 138] onto the error sources in the TSE/TDQ frameworks. In the following, we only discuss some bigger picture takeaways. First, we note that there is not a 1:1 relationship between the concepts in fair ML and the TSE/TDQ frameworks – hence the latter do not simply give new names to already existing concepts, but provide additional insights and perspectives. Second, some mappings cannot be established with complete confidence: for instance, it depends on the literature source whether 'representation bias' maps only onto 'sampling error' or also includes, in particular, 'coverage'. As we have not seen this issue discussed, the recognition of such differences might not have fully permeated the literature, highlighting the value of the TSE/TDQ frameworks for structuring the existing concepts. Third, several bias sources map onto multiple mechanisms in the TSE/TDQ frameworks. Some of the bias sources even map onto both measurement and representation – or at least it is not directly clear if a bias concept is only to be applied to one of these sides. Fourth, the bias source concepts are not all on the same hierarchical level: some are subtypes of others, such as self-selection (within selection) or those related to contexts (social contexts, platform contexts, and so on). Fifth, while 'measurement bias' corresponds to the whole measurement side of the TSE/TDQ frameworks, it is unclear if there exists a concept that covers the representation side completely. Finally, the last two bias sources are related to omitted variables – possibly because they were not included in modeling, but also possibly because these variables were simply not collected. They do not directly map onto the TSE/TDQ frameworks,[12] which we address at the beginning of the next section.

8 Limitations, Implications, and Future Research Directions

The TSE/TDQ communities are aware of interactions between the measurement and the representation side (as evidenced by the interplay between measurement errors and nonresponse errors mentioned below on fp. 15). However, this is not directly visible in Figure 1. As such, we would like to see a more structured, semi-formalized approach and a visualization added to aid customary workflows. A similar argument pertains to missing variables, as was discussed at the end of Section 7. This would then complete the mental picture of looking at the components of an imaginary data matrix: its rows (representation), the cells for a given variable/column (measurement), and its columns (missing variables).

We gave many examples of the different error mechanisms but focused on three important data types. West et al. discuss their TDQ framework based on 'designed data' versus 'gathered data', which is not completely congruent with our three data types. Our presentation tried to consider the additional challenges when one gathers data designed by someone else – choosing the most suitable dataset for a given task – but see [151] for more.

Our focus was on group and subgroup fairness as their connection to representation is immediate. Concepts of 'individual fairness' have been primarily viewed through the lens of an individual receiving similar predictions as similar other individuals and with the right to algorithmic recourse otherwise [73, 148]. Yet, the individual concept does connect to the group level and to representation: Are individually unfair cases concentrated in some groups and thus may need solutions on the group level? If group membership is relevant, do enough 'similar individuals' even exist in the training data and in the test data?

We hope that our methodological paper has underscored the need for integrating representation considerations alongside the whole pipeline, from planning to designing to conducting a data collection, to processing and analyzing the data. Having focused more on the research than the applied side, we tried to highlight why we think the total error frameworks can provide additional methodological insight. One next logical step would then be to examine these frameworks with a real-life data analysis task or a simulation study.

[12] A rather indirect connection exists: the analyst may drop variables with missing values completely from the model.

The frameworks also offer the opportunity to locate and motivate where and why participatory approaches [81, 135] are most needed. We think that on the measurement side, the importance tends to decrease from top to bottom, with the greatest importance for the construct and the measurement, including when, e.g., annotators or LLMs are employed to collect, synthesize, or label data [37, 80]. This issue needs further work and formalization.

The evaluation of data (quality) with regard to fairness issues that we propose implies the need for points of comparison. For *representation*, general reference points can be information about the multivariate distribution of protected attributes and, ideally, the to-be-predicted outcome variable. International and national statistical offices as well as similarly equipped and trusted institutions should be able to provide such gold-standard reference points in a future that will become even more data- and machine-learning-heavy.[13] Meanwhile, reference points that are specific to a particular data analysis are implied by the sequential stages that are visible in the TSE/TDQ: more concretely, the composition of the data after having finished a stage should be compared to the composition after the previous stage [124, 125]. For *measurement*, however, statistical methods and comparison data alone may not suffice. Yes, it is possible to assess the random variability introduced by, e.g., annotators: have multiple annotators label the same instances and look at the inter-annotator reliability. In some circumstances, gold-standard information might exist for a small set of instances, and by merging these (hopefully) true values with the values generated by another method, bias and variability can be assessed.[14] However, such additional data may often be unavailable or too costly to produce. Measurement issues, more than representation issues, thus may require expertise: (1) substantive domain knowledge, (2) expertise in testing for, e.g., errors of validity [54, ch. 8], and (3) the specific knowledge of how the training data were generated and how that may differ in deployment.

Under-representation generally increases random variability. It can, but does not have to, lead to biased results: more missingness increases the theoretically possible maximum bias, but how much bias there actually is depends on the exact selection mechanism [56, 57, 90]. For traditional statistical analysis [90], under-representation may be either safely ignored outright (if the mechanism has no relation to the studied patterns), or ignored after applying correction methodology (re-weighting or imputation of missing values), or be non-ignorable. This triad, however, has been shown to be rather unhelpful for prediction-focused, supervised ML [57]: here, the most relevant question is whether there is selection on the outcome variable. The connection to fairness becomes most apparent by considering that how exactly the selection depends on the outcome may differ between fairness-relevant groups. This key characteristic – the importance of the outcome variable – has not yet been widely recognized in the studies of fairness: e.g., yes, there is misrepresentation with respect to sex and race in the data in the well-known Gender Shades study [19] – but that sex is also the outcome variable for the trained classification algorithm falls by the wayside in the typical mentions of this seminal study.

Much trust has been placed in more or big(ger) data, whether through increased sample size or the inclusion of additional variables [3, 59, 95]. However, suppose that you want to predict teen pregnancies, but your data consist only of elderly males: no amount of additional sample size or additional variables will fix the lack of representation. For less drastic cases of racial or gender under-representation, the results can still be quite deleterious [137]. Furthermore, if a subpopulation is underrepresented, e.g., because it is hard-to-reach or hard-to-recruit, attempts

[13] Fortunately, such distributional information is distinctly less privacy-sensitive than releasing the full microdata would be. For the data, the privacy of fairness evaluations is indeed an important practical issue beyond the scope of our paper [see 130], that at least does not diminish the theoretical conceptualization of error mechanisms.

[14] Ditto for measurement differences between training and deployment.

to brute-force a larger sample may actually worsen representation as for the additional observational units it may get increasingly harder to recruit from that subpopulation [141]. Moreover, the interplay between the effects of nonresponse/missingness and of measurement error can be quite complex: e.g., increasing the data collection effort to add new, harder-to-sample observational units may increase sample size and even decrease nonresponse bias, but at the cost of these new units providing lower quality data (more measurement error), with an ex ante unclear net effect [84, 106]. Omitting variables yields predictions that amount to averaging out over the omitted variables – generally a disadvantage for all population groups, but more so for groups that are smaller and that are further from that population average [57]. In addition, as evidenced by Simpsons's paradox, leaving out features may produce models in which the influence of a particular feature is estimated with the wrong sign – unfairly punishing some individuals with a 'better' set of feature values (who, in actuality, have better true risk score), thus setting up a situation for algorithmic recourse [73, 148]. Finally, missing variables may be forgotten because they are mostly relevant for only some parts of the population, highlighting the value of expertise and participatory approaches.

If, however, fairness-relevant variables are missing from the data, auxiliary data about the data ('metadata') or about the data collection process ('paradata') may help: e.g., paradata have been used to model measurement and representation errors [126]. However, these concepts are very underutilized in the field of fairness or parent fields such as trustworthy ML or algorithmic accountability (but see [143]). As the generation of metadata and paradata is often easy to automate, we call for their potential to be explored more outside of survey contexts [20, 126].

The fairness of a trained model is usually evaluated by calculating a fairness metric on test/evaluation data. However, these data are themselves usually not checked for their representativeness. Often, they come from setting aside a part of the training data just for the calculation of performance and fairness metrics. Thus, these test/evaluation data can be expected to inherit the misrepresentation from the training data, perhaps magnified by random representation errors introduced by the data splitting.

The investigation of errors of representation in a total error framework consists of comparing the current composition of the data with that of a target population. Without knowing the deployment scenario for the to-be-trained algorithm, this evaluation cannot be carried out. Conversely, it should be possible to look at a particular data product and describe the uses for which it is (likely) fit. This information, or at least the necessary components to compute it, could be appended as metadata when that dataset is added to a data archive [28, 40] such as at the ICSPR, LIS, Roper Center, or UK Data Archive. Then, researchers could search in these data archives for data that fit the fairness criteria for their specific data analysis goals and contexts, lowering the chances of unfair models.

Acknowledgments

POS has been partially supported by the Federal Statistical Office of Germany.

A Appendix

Bias	Short Definition	Errors of Measurement			Errors of Representation			
		Validity	Measurement/ Data Origin	Processing	Coverage/ Data Access	Sampling/ Data Source	Nonresponse/ Missing Data	Adjustment
Measurement Bias^{DA}	Choosing, measuring, processing, and utilizing variables (features)	✓	✓	✓				
Representation Bias^{DA}	Data do not match population because of sampling or data access [96], but sometimes also includes other steps such as defining the target population [138]				(✓)	✓	?	?
Selection/Sampling Bias^{DA}	Due to non-random sampling from population					✓		
Self-selection Bias^{UD}	Sampling/selection bias driven the observational units themselves					✓		
Historical Bias^{UD}	already existing human and socio-technical biases in the world	✓	✓	✓	(✓)	(✓)	(✓)	(?)
Population Bias^{UD}	differences between training and target populations (perhaps restricted to some kinds of variables [107])				✓			
Social Bias^{UD}	Our behavior (e.g., survey responses or annotation choices) is modified by social processes and norms	✓	✓	✓	(✓)	(✓)	(✓)	
Behavioral Bias^{UD}	Our behavior is modified by contexts (e.g., on which platform we are)	✓	✓	✓	(✓)	(✓)	(✓)	
Content Production Bias^{UD}	Behavioral bias: lexical, syntactic, semantic, and structural differences in user-generated context	(?)			(✓)			
Temporal Bias^{UD}	Differences in populations and behaviors over time				✓			
Linking Bias^{DA}	Arises when re-constructed personal networks (from online social networks) and true networks differ	(✓)	(✓)	(✓)	(?)	(?)	?	
Redundancy	Duplicated observational units				✓	✓		(✓)
Evaluation Bias^{AU}	Inadequate data used for evaluation and benchmarking	?	?	?	?	?	?	?
Omitted Variable Bias^{DA}	Incorrect association because of variables omitted from model							
Aggregation Bias^{DA}	Ecological fallacy: e.g., Simpson's Paradox. Population average conclusions do not apply to all individuals/groups							

Table 1. Relationships between Known Data-based Biases (rows; see [96, 107, 138]) and the Error Mechanisms in the TSE/TDQ Frameworks (columns)

Notes: Our confidence in and the consistency in the literature regarding these mappings are indicated, in descending order, by the symbols ✓, (✓), ?, (?). Empty cells indicate no relationship. Biases are re-ordered from [96] for additional cohesion. Superscripts denote the three bias categories used by [96] (DA: Data to Algorithm, AU: Algorithm to User; UD: User to Data). 'Redundancy' is not listed in [96].

References

[1] Jose M. Alvarez, Alessandro Fabris, Christoph Heitz, Corinna Hertweck, Michele Loi, and Meike Zehlike (Eds.). 2023. *Proceedings of the 2nd European Workshop on Algorithmic Fairness (EWAF 2023)*. Number 3442 in CEUR Workshop Proceedings. Aachen, Germany. https://ceur-ws.org/Vol-3442/

[2] Ashley Amaya, Paul P Biemer, and David Kinyon. 2020. Total Error in a Big Data World: Adapting the TSE Framework to Big Data. *Journal of Survey Statistics and Methodology* 8, 1 (2020), 89–119.

[3] Chris Anderson. 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine* 16, 07 (23 June 2008).

[4] Chee Siang Ang, Ania Bobrowicz, Diane J Schiano, and Bonnie Nardi. 2013. Data in the Wild: Some Reflections. *Interactions* 20, 2 (2013), 39–43.

[5] Julia Angwin, Surya Mattu, and Lauren Kirchner. 2016. *Machine Bias*. Technical Report. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[6] Marten Appel and Nicholas Haas. 2024. @Who? Investigating Possible Errors in Studies Linking Survey and Twitter Data. *Journal of Quantitative Description: Digital Media* 4 (2024). https://doi.org/10.51685/jqd.2024.002

[7] Ruben L Bach, Carina Cornesse, and Jessica Daikeler. 2024. Equipping the Offline Population with Internet Access in an Online Panel: Does It Make a Difference? *Journal of Survey Statistics and Methodology* 12, 1 (2024), 80–93. https://doi.org/10.1093/jssam/smad003

[8] Utsav Bahl, Chad Topaz, Lea Obermuller, Sophie Goldstein, and Mira Sneirson. 2023. Algorithms in Judges' Hands: Incarceration and Inequity in Broward County, Florida. *UCLA L. Rev. Discourse* 71 (2023), 246–280.

[9] Sebastian Bähr, Georg-Christoph Haas, Florian Keusch, Frauke Kreuter, and Mark Trappmann. 2022. Missing Data and Other Measurement Quality Issues in Mobile Geolocation Sensor Data. *Social Science Computer Review* 40, 1 (2022), 212–235. https://doi.org/10.1177/0894439320944118

[10] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org.

[11] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104, 3 (2016), 671–732. http://www.jstor.org/stable/24758720

[12] Israel Lucas Barros De Amorim, Suzane Santos Dos Santos, Ingrid Moreira Miranda Da Silva, Kamila Rios Da Hora Rodrigues, and Marcelle Pereira Mota. 2024. Gender Nuances in Human-Computer Interaction Research. In *Proceedings of the XXII Brazilian Symposium on Human Factors in Computing Systems* (Maceió, Brazil) *(IHC '23)*. Association for Computing Machinery, New York, NY, Article 54, 12 pages. https://doi.org/10.1145/3638067.3638077

[13] Jacob Beck. 2023. Quality aspects of annotated data: A research synthesis. *AStA Wirtschafts-und Sozialstatistisches Archiv* 17, 3 (2023), 331–353. https://doi.org/10.1007/s11943-023-00332-y

[14] Jacob Beck, Stephanie Eckman, Rob Chew, and Frauke Kreuter. 2022. Improving Labeling Through Social Science Insights: Results and Research Agenda. In *HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence*, Jessie Y. C. Chen, Gino Fragomeni, Helmut Degen, and Stavroula Ntoa (Eds.). Springer, Cham, Switzerland, 245–261.

[15] Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. 2020. Are we done with ImageNet? arXiv:2006.07159

[16] Paul P Biemer, Edith D de Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars E Lyberg, N Clyde Tucker, and Brady T West (Eds.). 2017. *Total Survey Error in Practice*. John Wiley & Sons, Hoboken, NJ.

[17] Annelies G. Blom, Jessica M. E. Herzing, Carina Cornesse, Joseph W. Sakshaug, Ulrich Krieger, and Dayana Bossert. 2017. Does the Recruitment of Offline Households Increase the Sample Representativeness of Probability-Based Online Panels? Evidence From the German Internet Panel. *Social Science Computer Review* 35, 4 (2017), 498–520.

[18] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li,

Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258

[19] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a

[20] Mario Callegaro, Katja Lozar Manfreda, and Vasja Vehovar. 2015. *Web Survey Methodology*. Sage. https://www.websm.org/

[21] Michael Cassidy and Carole Gibbs. 2019. Examining Sentencing Patterns and Outcomes for White-Collar and Property Crime Offenders. *Victims & Offenders* 14, 1 (2019), 75–95. https://doi.org/10.1080/15564886.2018.1547996

[22] Simon Caton and Christian Haas. 2024. Fairness in Machine Learning: A Survey. *Comput. Surveys* 56, 7, Article 166 (2024), 38 pages. https://doi.org/10.1145/3616865

[23] Eduardo Cebrián and Josep Domenech. 2024. Addressing Google Trends inconsistencies. *Technological Forecasting and Social Change* 202 (2024), 123318. https://doi.org/10.1016/j.techfore.2024.123318

[24] Gregor Čehovin, Michael Bosnjak, and Katja Lozar Manfreda. 2023. Item Nonresponse in Web Versus Other Survey Modes: A Systematic Review and Meta-Analysis. *Social Science Computer Review* 41, 3 (2023), 926–945. https://doi.org/10.1177/08944393211056229

[25] Peter Christen. 2012. *Data Matching*. Springer, Berlin/Heidelberg, Germany.

[26] Line H. Clemmensen and Rune D. Kjærsgaard. 2023. Data Representativity for Machine Learning and AI Systems. arXiv:2203.04706

[27] Carina Cornesse and Ines Schaurer. 2021. The Long-Term Impact of Different Offline Population Inclusion Strategies in Probability-Based Online Panels: Evidence From the German Internet Panel and the GESIS Panel. *Social Science Computer Review* 39, 4 (2021), 687–704.

[28] Louise Corti. 2012. Recent developments in archiving social research. *International Journal of Social Research Methodology* 15, 4 (2012), 281–290.

[29] Kathleen Creel and Deborah Hellman. 2022. The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems. *Canadian Journal of Philosophy* 52, 1 (2022), 26–43. https://doi.org/10.1017/can.2022.3

[30] Caroline Criado Perez. 2019. *Invisible Women: Exposing Data Bias in a World Designed for Men*. Vintage, Penguin Random House, London, UK.

[31] Jessica Daikeler, Michael Bošnjak, and Katja Lozar Manfreda. 2020. Web Versus Other Survey Modes: An Updated and Extended Meta-Analysis Comparing Response Rates. *Journal of Survey Statistics and Methodology* 8, 3 (2020), 513–539. https://doi.org/10.1093/jssam/smz008

[32] Vanessa Di Lego. 2023. Uncovering the gender health data gap. *Cadernos de Saúde Pública* 39, 7 (2023), e00065423. https://doi.org/10.1590/0102-311XEN065423

[33] Yixiao Dong and Denis Dumas. 2020. Are personality measures valid for different populations? A systematic review of measurement invariance across cultures, gender, and age. *Personality and Individual Differences* 160 (2020), 109956. https://doi.org/10.1016/j.paid.2020.109956

[34] David Dutwin and Trent D. Buskirk. 2023. A Deeper Dive into the Digital Divide: Reducing Coverage Bias in Internet Surveys. *Social Science Computer Review* 41, 5 (2023), 1902–1920. https://doi.org/10.1177/08944393221093467

[35] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12*. ACM Press, Cambridge, Massachusetts, 214–226. https://doi.org/10.1145/2090236.2090255

[36] Stephanie Eckman. 2016. Does the Inclusion of Non-Internet Households in a Web Panel Reduce Coverage Bias? *Social Science Computer Review* 34, 1 (2016), 41–58.

[37] Stephanie Eckman, Bolei Ma, Christoph Kern, Rob Chew, Barbara Plank, and Frauke Kreuter. 2025. Correcting Annotator Bias in Training Data: Population-Aligned Instance Replication (PAIR). arXiv:2501.06826

[38] Stephanie Eckman, Barbara Plank, and Frauke Kreuter. 2024. Position: Insights from Survey Methodology can Improve Training Data. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*, Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR, 12268–12283. https://proceedings.mlr.press/v235/eckman24a.html

[39] Yotam Elor and Hadar Averbuch-Elor. 2022. To SMOTE, or not to SMOTE? arXiv:2201.08528

[40] Kristin R. Eschenfelder, Kalpana Shankar, and Greg Downey. 2022. The financial maintenance of social science data archives: Four case studies of long-term infrastructure work. *Journal of the Association for Information Science and Technology* 73, 12 (2022), 1723–1740. https://doi.org/10.1002/asi.24691

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

[42] Marco Favier, Toon Calders, Sam Pinxteren, and Jonathan Meyer. 2023. How to be fair? a study of label and selection bias. *Machine Learning* 112, 12 (2023), 5081–5104.

[43] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. arXiv:2305.08283

[44] Jens Foerderer. 2023. Should we trust web-scraped data? arXiv:2308.02231

[45] Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation* (2024), 1–28. https://doi.org/10.1007/s10579-024-09766-4

[46] David O Friedrichs. 2019. White Collar Crime and the Class-Race-Gender Construct. In *Race, Gender, and Class in Criminology. The Intersections*, Martin D. Schwartz and Dragan Milovanovic (Eds.). Routledge, 141–158.

[47] Howard N. Garb. 2021. Race bias and gender bias in the diagnosis of psychological disorders. *Clinical Psychology Review* 90 (2021), 102087. https://doi.org/10.1016/j.cpr.2021.102087

[48] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92. https://doi.org/10.1145/3458723

[49] Michael Gelman, Shachar Kariv, Matthew D. Shapiro, Dan Silverman, and Steven Tadelis. 2014. Harnessing naturally occurring data to measure the response of spending to income. *Science* 345, 6193 (2014), 212–215.

[50] Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 1161–1166. https://doi.org/10.18653/v1/D19-1107

[51] Francesco Giusti, Carmen Martos, Annalisa Trama, Manola Bettio, Arantza Sanvisens, Riccardo Audisio, Volker Arndt, Silvia Francisci, Carine Dochez, Josepa Ribes, Laura Pareja Fernández, Anna Gavin, Gemma Gatta, Rafael Marcos-Gragera, Yolande Lievens, Claudia Allemani, Roberta De Angelis, Otto Visser, Liesbet Van Eycken, and the ENCR Working Group on Treatment Data Harmonisation . 2023. Cancer treatment data available in European cancer registries: Where are we and where are we going? *Frontiers in Oncology* 13 (2023), 1109978. https://doi.org/10.3389/fonc.2023.1109978

[52] Robert M Groves. 2011. 'Designed Data' and 'Organic Data'. https://www.census.gov/newsroom/blogs/director/2011/05/designed-data-and-organic-data.html US Census, Director's Blog, entry on 2011-05-31, accessed 2022-07-27.

[53] Robert M Groves. 2011. Three Eras of Survey Research. *Public Opinion Quarterly* 75, 5 (2011), 861–871.

[54] Robert M Groves, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey Methodology* (2nd ed.). John Wiley & Sons, Hoboken, NJ.

[55] Robert M Groves and Lars Lyberg. 2010. Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly* 74, 5 (2010), 849–879. https://doi.org/10.1093/poq/nfq065

[56] Robert M. Groves and Emilia Peytcheva. 2008. The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. *Public Opinion Quarterly* 72, 2 (2008), 167–189. https://doi.org/10.1093/poq/nfn011

[57] Cornelia Gruber, Patrick Oliver Schenk, Malte Schierholz, Frauke Kreuter, and Göran Kauermann. 2023. Sources of Uncertainty in Machine Learning – A Statisticians' View. arXiv:2305.16703

[58] Furkan Gursoy and Ioannis A. Kakadiaris. 2022. Equal Confusion Fairness: Measuring Group-Based Disparities in Automated Decision Systems. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. 137–146. https://doi.org/10.1109/ICDMW58026.2022.00027

[59] Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems* 24, 2 (2009), 8–12.

[60] Katarina Hamberg. 2008. Gender Bias in Medicine. *Women's Health* 4, 3 (2008), 237–243. https://doi.org/10.2217/17455057.4.3.237

[61] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 501—-512. https://doi.org/10.1145/3351095.3372826

[62] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, Stockholmsmässan, Stockholm, Sweden, 1939–1948.

[63] Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. 2023. Foundation Models and Fair Use. *Journal of Machine Learning Research* 24, 400 (2023), 1–79. http://jmlr.org/papers/v24/23-0569.html

[64] Thomas N. Herzog, Fritz J. Scheuren, and William E. Winkler. 2007. *Data Quality and Record Linkage Techniques*. Springer, New York, NY.

[65] Johanna Hölzl, Florian Keusch, and John Collins. 2024. Sample Inconsistencies of Google Trends Data Across Download Locations. https://doi.org/10.17605/OSF.IO/CD6RW

[66] Johanna Hölzl, Florian Keusch, and Christoph Sajons. 2024. The (mis) use of Google Trends data in the social sciences – A systematic review, critique, and recommendations. *Social Science Research* 126 (2024), 103099. https://doi.org/10.1016/j.ssresearch.2024.103099

[67] Itay Itzhak, Gabriel Stanovsky, Nir Rosenfeld, and Yonatan Belinkov. 2023. Instructed to bias: instruction-tuned language models exhibit emergent cognitive bias. arXiv:2308.00225

[68] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, 375—385. https://doi.org/10.1145/3442188.3445901

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

[70] Lilli Japec, Frauke Kreuter, Marcus Berg, Paul P Biemer, Paul Decker, Cliff Lampe, Julia Lane, Cathy O'Neil, and Abe Usher. 2015. Big Data in Survey Research: AAPOR Task Force Report. *Public Opinion Quarterly* 79, 4 (2015), 839–880.

[71] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, 306—-316. https://doi.org/10.1145/3351095.3372829

[72] Alexandra D. Kaplan, Theresa T. Kessler, J. Christopher Brill, and P. A. Hancock. 2023. Trust in Artificial Intelligence: Meta-Analytic Findings. *Human Factors* 65, 2 (2023), 337–359. https://doi.org/10.1177/00187208211013988

[73] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2021. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. arXiv:2010.04050

[74] Alexander Katalinic, Marco Halber, Martin Meyer, Maren Pflüger, Andrea Eberle, Alice Nennecke, Soo-Zin Kim-Wanner, Tobias Hartz, Kerstin Weitmann, Andreas Stang, Christina Justenhoven, Bernd Holleczek, Daniela Piontek, Ian Wittenberg, Annika Heßmer, Klaus Kraywinkel, Claudia Spix, and Ron Pritzkuleit. 2023. Population-Based Clinical Cancer Registration in Germany. *Cancers* 15, 15 (2023). https://doi.org/10.3390/cancers15153934

[75] Christoph Kern, Stephanie Eckman, Jacob Beck, Rob Chew, Bolei Ma, and Frauke Kreuter. 2023. Annotation Sensitivity: Training Data Collection Methods Affect Model Performance. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 14874–14886. https://doi.org/10.18653/v1/2023.findings-emnlp.992

[76] Christoph Kern, Frederic Gerdon, Ruben L. Bach, Florian Keusch, and Frauke Kreuter. 2022. Humans versus machines: Who is perceived to decide fairer? Experimental evidence on attitudes toward automated decision-making. *Patterns* 3, 10 (2022), 100591.

https://doi.org/10.1016/j.patter.2022.100591

[77] Florian Keusch and Frauke Kreuter. 2021. Digital trace data: Modes of data collection, applications, and errors at a glance. In *Handbook of Computational Social Science, Vol 1*, Uwe Engel, Anabel Quan-Haase, Sunny Xun Liu, and Lars Lyberg (Eds.). Taylor & Francis. https://doi.org/10.4324/9781003024583-8

[78] Michael P. Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. Association for Computing Machinery, New York, NY, USA, 247–254. https://doi.org/10.1145/3306618.3314287

[79] Yoonsang Kim, Rachel Nordgren, and Sherry Emery. 2020. The Story of Goldilocks and Three Twitter's APIs: A Pilot Study on Twitter Data Sources and Disclosure. *International Journal of Environmental Research and Public Health* 17, 3 (2020), 864. https://doi.org/10.3390/ijerph17030864

[80] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. arXiv:2404.16019

[81] Lauren Klein and Catherine D'Ignazio. 2024. Data Feminism for AI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, 100–112. https://doi.org/10.1145/3630106.3658543

[82] Ulrich Kohler, Frauke Kreuter, and Elizabeth A Stuart. 2019. Nonprobability Sampling and Causal Analysis. *Annual Review of Statistics and Its Application* 6, 1 (2019), 149–172. https://doi.org/10.1146/annurev-statistics-030718-104951

[83] Dagmar Krebs and Jan Karem Höhne. 2020. Exploring Scale Direction Effects and Response Behavior across PC and Smartphone Surveys. *Journal of Survey Statistics and Methodology* 9, 3 (2020), 477–495. https://doi.org/10.1093/jssam/smz058

[84] Frauke Kreuter, Gerrit Müller, and Mark Trappmann. 2014. A Note on Mechanisms Leading to Lower Data Quality of Late or Reluctant Respondents. *Sociological Methods & Research* 43, 3 (2014), 452–464. https://doi.org/10.1177/004912411350809

[85] Matthias Kuppler, Christoph Kern, Ruben Bach, and Frauke Kreuter. 2022. From fair predictions to just decisions? Conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making. *Frontiers in Sociology* 7 (2022). https://doi.org/10.3389/fsoc.2022.883999

[86] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science* 343, 6176 (2014), 1203–1205. https://doi.org/10.1126/science.1248506

[87] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. 2022. A survey on datasets for fairness-aware machine learning. *WIREs Data Mining and Knowledge Discovery* 12, 3 (2022), e1452. https://doi.org/10.1002/widm.1452

[88] Yelda A Leal, Leticia M Fernández-Garrote, Alejandro Mohar-Betancourt, and Abelardo Meneses-García. 2016. The importance of registries in cancer control. *Salud Pública de México* 58, 2 (2016), 309–316. https://www.scielo.org.mx/pdf/spm/v58n2/0036-3634-spm-58-02-00309.pdf

[89] Fumin Li, Yisu Zhou, and Tianji Cai. 2021. Trails of Data: Three Cases for Collecting Web Information for Social Science Research. *Social Science Computer Review* 39, 5 (2021), 922–942. https://doi.org/10.1177/0894439319886019

[90] Roderick JA Little and Donald B Rubin. 2019. *Statistical analysis with missing data*. John Wiley & Sons, Hoboken, NJ.

[91] Edward R Maguire and Rebecca Schulte-Murray. 2001. Issues and Patterns in the Comparative International Study of Police Strength . *International Journal of Comparative Sociology* 42, 1/2 (2001), 75–100. https://doi.org/10.1163/156851801300171715

[92] Roland Maio. 2024. Machine-Learning Fairness in Data Markets: Challenges and Opportunities. PhD thesis, Columbia University. https://doi.org/10.7916/keek-wb58

[93] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2021. Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management* 5 (2021), 102642. https://doi.org/10.1016/j.ipm.2021.102642

[94] Alessandro Mantelero. 2018. AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review* 34, 4 (Aug. 2018), 754–772. https://doi.org/10.1016/j.clsr.2018.05.017

[95] Viktor Mayer-Schönberger and Kenneth Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt, Boston, MA.

[96] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6 (2021), 1–36. https://doi.org/10.1145/3457607

[97] Katharina M. Meitinger and Timothy P. Johnson. 2020. *Power, Culture and Item Nonresponse in Social Surveys*. Springer, Cham, Switzerland, 169–191. https://doi.org/10.1007/978-3-030-47256-6_8

[98] Danaë Metaxa, Michelle A. Gan, Su Goh, Jeff Hancock, and James A. Landay. 2021. An Image of Society: Gender and Racial Representation and Impact in Image Search Results for Occupations. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 26 (2021). https://doi.org/10.1145/3449100

[99] Ishan Misra, C Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. 2016. Seeing Through the Human Reporting Bias: Visual Classifiers From Noisy Human-Centric Labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2930–2939. https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Misra_Seeing_Through_the_CVPR_2016_paper.html

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

[101] Kari Lock Morgan and Donald B. Rubin. 2012. Rerandomization to improve covariate balance in experiments. *The Annals of Statistics* 40, 2 (2012), 1263 – 1282. https://doi.org/10.1214/12-AOS1008

[102] Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. arXiv:2004.09456

[103] National Academies of Sciences, Engineering, and Medicine (NAS). 2022. *Transparency in Statistical Information for the National Center for Science and Engineering Statistics and All Federal Statistical Agencies*. The National Academies Press, Washington, DC. https://doi.org/10.17226/26360

[104] Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. arXiv:2103.14749

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

[106] Kristen Olson. 2006. Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias. *Public Opinion Quarterly* 70, 5 (2006), 737–758. https://doi.org/10.1093/poq/nfl038

[107] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data* 2 (2019). https://doi.org/10.3389/fdata.2019.00013

[108] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (2021), 100336. https://doi.org/10.1016/j.patter.2021.100336

[109] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative Prediction. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 7599–7609. https://proceedings.mlr.press/v119/perdomo20a.html

[110] X Developer Platform. [n. d.]. X API v2: Timelines. https://developer.x.com/en/docs/x-api/tweets/timelines/introduction. Accessed: 2024-12-01.

[111] Nabeel Qureshi, Maria Edelen, Lara Hilton, Anthony Rodriguez, Ron D Hays, and Patricia M Herman. 2022. Comparing data collected on Amazon's Mechanical Turk to national surveys. *American Journal of Health Behavior* 46, 5 (2022), 497–502. https://doi.org/10.5993/AJHB.46.5.1

[112] Vikram V. Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron B. Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. 2023. GeoDE: a Geographically Diverse Evaluation Dataset for Object Recognition. arXiv:2301.02560

[113] Susanne Rässler. 2016. Data Fusion: Identification Problems, Validity, and Multiple Imputation. *Austrian Journal of Statistics* 33, 1&2 (2016), 153—171. https://doi.org/10.17713/ajs.v33i1&2.436

[114] Filipe N. Ribeiro, Fabrício Benevenuto, and Emilio Zagheni. 2020. How Biased is the Population of Facebook Users? Comparing the Demographics of Facebook Users with Census Data to Generate Correction Factors. In *Proceedings of the 12th ACM Conference on Web Science* (Southampton, United Kingdom) *(WebSci '20)*. Association for Computing Machinery, New York, NY, 325—334. https://doi.org/10.1145/3394231.3397923

[115] Aya Rizk and Ida Lindgren. 2024. Automated Decision-Making in the Public Sector: A Multidisciplinary Literature Review. In *International Conference on Electronic Government (EGOV 2024, Vol. 14841)*, Marijn Janssen, Joep Crompvoets, J Ramon Gil-Garcia, Habin Lee, Ida Lindgren, Anastasija Nikiforova, and Gabriela Viale Pereira (Eds.). Springer, Cham, Switzerland, 237–253. https://doi.org/10.1007/978-3-031-70274-7_15

[116] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, et al. 2021. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence* 3, 3 (2021), 199–217.

[117] Esther Rolf, Theodora T Worledge, Benjamin Recht, and Michael Jordan. 2021. Representation matters: Assessing the importance of subgroup allocations in training data. In *International Conference on Machine Learning*. PMLR, 9040–9051.

[118] Derek Ruths and Jürgen Pfeffer. 2014. Social media for large studies of behavior. *Science* 346, 6213 (2014), 1063–1064. https://doi.org/10.1126/science.346.6213.1063

[119] Joseph W Sakshaug, Alexandra Schmucker, Frauke Kreuter, Mick P Couper, and Eleanor Singer. 2019. Effect of Framing and Placement on Linkage Consent. *Public Opinion Quarterly* 83 (2019), 289 – 308.

[120] Joseph W. Sakshaug, Jens Stegmaier, Mark Trappmann, and Frauke Kreuter. 2019. Does Benefit Framing Improve Record Linkage Consent Rates? A Survey Experiment. *Survey Research Methods* 13, 3 (2019), 289–304. https://doi.org//10.18148/srm/2019.v13i3.7391

[121] Matthew J Salganik. 2017. *Bit by bit: Social research in the digital age*. Princeton University Press, Princeton, NJ.

[122] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work"': Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, Article 39, 15 pages. https://doi.org/10.1145/3411764.3445518

[123] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect?. In *International Conference on Machine Learning*. PMLR, 29971–30004.

[124] Patrick Oliver Schenk and Christoph Kern. 2024. Connecting Algorithmic Fairness to Quality Dimensions in Machine Learning in Official Statistics and Survey Production. *AStA Wirtschafts-und Sozialstatistisches Archiv* 18 (2024), 131–184. https://doi.org/10.1007/s11943-024-00344-2

[125] Patrick Oliver Schenk, Christoph Kern, and Frauke Kreuter. 2025. Fairness in machine learning for national statistical organizations. In *Foundations and Advances of Machine Learning in Official Statistics*, Florian Dumpert (Ed.). Springer.

[126] Patrick Oliver Schenk and Simone Reuß. 2024. Paradata in Surveys. In *Perspectives on Paradata: Research and Practice of Documenting Process Knowledge*, Isto Huvila, Lisa Andersson, and Olle Sköld (Eds.). Springer, Cham, Switzerland, 15–43. https://doi.org/10.1007/978-3-031-53946-6_2

[127] Ari Schlesinger, W. Keith Edwards, and Rebecca E. Grinter. 2017. Intersectional HCI: Engaging Identity through Gender, Race, and Class. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, 5412—5427. https://doi.org/10.1145/3025453.3025766

[128] Carsten Schwemmer, Carly Knight, Emily D. Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W. Lockhart. 2020. Diagnosing Gender Bias in Image Recognition Systems. *Socius* 6 (2020). https://doi.org/10.1177/2378023120967171

[129] Indira Sen, Fabian Flöck, Katrin Weller, Bernd Weiß, and Claudia Wagner. 2021. A Total Error Framework for Digital Traces of Human Behavior on Online Platforms. *Public Opinion Quarterly* 85, S1 (2021), 399–422.

[130] Sina Shaham, Arash Hajisafi, Minh K. Quan, Dinh C. Nguyen, Bhaskar Krishnamachari, Charith Peris, Gabriel Ghinita, Cyrus Shahabi, and Pubudu N. Pathirana. 2025. Privacy and Fairness in Machine Learning: A Survey. *IEEE Transactions on Artificial Intelligence* (2025), 1–21. https://doi.org/10.1109/TAI.2025.3531326

[131] Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and H. V. Jagadish. 2023. Representation Bias in Data: A Survey on Identification and Resolution Techniques. *Comput. Surveys* 55, 13s, Article 293 (2023), 39 pages. https://doi.org/10.1145/3588433

[132] Ben Shneiderman, Maxine Cohen, Steven Jacobs, Catherine Plaisant, Nicholas Diakopoulos, and Niklas Elmqvist. 2017. *Designing the User Interface Strategies for Effective Human-Computer Interaction* (6th ed.). Pearson.

[133] Asif H Siddiqui and Syed Nabeel Zafar. 2018. Global Availability of Cancer Registry Data. *Journal of Global Oncology* 4 (2018). https://doi.org/10.1200/JGO.18.00116

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

[135] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2022. Participation Is not a Design Fix for Machine Learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22)*. Association for Computing Machinery, New York, NY, Article 1, 6 pages. https://doi.org/10.1145/3551624.3555285

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

[137] Isabel Straw, Geraint Rees, and Parashkev Nachev. 2024. Sex-Based Performance Disparities in Machine Learning Algorithms for Cardiac Disease Prediction: Exploratory Study. *Journal of Medical Internet Research* 26 (2024), e46936. https://doi.org/10.2196/46936

[138] Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (New York, NY) *(EAAMO '21)*. Association for Computing Machinery, New York, NY, Article 17, 9 pages. https://doi.org/10.1145/3465416.3483305

[139] Harini Suresh and John V. Guttag. 2020. A Framework for Understanding Unintended Consequences of Machine Learning. arXiv:1901.10002

[140] Daria Szafran and Ruben L Bach. 2024. "The Human Must Remain the Central Focus"': Subjective Fairness Perceptions in Automated Decision-Making. *Minds and Machines* 34, 3 (2024), 24. https://doi.org/10.1007/s11023-024-09684-y

[141] George Terhanian and John Bremer. 2012. A Smarter Way to Select Respondents for Surveys? *International Journal of Market Research* 54, 6 (2012), 751–780. https://doi.org/10.2501/IJMR-54-6-751-780

[142] Vera Toepoel and Yanniek Hendriks. 2016. The impact of non-coverage in web surveys in a country with high internet penetration: Is it (still) useful to provide equipment to non-internet households in the Netherlands? *International Journal of Internet Science* 11, 1 (2016).

[143] Ciaran B Trace and James A Hodges. 2024. The Role of Paradata in Algorithmic Accountability. In *Perspectives on Paradata: Research and Practice of Documenting Process Knowledge*, Isto Huvila, Lisa Andersson, and Olle Sköld (Eds.). Springer, Cham, Switzerland, 197–213. https://doi.org/10.1007/978-3-031-53946-6_11

[144] Cuong Tran, Ferdinando Fioretto, Pascal Van Hentenryck, and Zhiyan Yao. 2021. Decision Making with Differential Privacy under a Fairness Lens. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 560–566. https://doi.org/10.24963/ijcai.2021/78 Main Track.

[145] Richard Valliant, Jill A Dever, and Frauke Kreuter. 2018. *Practical tools for designing and weighting survey samples* (2nd ed.). Springer, Cham, Switzerland.

[146] Ruben van den Goorbergh, Maarten van Smeden, Dirk Timmerman, and Ben Van Calster. 2022. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *Journal of the American Medical Informatics Association* 29, 9 (2022), 1525–1534. https://doi.org/10.1093/jamia/ocac093

[147] Niek Van Wettere. 2022. How to implement the metadata standard "Data Documentation Initiative (DDI)"? https://doi.org/10.5281/zenodo.5841664

[148] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E. Hines, John P. Dickerson, and Chirag Shah. 2022. Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. arXiv:2010.10596

[149] Gregory Vial. 2019. Reflections on quality requirements for digital trace data in IS research. *Decision Support Systems* 126 (2019), 113–133. https://doi.org/10.1016/j.dss.2019.113133

[150] Kelly Walters, Dimitri A Christakis, and Davene R Wright. 2018. Are Mechanical Turk worker samples representative of health status and health behaviors in the US? *PLOS One* 13, 6 (2018), e0198835. https://doi.org/10.1371/journal.pone.0198835

[151] Brady T. West, James Wagner, Jinseok Kim, and Trent D. Buskirk. 2023. The Total Data Quality Framework. https://www.coursera.org/specializations/total-data-quality.

[152] Ting Yan and Richard Curtin. 2010. The Relation Between Unit Nonresponse and Item Nonresponse: A Response Continuum Perspective. *International Journal of Public Opinion Research* 22, 4 (2010), 535–551. https://doi.org/10.1093/ijpor/edq037

[153] Vesna Zadnik. 2023. Europe. Update for International Association of Cancer Registries 2022-2023. https://www.encr.eu/sites/default/files/inline-files/EUROPE_IACR%202023_report_final.pdf