AStA Wirtschafts- und Sozialstatistisches Archiv (2024) 18:131–184
https://doi.org/10.1007/s11943-024-00344-2

ORIGINALVERÖFFENTLICHUNG

Connecting algorithmic fairness to quality dimensions in machine learning in official statistics and survey production

Patrick Oliver Schenk · Christoph Kern

Received: 5 May 2023 / Accepted: 23 July 2024 / Published online: 7 October 2024

Abstract National Statistical Organizations (NSOs) increasingly draw on Machine Learning (ML) to improve the timeliness and cost-effectiveness of their products. When introducing ML solutions, NSOs must ensure that high standards with respect to robustness, reproducibility, and accuracy are upheld as codified, e.g., in the Quality Framework for Statistical Algorithms (QF4SA; Yung et al. 2022, *Statistical Journal of the IAOS*). At the same time, a growing body of research focuses on fairness as a pre-condition of a safe deployment of ML to prevent disparate social impacts in practice. However, fairness has not yet been explicitly discussed as a quality aspect in the context of the application of ML at NSOs. We employ the QF4SA quality framework and present a mapping of its quality dimensions to algorithmic fairness. We thereby extend the QF4SA framework in several ways: First, we investigate the interaction of fairness with each of these quality dimensions. Second, we argue for fairness as its own, additional quality dimension, beyond what is contained in the QF4SA so far. Third, we emphasize and explicitly address data, both on its own and its interaction with applied methodology. In parallel with empirical illustrations, we show how our mapping can contribute to methodology in the domains of official statistics, algorithmic fairness, and trustworthy machine learning.

Little to no prior knowledge of ML, fairness, and quality dimensions in official statistics is required as we provide introductions to these subjects. These introductions are also targeted to the discussion of quality dimensions and fairness.

✉ Patrick Oliver Schenk · Christoph Kern
Department of Statistics, LMU Munich, Ludwigstr. 33, 80539 Munich, Germany
E-Mail: patrick.schenk@stat.uni-muenchen.de; p.o.s.on.stats@gmail.com

Christoph Kern
E-Mail: christoph.kern@stat.uni-muenchen.de

Christoph Kern
Munich Center for Machine Learning (MCML), Munich, Germany

Keywords Algorithmic Fairness · Quality Dimensions · Machine Learning · Official Statistics · Trustworthy Machine Learning

JEL classification C80 · C01 · C52 · C53 · Y80

1 Introduction

Official Statistics, Other Data Producers, and Machine Learning Machine Learning (ML, see Table 1 for a list of abbreviations) is now widely used in government, state, federal, and similar agencies (Engstrom et al. 2020; IPS Observatory 2024; TAG Register 2024; AlgorithmWatch 2019; Domscheit-Berg 2024). Official Statistics, e.g., in International, State, and National Statistical Organizations (NSOs for short), is one such area (see Beck et al. 2018a, Chap. 2 and Sect. 2). The introduction of ML can be seen as part of the modernization efforts at NSOs: these happen on the (cross-)organizational level (e.g., the UNECE High-Level Group for the Modernisation of Official Statistics, see https://statswiki.unece.org/display/hlgbas) but also within organizations because of their mandates for ongoing revision of methods, data sources, and products and, more indirectly, because of their operating principles of e.g., timeliness, and cost-effectiveness (Eurostat 2017). In addition, there is increased competition from other producers of data and of statistics who offer products that are, e.g., new or more timely, often made possible by gained innovation advantages or because they are less bound by quality principles (Julien 2020, Chap. 2). Thus, NSOs strive to improve by offering new or refined products (i.e., data or statistics). The latter can be described as doing better on at least one of their quality dimensions (see Sect. 4) and not (meaningfully) worse on the others (Julien 2020, p. 12): e.g., producing the 'same' data more cheaply or releasing the 'same' statistic more timely. In this endeavor, new methods, particularly ML, and new data sources, including those that require ML, are not an end in themselves but must serve the business needs of NSOs and, by extension, their audience (Measure 2020, p. 6), with demonstrated added value (Julien 2020, p. 1).

While the exact conditions for and tasks of NSOs may be laid down in local laws (e.g., the German *BStatG*), by and large NSOs follow the fundamental principles adopted by the United Nations Economic and Social Council (UNECE 2013) and the European Statistics Code of Practice (Eurostat 2017): NSOs are impartial, credible producers of relevant data and of statistics, based on high professional, ethical, and scientific standards, working according to quality dimensions that we consider in Sect. 4. The same dual roles (producers and analysts of data), a similar audience, and many of the same quality considerations are shared by (the respective data units within), e.g., central banks, federal research institutes, research data centers, units of governmental departments (e.g., the Bureau of Labor Statistics within the U.S. Department of Labor), and the big producers of administrative data such as government unemployment services, pension funds, and health insurance providers.[1] There is also great overlap with the survey world: NSOs are one of the big conductors

[1] The data analyses of NSOs may tend to be more basic and descriptive.

and sponsors of surveys and have contributed much to survey methodology. Also, quality considerations for surveys and official statistics have much in common (compare Groves et al. 2009, Chap. 2.6 and Sect. 4). Even more importantly, several of the applications of ML in NSOs that we consider (see Sect. 2.4) specifically concern the administration of surveys and the processing of survey data. Therefore, much of what we will discuss also applies to survey organizations and other data producers, so that they are included here when, for brevity, we only speak of NSOs.

Fairness In parallel to the introduction of ML at NSOs, the increasing use of prediction algorithms in the private and in the public sector has sparked a wide range of research on algorithmic fairness[2]. We posit that algorithmic fairness is of particular relevance to NSOs as it touches on ethical considerations, quality dimensions, and their interactions. The importance of (algorithmic) bias and fairness for the work of NSOs is not completely unrecognized (e.g., Helwegen and Braaksma 2020), but treatments are sparse and more high-level. Also, fairness may get subsumed under ethics or there may be a general discussion of ethics, but not algorithmic fairness in particular (e.g., UK Statistics Authority 2021). Considering fairness solely within legal mandates and as an aspect of ethics (Julien 2020, p. 6f.) also influences the types of fairness one considers as well as, e.g., which groups are investigated. This may not be the most suitable approach for the work of NSOs. Therefore, we suggest considering fairness also within the quality frameworks of NSOs: both, as its own quality dimension and how it interacts with the other quality dimensions. In this paper, we highlight these interactions by discussing how each quality dimension of the Quality Framework for Statistical Algorithms (QF4SA; Yung et al. 2022) maps to algorithmic fairness.

Fairness is not the only relevant dimension beyond performance (Yung et al. 2022, p. 8). Frameworks such *Trustworthy ML* aim for explainable, fair, privacy-preserving, causal, and robust systems (Varshney et al. 2022; TrustML 2024).[3] Thus, the overlap to the quality dimensions for NSOs as described by the QF4SA is very strong (see Sect. 4).

Contribution Mapping quality dimensions of official statistics to fairness considerations leads to contributions that are of relevance to both communities. Our contribution to the literature for official statistics, survey organizations, and similar data producers includes expanding the current QF4SA framework by highlighting connections to the extensive literature on algorithmic fairness and illustrating how these connections may be exploited in practice. We thereby shed new light on the existing quality dimensions with a particular focus on how they can cater towards a safe

[2] No connection to the acronym FAIR (i.e., findability, accessibility, interoperability, and reusability of data; Wilkinson et al. 2016).

[3] There exist similar concepts and terms such as ethical or responsible computing/ML/AI (see Díaz-Rodríguez et al. 2023 for an overview and mapping of different concepts in trustworthy ML). As these terms are unfortunately not used consistently, we can only point the interested reader to Díaz-Rodríguez et al. (2023) and TrustML (2024) as starting points.

deployment of ML at NSOs from a fairness perspective. The proposed mapping[4] can sharpen requirements that are made in the current QF4SA – e.g., by expanding overall accuracy assessments to notions of multi-group fairness. Nonetheless, the roles, goals, and tasks of NSOs in part differ from what the traditional fair ML literature focuses on. Thus, ML applications in NSO in turn motivate contributions to the algorithmic fairness literature: e.g., we suggest the use of heterogeneity-finding ML machinery for more fair reporting of results, for finding unfairly treated groups, and for finding biases in the data (production process). We further discuss fairness implications of (temporal) data drift, a setting that can be common with administrative data sources of NSOs, but less frequently considered in the fairness in ML community. Our contribution to the Trustworthy ML literature is the discussion of the interconnections among the quality dimensions, both with and without fairness.

Structure and Overview Given the heterogeneity of audiences and backgrounds, we strive for a self-contained article so that detailed background knowledge on the topics of our article is not required. Readers who are completely new to ML might want to briefly familiarize themselves with the easy-to-understand decision trees (e.g., Molnar 2020, Chap. 5.4) as we use them as examples on several occasions. We begin (see Fig. 1) by providing background on ML (Sect. 2.1) and distinguishing procedural from methodological benefits of ML (Sect. 2.2). After illustrating the main drivers behind the interest in ML by NSOs (Sect. 2.3), we highlight ML applications (Sect. 2.4). Next, we provide background on algorithmic fairness (Sect. 3.1) and the human components in fair ML (Sect. 3.2). We note that these introductions

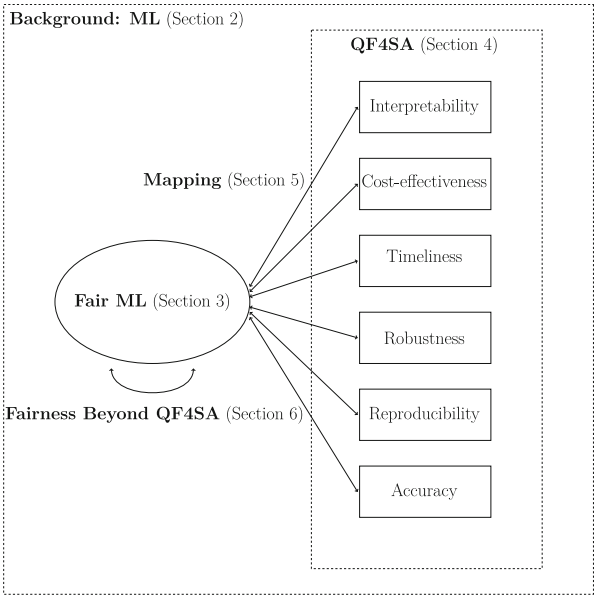

Fig. 1 Structure and Overview

[4] We kindly ask the mathematically inclined reader not to take the term 'mapping' in the most formal sense.

in Sects. 2 and 3 are explicitly tailored to our subsequent study of fair machine learning in official statistics.

We proceed by summarizing the quality dimensions of the QF4SA framework and their interconnections (Sect. 4), before turning to our mapping of quality dimensions and fairness (Sect. 5). Our mapping is complemented by a presentation of fairness considerations that extend beyond the dimensions of the QF4SA framework (Sect. 6). We close with a discussion and outlook (Sect. 7).

2 Background: Machine Learning

We provide a (high-level) introduction to ML, targeted to the focus of this article. As we focus on ML, readers who are interested in a comparison of the "two cultures" – that is, traditional statistics and machine learning – beyond what we provide in Sect. 2.2 are pointed to Breiman (2001)'s seminal articulation of this topic, recently reflected upon by Raper et al. (2020) and in a 2021 special issue of *Observational Studies* edited by Nandita Mitra (2021).

2.1 ML and Statistics, Supervised and Unsupervised Learning

We discuss the use of ML in NSOs in their roles as producers and analysts of data.[5] Most if not all of these ML applications fall under either supervised learning or unsupervised learning.[6] In *supervised learning*, the goal is to learn the functional relationship f between inputs or *features* $X_1, \ldots, X_p$ and the outcome or *label* Y, both of which are contained in the training data. In the supervised ML paradigm, illuminated further in Sect. 2.2, the focus is on *prediction*, i.e., the ability to predict the outcome from the feature values for new data points (i.e., not used for training the model): $\hat{y} = f(x_1, \ldots, x_p)$.[7] In other words, the rationale for supervised ML is *deployment*: being able to use the learned model $\hat{f}$ to predict the outcome for new units for which the outcome is unknown. This is in contrast with the traditional inferential statistics approach: there, the goal is the *estimation* of (population) parameters θ, typically in order to answer substantive questions about the world which were translated into statistical parameters, $\mathbb{E}(y|x_1, \ldots, x_p) = f(x_1, \ldots, x_p; \theta)$. Additionally, compared to prediction, population-based inferential statistics is less focused on the individual (Breiman 2001).

Supervised learning tasks model either a qualitative outcome (called *classification*) or a quantitative outcome (called *regression* in the ML world, regardless of

[5] Similar to others (e.g., Beck et al. 2018a and Beck et al. 2018b, Chap. 1.2), we focus on ML and neither AI that is not ML nor ML for other uses, such as virtual assistants facilitating users' interaction with a NSO's website or data. We briefly comment on current developments such as Large Language Models (LLMs) in Sect. 2.4.4.

[6] We are not aware of applications using reinforcement learning which is often distinguished as a third category (e.g., Molnar 2022, Chap. 10) and we think that NSOs' work typically does not lend itself to this approach.

[7] Note that the term prediction is not used in the sense of making statements about the future, although some supervised learning tasks are such forecasting tasks.

whether statistical regression or other methods are used). Classification predominates in the theoretical literatures (including that on fairness, see Sect. 3), in ML applications in general, and also in ML applications within NSOs (e.g., Beck et al. 2018a, Chap. 4).

Unsupervised learning comprises a very heterogeneous set of methods and tasks: clustering, dimensionality reduction, and outlier/anomaly detection, but also latent variables, archetypes, association rule learning, and more (Molnar 2022, Chap. 9). In contrast to supervised learning, unsupervised tasks lack an outcome variable Y in the data; only $X_1, \ldots, X_p$ are available. This also means that measuring performance is much more difficult. The common mission in the diverse collection of unsupervised tasks is to "find hidden patterns" (ibid) or to "discover interesting things" (James et al. 2021, Chap. 12) about the units, about the features, or, more generally, about the data. It is possible that the output of unsupervised models is the endpoint of the data analysis: e.g., one may be satisfied to learn how many 'groups' a clustering algorithm has detected or whether particular units are placed in the same cluster. Often, however, unsupervised methods are applied to pre-process or transform the data before they are fed into other, typically supervised models (James et al. 2021, Chap. 12): e.g., a high-dimensional set of variables can be reduced to a few, more manageable, perhaps more interpretable set of 'principal components' which are then employed as features in a prediction model (Bach et al. 2022). For some applications, both supervised and unsupervised learning may be useful, depending on the particular situation, goals, and available data: e.g., in the identification of the same units in two disparate data sources, one may or may not have gold-standard information about true matches via a unique identifier.

2.2 The Machine Learning Mindset, Procedural and Methodological Benefits, and a Comparison to Statistics

We believe it is important to reflect upon how ML proceeds, how and why ML is successful, and what the field has brought to data analyses in general. Only based on this explicitly articulated understanding does it make sense to discuss how ML relates to the quality dimensions of NSOs and to fairness. ML-based data analysis is characterized by two somewhat separate aspects. First, the *ML mindset*, paradigm, or approach which, along with its procedural contributions, we discuss in the next two paragraphs. Second, the increased use of *ML methods* (or model classes) in a stricter sense, which is discussed in the subsequent paragraph. Some model classes clearly fall under traditional statistical methods[8] and others are considered machine learning in a stricter sense (e.g., decision trees, random forests, and neural networks). Thus, while there is no sharp, universally agreed-upon boundary separating the two, the notion of ML methods is still useful. The following discussion of why and how ML succeeds and the contributions it has brought shall be informative in several ways: to help decide whether a ML mindset and ML-based methods fit a particular

[8] By traditional or classical statistical methods, we and others (e.g., Dumpert 2020, p. 8) do not mean historical or outdated methods but those that are part of statistical education: e.g. linear regression, but also generalized linear models, additive models, and so on.

application in official statistics and as necessary background information for our discussion of ML in official statistics.

With ML *procedures* we refer to four practices, discussed in the next paragraph, that, empirically, are most associated with sound ML-based data analysis, but that work largely independently from whether the considered model classes are statistical or ML. At the heart of the ML mindset is the rigorous evaluation of performance on a particular task.[9] In supervised learning, evaluation is about how well a model is able to predict on new data, i.e., the expected out-of-sample prediction error (*generalization error*). Thus, the ML mindset is one of competition, with the best-performing model being chosen. One typically considers several model classes (e.g., linear regression, LASSO, decision trees, random forests, and XGBoost) and several models within each class. The latter correspond to different selected features and 'parameters' (e.g., the features and splits in a decision tree, respectively) and different tuning or hyperparameters (e.g., the depth in a decision tree or the regularization penalty in LASSO; Hastie et al. 2009, Chaps. 3.4 and 9).

First, for both, model selection and assessment of the final model, unbiased estimation of a model's generalization error is paramount. The main procedural building block helping to ensure this is *data splitting* into two parts: the *training data*, which are only used for training the model, and the *evaluation data*, which are only used to evaluate the predictive performance of the trained model.[10] The error on the training data is a systematically over-optimistic measure. In contrast, the error on the separate, fresh evaluation data provides a valid estimate of the generalization error which also guards against overfitting (i.e., fitting too closely to the observed data, thus fitting partly to random noise inherent in the training observations). A second procedural contribution from supervised ML is that information that, in reality, would not be available at the time of the prediction may typically not be used during model training and previous steps (see Ghani and Schierholz 2020, Chap. 7.8.1 and Guts 2020): *Data leakage* occurs when any information from the supposedly separate, unseen evaluation data is used in some form, hurting the freshness of the evaluation data. *Target leakage* is about using the values of the outcome Y.[11] Both open the door for over-optimistic performance evaluations. Leakage can sneak in very subtly, as when the information is used for pre-processing the data, e.g., in feature engineering or imputation of missing data. Third, the centrality of performance comparisons in the ML approach brought focus to the issue of *metrics* used for model evaluation and during model training (i.e., the loss function to be optimized).[12] This is not unrecognized in traditional statistics, but the choice of metrics is more active and task-driven in the ML approach. In classification in particular, false negatives may be

[9] For explanations and comparisons of the different cultures and mindsets in data analysis, see Molnar (2022, Chaps. 2, 7, and 8).

[10] In model training, this typically takes the form of repeated data splitting via cross-validation. In model assessment, there is one split into training data and test data. See Hastie et al. (2009, Chap. 7).

[11] In addition, there is a practical concern: such information may be available during training, but it would not be available when the model is actually deployed – after all, one trains a prediction model precisely because the deployment data do not contain Y.

[12] It is possible to use a different metric during training than for model assessment, e.g., for computational reasons.

much more important relative to false positives for one application than for another. Recognizing the trade-off between the two error rates and choosing a task-suitable metric is an improvement over always employing overall accuracy. Fourth, as the ML approach involves the consideration of several models, one natural procedural extension was to combine several models, sometimes from different model classes, into one *ensemble* (via, e.g., bagging, boosting, stacking, or simpler methods such as averaging or majority vote; see Hastie et al. 2009, Chap. 8). The intuition for improved predictive performance is two-fold: for different data points a different model (class) may be closest to the truth and an ensemble of models is more stable than any one single model would be.

We now turn to ML *methods* or model classes and three of their reputed benefits, particularly relative to traditional statistical methods. First, flexibility, which in supervised learning is about the functional forms of the relationships between the features and the outcome as well as about interactions among the features. This actually entails two components:

(a) the ability to accommodate complex functional forms, which pertains mainly to quantitative features, and
(b) the automatic recognition of the (approximate) functional form and of interactions.

The former is afforded by ML methods that can be quite complex; however, typically the more flexible, the more data are required (James et al. 2021, Chap. 2.1.2). A further reason for (a) lies in the ML-based *approach* 'trying out' many ML model classes. For a different true functional form, a different model class is the most natural fit: e.g., trees are most suitable for step functions.[13] However, to compare the whole basket of ML model classes with just one statistical model class and conclude that statistics (every statistical method) is less flexible than ML (every ML method) is not fair. It also not accurate: in particular, Generalized Additive Models (James et al. 2021, Chap. 7) are a statistical model class that is able to automatically adapt to non-linear relationships and can accommodate interactions. In comparison studies, particularly for small and medium sample sizes, simpler and traditional statistical methods are often not inferior (Christodoulou et al. 2019; Grinsztajn et al. 2022).[14] Thus, even for performance reasons alone, simple methods and traditional statistical model classes should always be among those tried out; we will address other rationales such as interpretability in Sect. 4. Second, automatic feature selection is built into some ML model classes: e.g., in trees, at each split, only one variable is chosen. In high-dimensional settings, feature selection helps to stabilize the model estimation and, especially for traditional model classes, is even necessary when the number of predictors exceeds the number of observations ("$p > N$", James et al.

[13] To their credit, tree-based methods can approximate polynomial and other smooth relationships, but at the cost of increased complexity (many splits per tree or many trees in an ensemble), making them less sample-efficient for certain cases than more suitable model classes, including traditional statistical methods.

[14] Also, the reported superior performance by complex or ML methods has sometimes been found to be an artifact of flawed data splitting, leakage, and other violations of the good practices discussed above (e.g., Kapoor and Narayanan 2022 and Roberts et al. 2021).

2021, Chap. 6). However, traditional statistics is not without methods for feature selection (Hastie et al. 2009, Chaps. 3.2ff.): e.g., subset selection procedures or, more modern, via regularization such as the LASSO. Third, traditional statistics is geared towards what the ML culture calls structured or tabular data: e.g., for survey data represented in a matrix format, each row corresponds to exactly one respondent and each column corresponds to one survey question. It is undeniable that ML has made great progress regarding un- and semi-structured data such as (a collection of) images, audio or video data, texts, or even multimodal combinations thereof. In particular, Deep Learning is able to process unstructured data end-to-end: the raw input data are fed into the network – no feature engineering needed on the part of the data analyst (Molnar 2022, Chap. 11).

We conclude with two remarks. First, when the ML mindset fits an application, e.g., when prediction is the focus, then the procedural and methodological lessons discussed above are also relevant when traditional statistical model classes are used. Thus, much of the rest of this paper is not just relevant to the use of ML models. It is true, however, that more complex model classes have more potential for overfitting, i.e., over-adapting to their training data, so adhering to good practices tends to be more important (James et al. 2021, Chap. 2.1.2). Second, for a quantitative outcome, the expected squared out-of-sample prediction error (ESPE) at a point x_0 in the feature space can be decomposed (Hastie et al. 2009, Chap. 7.3): $E((Y - \hat{f}(x_0))^2|x_0) = \text{Var}(Y|x_0) + \text{Bias}(\hat{f}(x_0))^2 + \text{Var}(\hat{f}(x_0))$. The first term, $\text{Var}(Y|x_0) = E(Y - f(x_0)|x_0)^2$, is the conditional variance of the outcome around its true conditional mean $f(x_0)$: for given features, it cannot be reduced and is independent of choices made by the data analyst. The second term depicts the squared bias of $\hat{f}$, i.e., the expected squared deviation of the learned model from the (unknown) true conditional mean or true model $f(x_0)$. It is typically monotonously decreasing in model complexity, quickly at first and then leveling off (James et al. 2021, Chap. 2.2.2). The greater the (allowed) complexity, the greater the set of possible models, and thus the smaller bias, i.e., the distance of the best model in the consideration set to the true model (see Hastie et al. 2009, Chap. 7.3). The third term, $\text{Var}(\hat{f}(x_0)) = E(\hat{f}(x_0) - E(\hat{f}(x_0)))^2$, is the variance of $\hat{f}$, denoting the variation in the learned model when learned on different training data sets (same population, same sample size). In general, this estimation uncertainty is monotonously increasing in model complexity due to higher susceptibility to small perturbations in the data (James et al. 2021, Chap. 2.2.2). This is why, typically, the more flexible a model class, the more training observations are needed (James et al. 2021, Chap. 2.1.2). Note that the implied bias-variance trade-off is due to the focus of the supervised ML approach on prediction error: any model (class), ML or not, trained to minimize ESPE will exhibit at least a small amount of bias as long as it results in a larger decrease in variance. This results in a U-shaped relation of model complexity and ESPE (at least for the typical, 'under-parameterized' models, see Belkin et al. 2019). ML-based data analysis operates on both sides of this trade-off: flexible model classes and trying out different models mean more complexity (low bias, high variance) whereas feature selection (more bias, less variance) reduces model complexity.

2.3 Overarching Drivers and Goals of NSOs

The primary drivers are NSOs striving for improved products and processes (according to the quality dimensions, see Sect. 4), new products, new applications, new data, and an interest in new methods. These are not isolated aspects: e.g., some new products (e.g., data about very dynamic sectors or companies) make only sense when released very frequently or timely and some new data are too voluminous or unstructured for existing methods (based on traditional statistics or human work). We consider two of these drivers in more detail.

First, desired *improvements* include producing data and statistics more cheaply, releasing them more frequently, more timely, or on a more granular level, making them more accurate, or lowering response burden. Partial automation is seen as a vehicle for such improvements: e.g., algorithms can handle the easy cases, allowing staff to focus on cases that are hard to classify (Coronado and Juárez 2020) or important or influential (Dumpert 2020, p. 2), or to contribute to other activities (Coronado and Juárez 2020). Alternatively, algorithmic assistance can take on the form of providing a model's most likely outcomes for a given data point as suggestions in, e.g., human coding tasks (Measure 2020, p. 7 and Sthamer 2020b, Chap. 7).

Second, *new data* are considered to complement and, in part, to replace some of the traditional main data sources of NSOs – censuses, surveys, registers, and administrative data. We would like to remind that survey data have already been more than just the responses to the survey items: e.g., respondents and interviewers can provide samples (soil, saliva, blood, etc.) and measurements (Groves et al. 2009, Chap. 2.2.2), information from digital devices can be used (Keusch et al. 2024), and paradata about the data collection process are captured (Kreuter 2013; Schenk and Reuß 2024). How surveys will evolve in the era of, in particular, Big Data has received increasing attention since the second part of the 2010s (e.g., Baker 2017 and the BigSurv (https://www.bigsurv.org) conferences, see Hill et al. 2019): where they can replace survey data (Couper 2017, p. 134f.), where and how the two can complement one another (ibid; Japec et al. 2015, p. 873), and what can be learned methodologically from each other (Hill et al. 2021). The community appears to agree that surveys are here to stay: in contrast to most other data, surveys can be designed to give the desired breadth, level of detail, and fitness for a specific use, and to control the various error sources better. New data types may be collected in conjunction with a survey[15] or without it. Given user consent, wearables, apps, and sensors are emerging sources (Keusch et al. 2024), as are data donation and (screen) tracking (Ohme et al. 2024). Instead of single values, these data exhibit complex measurement series.

Another important new data source is images – so far mostly aerial images and other kinds of remote sensing (Coronado and Juárez 2020, p. 4 and 9): in particular, there have been vast improvements in the frequency and availability,

[15] This has two benefits: the survey and the other data can be designed to complement one another more optimally, and the already linked data collection makes tedious, error-prone record linkage (see footnote 16) unnecessary.

level of detail, and costs of satellite images. The volume is too much to handle for classical processes (i.e., involving traditional statistics or human work), making ML approaches a virtual necessity. Another reason is that in some cases multiple spectra or sensing technologies, going beyond wavelengths that humans can perceive, can be combined for the same object.

Textual data are a further new avenue (Text Classification Theme Group 2022). They range from open-text responses in surveys, to traffic, coroners', or police reports, to complaint filings, building permits, and other legal documents. Some of these are acquired via web scraping, as is information from company websites, online shops, news reports, job ads, and social media posts.

2.4 Applications and Tasks for ML in NSOs

Before and During Collection of (Traditional) Data The ability to acquire representative samples depends on having high-quality sampling frames. The necessary contact and other information can come from, e.g., registers or population-wide administrative data. Automated image recognition can help in keeping the addresses up-to-date (Coronado and Juárez 2020), as can information scraped from company websites. The latter are also helpful for making necessary additions to and deletions from the list (e.g., new and dissolved companies, respectively). Duplicates on the sampling frame are another problem and they can be detected and eliminated with the help of models for *identification* of units.[16]

In general, the empirics of coverage errors are understudied (Eckman 2013) and such new approaches are a welcome addition to the toolbox for improving sampling frames or, at least, to be able to evaluate them better.

Being uniquely suited to prediction tasks, supervised ML is the approach for forecasting (or nowcasting) what happens during data collection. Of particular interest are problems with the sampling units (likely nonrespondents, break-offs, and panel dropout) and their responses (e.g., problems understanding prompts or satisficing behavior producing subpar answers). Good predictions of these problems form the basis for interventions (e.g., via Adaptive Survey Design, see Wagner et al. 2008) that in turn help to increase the cost-effectiveness of the data collection and to prevent errors in the data.

Common to the mentioned tasks so far is that few features are available at the time of prediction, making paradata (Kreuter 2013) and other auxiliary information attractive. Schenk and Reuß (2024, Chap. 5) provide an introduction to paradata-based applications and interventions, but mention that ML is only starting to be embraced by survey methodologists. One type of paradata are observations from the interviewers (or address listers, recruiters, or others working on the ground) about the

[16] For two databases $\mathcal{A}$ and $\mathcal{B}$, identification has the goal to find the common units: e.g., for each record in database $\mathcal{A}$, it must be determined whether $\mathcal{B}$ has a corresponding entry and, if so, which one. In the statistical literature, this is mostly associated with record linkage (Herzog et al. 2007): e.g., for each survey respondent, one wants to identify the entry in administrative data belonging to the very same person, so as to merge the survey data and the administrative data sets. Depending on the scientific field, particularly within computer science, and the specific goal, this has many different names such as entity resolution and duplicate detection. For the latter, $\mathcal{A} = \mathcal{B}$.

particular dwelling and the neighborhood. Cartographic, satellite, or 'Street View' information is available online but has only been modestly explored with computer vision (instead of humans) for surveys. While these data sources are in principle available upfront, they may also be outdated or unavailable for most places. We suggest that pictures are easily captured with smartphones, by interviewers or address listers, and can be processed automatically in lieu of interviewers' judgment on what to record.

Finally, expert interviewers, especially in partly open or fully qualitative interviews, can also better prepare for visits with, e.g., web-scraped and condensed company information.

Processing and Adjusting Data *Editing* is the identification of data (cells, but also variables and units) that are problematic in one of two ways (Dumpert 2020, p. 1f.): Either information is missing, such as in voluntary survey responses (e.g., working hours and experience, income, or nationality and migration background; Beck et al. 2018a) or because multiple data sources were linked and a unit was not present in all of them. Or values are implausible, contradictory, or otherwise suspicious based on general logic, specific domain knowledge, or statistical patterns/distributions, such as survey responses suffering from satisficing or unverified parts of administrative data. *Imputation* is the filling in of missing or the alteration/replacing of suspicious values (Dumpert 2020, p. 1). Supervised learning on past edited data amounts to the search for the rules that govern the existing editing process: i.e., the outcome variable for such models is whether a particular value was flagged, edited, or imputed (or non-binary variants thereof). A trained model might then come somewhat close to replicating the performance of the editing process, but should not be expected to be more accurate (Dumpert 2020, p. 1). If instead true values (or some gold standard data that are better than the edited data) are available, a model trained on them may surpass the existing editing process. However, even for such data, there may be too few (documented) cases for each type of problem to be learned by supervised ML unless the mechanisms are very simple or the number of observations is enormous. Deviant interviewer behavior, up to complete fabrications, is one such example for which unsupervised learning may therefore be the better choice (Schwanhäuser et al. 2022): e.g., clustering and outlier detection. Finally, if the discovery of editing and imputation rules is a primary goal (Dumpert 2020, p. 2), we suggest that one might also turn to the field of rule induction or (association) rule learning (see Fürnkranz et al. 2012).

In NSOs' work, *outlier* or *anomaly detection*, i.e., the finding of unusual or extreme data points, is typically an unsupervised task: thus, among the many methods, e.g., clustering-based algorithms exist. Data analysts usually have four choices: to ignore outliers, to remove them altogether, to impute the suspicious values, or to use robust analysis methods. In contrast, data producers can sometimes investigate the flagged data points: they may be able to confirm or correct the information.[17]

[17] E.g., if a survey respondent is listed with extreme height, the interviewer can be asked if they recall such an occurrence. In company surveys, one may contact the respondent with a request for clarification or use web-scraped or other data sources that should contain the same information.

Identification of units is a crucial step for record linkage or for the identification of duplicates (see footnote 16). While there may be cases where supervised (machine) learning can be employed (Tokle and Bender 2020, Chap. 3.5.3), this is mostly an unsupervised task: in essence, for each pairing of records a and b from data sources $\mathcal{A}$ and $\mathcal{B}$, respectively, one wants to know their similarity in order to judge whether the belong to the same underlying unit. ML may help to implement different, data-driven distance metrics; also, as the computation of all pairwise comparisons is often infeasible, clustering or other methods may be used to replace the blocking of traditional identification units, which reduces the number of necessary operations as only units within blocks or clusters are compared.

Nonresponse is one of the sources that can bias data. This is often countered in data analysis by employing weights that are inversely proportional to the response propensity. Predicting these response propensities is a supervised learning task (see prediction during data collection in Sect. 2.4).

Some tasks can be seen as an example of both, processing and data analysis: e.g., the generation of new features. Clustering is an example of ML-based feature generation. Meanwhile, while textual data may often be fed to ML algorithms, the traditional processing steps (e.g., removing stop words, stemming, and turning text into a frequency matrix) themselves often do not involve ML.

Analysing Data ML, particularly Deep Learning, is very helpful with images. In NSOs, this has been mostly about satellite images to predict land cover and land use (agriculture, solar panels, etc.), for crop identification, monitoring of natural resources, growth of urban areas, and population distribution (Coronado and Juárez 2020). Such pattern recognition can be the basis for monitoring, e.g., wildlife populations (Bothmann et al. 2023) and indicators relating to climate change and the Sustainable Development Goals on agriculture, forests, and water (Holloway and Mengersen 2018).

Many classification efforts within NSOs have been on some kind of text (Reusens et al. 2022; Text Classification Theme Group 2022): e.g., occupation coding (from open-text survey responses or job listings to ISCO or other schemes), product categories (from household spending surveys, retail sales, scanner data, or web-scraped online shop information), and classifying enterprises according to their economic (NACE) or other activities (e.g., use of AI, engagement in research and development, innovativeness, corporate social responsibility, and social media presence) from web-scraped website information, financial publications, and news reports. Causes of death, accidents, crimes, etc., can be categorized from the respective text documents. The general classification from responses to open-ended items or transcripts in surveys is a frequent challenge. (Sthamer 2020b). Another task is the reclassification of past data when classification schemes are changed: e.g., historic or prior panel wave data need to be updated accordingly.

Economic and other time-series data exhibit potentially complex seasonal and other patterns which can be learned by flexible ML, given enough training data. Some of these data concern very dynamic settings (e.g., startups and high-growth firms), so prediction models can be used to extrapolate until the next data collection.

Occasionally, NSOs engage in forecasting (e.g., demographic developments) and nowcasting. GDP and other economic indicators may be released with much delay, and predicting these indicators with alternative data sources (e.g., Google Trends and traditional media information) is explored.

Outlook and Current Developments Based on a user's prompts, *generative AI* systems generate, e.g., new text or images.[18] Applications are not just limited to creative work, making such tools also interesting to NSOs. However, within the community, on this issue there exist mainly grey literature, plans, and prototypes (Curtin et al. 2023; Statistics Norway 2024). First, such systems may help to generate synthetic data both in the narrow sense (released data that do not violate confidentiality) and more broadly: e.g., chatbots powered by Large Language Models (LLMs) can be used to simulate respondents or interviewers (Argyle et al. 2023). While, during the pre-testing phase of a new survey, this may be a great additional tool to detect problems with a questionnaire, we are, however, more skeptical of the idea of using LLMs as a true replacement for (survey) data collection, i.e., prompting the LLM with respondent personas and asking for the persona's response to questions one would ask in a new survey. Particularly with regard to the quality expectations of NSOs' products,, the documented representativity issues can be stark (see, e.g., Heyde et al. 2024). These issues are subject of ongoing research (Agnew et al. 2024; Ma et al. 2024).

Second, LLM-based systems may function as a replacement of or supplement to human annotations (e.g., of open-ended surveys responses; Zenimoto et al. 2024). Third, such systems already can provide computer code for a desired task, suggesting time savings or improved quality of code and results. Such tools can make it easier for subject matter experts to interact with the data, e.g., during editing and imputation, without the need for assistance from other personnel. Time savings also imply that, e.g., more editing checks can be run. Such augmentation of human work will be particularly helpful in a data science world in which the roles of staff become blurred so that a single person has meaningful knowledge of not just one, but several or all of subject matter, statistics, ML, other methodology, and so on (Measure 2020, p. 6 and Julien 2020, p. 9).

So far, some data collection efforts come with great response burden: e.g., documenting all household spending or all food that is consumed is very time-consuming and error-prone. With the proliferation of smartphones, respondents can simply take pictures of their receipts or prepared food. Receipts, particularly when coupled with retail data, may even offer a greater level of detail on the specific products, their prices, and so on. Such solutions are actively studied in survey research (Struminskaya et al. 2021; Ilic et al. 2022) and official statistics (Benedikt et al. 2020) and are already being brought to market in the private sector (Page et al. 2023). Digital trace data, whether provided through a platform's API, tracking, or data donation, are another potentially interesting avenue (Keusch and Kreuter 2021; Ohme et al. 2024).

[18] Typically, these fall under the umbrella of very large, broadly trained*foundation models* (Bommasani et al. 2021) that may or may not require fine-tuning to tweak them towards one's specific application.

With the power of automation, ML, and the availability of additional data, ensuring confidentiality of released data and results released by NSOs is an increasingly difficult task. Yet, methods for attacking can also be used to improve defenses: e.g., generative adversarial networks to produce privacy-preserving synthetic data (see Neunhoeffer et al. 2021).

2.5 Compatibility

We close with a note on the compatibility of the combination of ML and traditional statistics and of data processing and eventual data analysis. We use the imputation of missing data during processing as an example. If the eventual, 'downstream' data analysis is traditional, it is well known that multiple imputation, rather than single imputation, is needed to properly quantify the uncertainty of parameter estimates (Little and Rubin 2019, Chap. 4f.). For that, multiple draws from the posterior predictive distribution are needed. A statistical imputation model provides such a distribution while ML models, even ensembles, typically do not. Conversely, suppose the eventual data analysis follows the supervised ML paradigm, i.e., prediction. If the downstream analyst is not interested in quantifying the uncertainty of the predictions, multiple imputation is not needed. However, a traditional statistical imputation model would be estimated on the whole data set while the downstream ML-user is only permitted to use the training data portion, but not the evaluation data, for learning how best to process the data (and for model training). In other words, this statistics-ML combination is likely to produce data leakage.[19] The downstream ML-user may also be affected by target leakage when the outcome variable was used in the imputation model (which a downstream statistics-user typically would not view as problematic). These are two examples in which using ML in data processing and traditional statistics in the eventual data analysis, or vice versa, can lead to problems.

3 Background: Fairness in Machine Learning

3.1 Algorithmic (Un)Fairness: Sources, Concepts, and Metrics

Following controversial applications of machine learning in high-stakes settings (Angwin et al. 2016; Buolamwini and Gebru 2018; Allhutter et al. 2020), fairness concerns have sparked a multi-faceted and multi-disciplinary research field centered around the social impacts of algorithmic decision-making (ADM). Research on fairness in machine learning (fair ML; see Mehrabi et al. 2021; Mitchell et al. 2021; Makhlouf et al. 2020; Caton and Haas 2024 for overviews) is thus typically focused on prediction models as part of larger socio-technical systems which may allocate access to positions, treatments or, more generally, valuable resources. The scope of fair ML, however, extends beyond ADM applications and includes fairness implications of the use of ML in other contexts, such as in data processing and survey production (Rodolfa et al. 2020).

[19] The ML-ML combination may be more likely to solve this problem by using the same data splitting.

A key concept in the fair ML literature is the notion of *protected attributes*. Protected attributes are inherent or ascribed characteristics of individuals (such as ethnic origin, gender, age, or religion – i.e., characteristics over which the the individual typically has no control), for which they can (or should) not be made responsible, but which nonetheless may be the grounds for differential treatment of individuals in the real world due to prejudice and discrimination. In a narrow sense, protected attributes may be defined based on anti-discrimination legislation (such as the Equal Credit Opportunity Act in the U.S., Mehrabi et al. 2021), but the eventual set of attributes that should be considered in a given application may be context-specific. Note that the adaptation of U.S.-centric concepts such as 'race' to other contexts is also non-trivial. The implication of introducing protected attributes is now *not* to ignore these features in the ML pipeline, but rather to faithfully acknowledge heterogeneity in data and to build subgroup-aware models that incorporate moral considerations on how to account for and resolve societal biases in a given context. To approach conceptions of fairness in machine learning, an initial, higher-level requirement could be based on the adaption of the disparate impact doctrine to data modeling – prevent outcomes or practices that have disproportionately adverse impacts on members of protected groups (Barocas and Selbst 2016). There are various pathways through which this principle may be violated in machine learning practice, the most prominent one being (different types of) *biases in data* (Mehrabi et al. 2021). Historical bias may be present in any data that result from social processes: administrative labor market records capture historical discrimination on the labor market, educational attainment histories are reflective of social biases in the education system, and (geospatial) records of criminal incidents are in part affected by decisions on which areas should be patrolled. Historical bias can easily be learned by and incorporated into ML models if data that reflect social processes is used for model training. Model training may, however, also be affected by measurement bias. In supervised learning, the outcome variable that is observed in the data may be a biased proxy for the actual outcome of interest such that social biases sneak into the model in the model specification step (Obermeyer et al. 2019): e.g., arrests are a biased proxy for criminal activity when, conditional on the same behavior, arrest probability is higher for some individuals or groups. Lastly, representation bias refers to deficits in the composition of the training data. Such deficits may refer to the (mis)representation of specific social subgroups in absolute or relative terms, or to the match between the data that is available for model training and the eventual target population more generally. We caution that very different meanings and haphazard usage of the term 'representativity' in the ML/AI community have been documented (Clemmensen and Kjærsgaard 2023), sometimes strongly diverging from the statistical notion. Regardless of these causes for biases in the data, there can be feedback loops: when (biased) predictions influence real-world outcomes, they may maintain or worsen biases in the next round of data, thereby sustaining or perpetuating biases in the predictions (Perdomo et al. 2020).

The fair ML literature notes that disparate impact may also be caused by other factors. This includes data pre-processing and modeling decisions along the ML pipeline which may operate next to or in interaction with existing data biases (Gerdon et al. 2022; Rodolfa et al. 2020). Examples include the compilation and match-

ing of information about subpopulations in the training data preparation step, the encoding of (correlates of) protected attributes and their use in model training, as well as decisions on how model outputs are eventually used downstream, e.g., for classification purposes.

In light of the various ways machine learning models may be affected by social biases, an abundance of *fairness notions* has been proposed in the literature which formalize different fairness conceptions and often imply corresponding *fairness metrics* to quantify adherence to a given notion in practice. Fairness notions typically focus on binary classification tasks and have been formulated on the group, subgroup, or individual level. Group fairness notions compare members of protected groups to their non-protected counterparts with respect to different prediction-based quantities. Given protected attribute A and predicted outcome $\hat{Y}$, independence-based group fairness notions require the predictions to be independent of group membership: $\hat{Y} \perp A$. The separation criterion additionally considers the observed outcome Y and requires independence conditionally on the true label: $\hat{Y} \perp A \mid Y$. Sufficiency-based notions, in contrast, condition on the predictions: $Y \perp A \mid \hat{Y}$ (Barocas et al. 2023; Makhlouf et al. 2020). Next to group fairness, subgroup fairness aims to provide stronger fairness guarantees by imposing fairness constraints on large collections of subgroups that may be defined by intersections of many (protected and non-protected) attributes (Hebert-Johnson et al. 2018; Kim et al. 2019; Kearns et al. 2018). Finally, individual fairness formulates requirements on the individual level, e.g., by mapping distances between individuals to distances in predictions (i.e., similar individuals should receive similar predictions; Dwork et al. 2012) or by drawing on causal reasoning (Kusner et al. 2018; Kilbertus et al. 2017).

While group-based fairness metrics are rather straightforward to compute and evaluate in practice, a central result of the fair ML literature has been that complying with multiple group-based fairness notions simultaneously along the dimensions discussed above is difficult. Except for highly stylized cases, a prediction model cannot fulfill independence, separation, and sufficiency criteria at the same time (Chouldechova 2016). Requesting group fairness thus comes with trade-offs, and considerations on which (group) fairness notion should be prioritized might be highly context-specific.

Valid criticisms of the fair ML literature must be acknowledged: e.g., some fairness notions may suggest changing a prediction model so as to provide worse predictions (for some groups) in order for some equality constraint to become satisfied. However, as discussed by Kuppler et al. (2022), this is an artifact of considering only ADM systems in which the decision is a function of solely the model's prediction $\hat{Y}$, ignoring the protected attribute A, other features used to predict the outcome W, and further information such as the accuracy of $\hat{Y}$ given A and W. Such systems exist, but the flaw is in their construction, not in the consideration of the fairness of algorithms per se. This is solved by splitting up the prediction task (i.e., building a good model) and the decision-making task: then, too, fairness notions for the former and justice notions for the latter can be cleanly separated. As suggested by Kuppler et al. (2022), *accuracy-based* or, conversely, *error-based fairness* metrics may be the most natural: For classification tasks, one can ask how rates of overall errors,

false positives, false negatives, 1-precision, or 1-recall, as well as miscalibration[20] differ across groups or subgroups. For regression tasks, bias and variance can be looked at. This is all the more important for the work of NSOs: they do not engage in ADM, they cannot know during data production what justice principles (and other goals) downstream users of the data will have, and accuracy is already one of the main quality dimensions they consider (see Sect. 4.6). Furthermore, error fairness translates more easily to data that are not about humans but about, e.g., companies – a large part of NSOs' work. We will therefore concentrate on error-based fairness notions later on.

3.2 The Human Component(s)

The catalog of fairness notions that have been proposed in the literature highlights that fairness can be conceptualized in various (and conflicting) ways. Given a fairness metric, additional parameters might need to be set to formalize the range of values deemed acceptable. Thus, technical measures which quantify whether a prediction model satisfies some fairness constraint do not substitute for human judgment and reflection. In contrast, fair ML implies moral reasoning and raises questions of distributive justice (Kuppler et al. 2022; Heidari et al. 2019; Loi et al. 2021; Binns 2018; Lee et al. 2020; Gajane and Pechenizkiy 2018): How should (different types of) prediction errors be distributed across social groups in a given context? Given fair predictions, which downstream allocation of resources do we perceive as just?

Committing to fairness in building and implementing machine learning systems thus requires developers and stakeholders to explicitly specify their goals. This inevitably includes engaging with various normative questions such as which attributes should be considered sensitive, which fairness concept should be prioritized, and how exactly deviations from 'optimal fairness' should be defined and potentially addressed (Bothmann et al. 2022). Some guidelines have been proposed to help navigate the fairness field: Makhlouf et al. (2020) and Saleiro et al. (2019), for example, structure fairness notions based on a set of selection criteria. Such templates can point out critical decision points and help in guiding discussions among stakeholders, but nonetheless require normative input and context-specific weightings of interests. This implies that NSOs may need to critically engage with downstream users, and reflect on whether the same product can meet heterogeneous needs in different contexts.

Recent research has started to focus on the human component in fair ML by studying human perceptions of algorithmic fairness. This line of work focuses on how design aspects of ADM systems or characteristics of the human evaluators affect individual fairness perceptions, or how algorithmic decisions are perceived in comparison to human decision-making (see the review by Starke et al. 2022). Studies that investigate which type of input data (Kern et al. 2022) or attributes (Grgic-Hlaca et al. 2018) are perceived as sensitive in a given context or which types of

[20] Calibration requires that the predicted probabilities $p(x)$ of a ML model 'mean what they say', i.e., correspond to the actual risk of observing the event that is predicted. That is, for any probability v, $E(y \mid x, p(x) \approx v) \approx v$.

prediction errors are evaluated as particularly problematic (Srivastava et al. 2019) by the general public may provide valuable input to tackle the normative dilemmas mentioned above.

Finally, characteristics of the individual decision-maker, the algorithm, and the context in which it is applied can affect "algorithm aversion" or "algorithm appreciation", i.e., the individual's under- or over-reliance on the algorithm's results (Burton et al. 2020; Jussupow et al. 2020; Hou and Jung 2021). While NSOs are typically not the place for ADM, the data they produce may very well be frequently employed for such purposes, e.g., by governmental bodies. Thus, the data and how they are produced as well as what information (documentation, metadata, etc.) is released can influence aversion to such downstream algorithms. The same can be said for the fairness perceptions discussed in the previous paragraph. In addition, the internal high-level decisions of whether and how to implement ML algorithms in a NSO's processes are likely affected by the very same characteristics, as are attitudes by other stakeholders (staff, recipients of statistics, data users, etc.).

4 Data Quality Framework Principles

A commitment to quality is one of the fundamental principles of NSOs (see Sect. 1). Specific, lower-level criteria are required in order to concretize and operationalize this overarching goal. The European Statistics Code of Practice (Eurostat 2017) in particular contains such principles for the institutional level, for the statistical processes, and for the outputs: relevance, accuracy, reliability, consistency and comparability (internally, over time, within and across regions), accessibility and clarity (clear, understandable, and documented), confidentiality, response burden (proportional and non-excessive), timeliness, and cost-effectiveness as 'quality dimensions'. Yung et al. (2022, p. 1), aiming to complement rather than replace existing quality frameworks, put forth a "Quality Framework for Statistical Algorithms" (QF4SA henceforth) consisting of five dimensions: accuracy, timeliness, cost-effectiveness, explainability, and reproducibility. The first three are visibly also part of the above list. Explainability is related to accessibility and clarity, but not fully contained within it. We agree with Salwiczek and Rohde (2022) that robustness is another aspect of reliability (in the sense of the European Statistics Code of Practice mentioned above, but not exactly in the strict statistical sense) and add it to the dimensions that we discuss. Yung et al. (2022, p. 1 and 4) chose the five dimensions in QF4SA because they find them particularly relevant when "intermediate outputs" (that are inputs for further processing or data analysis) are produced; these dimensions, however, should also be considered upstream (relating to data and data collection) and downstream (for final statistical outputs, whether created by NSOs or external data users). We will briefly touch on to what extent these dimensions are also more relevant or different in a world with ML and therefore should be singled out. While explainability and reproducibility connect to fundamental principles, in the presented form they are sufficiently distinct from them so that they can be considered missing from previous, pre-ML quality frameworks.

The importance of the above-mentioned quality dimensions has been established: they are central to credible, high-quality products and institutions. There is also a ML perspective on quality dimensions. Doshi-Velez and Kim (2017, Chap. 2) make the case that many problems stem from some form of incompleteness: models are optimized for predictive accuracy – one important goal –, but the deployer's or decision-maker's other desiderata typically do not enter model building and training at all (as would be possible, albeit non-trivial, by introducing formal constraints or multi-objective optimization). The trained models' performance on these criteria is thus completely unknown and must be explicitly evaluated. Differences in how well models do on these quality dimensions can then be used to choose among the (similarly predictive) trained models. In any case, it must be ascertained whether the selected model fulfills minimum standards. In addition, providing explanations, limitations, and suitable applications when releasing or deploying a model is encouraged (Mitchell et al. 2019; Richards et al. 2020).

For the remainder of this section, we consider these dimensions in a ML world and mention their interconnections. The respective interactions of these dimensions with fairness are addressed in Sect. 5. As we build on QF4SA and our remarks are complementary and typically higher-level, this chapter is best read in conjunction with Yung et al. (2022).

4.1 Explainability and Interpretability

We begin with explainability and interpretability which we address in somewhat more detail than the other quality dimensions.[21] As is widespread, we treat interpretability and explainability as synonyms (e.g., Miller 2017, Chap. 2.1.5 and Molnar 2020, Chap. 3.0). Interpretability has a dual role: it is a desirable property of a model and denotes a set of tools that can help to investigate other desirable properties. As this dimension is not explicitly part of pre-ML data quality frameworks (Yung et al. 2022, p. 4), we give an introduction to the field of Interpretable Machine Learning (IML)[22] – on a high level, without delving into specific methods, and to the extent useful for our later discussion. We refer the interested reader to Molnar (2020)'s excellent and accessible book on the subject. While the field has created a broad set of concepts and methods, it is still developing. We will also touch upon the fact that IML is not a monolith and the various concepts and methods are sometimes competing (Lipton 2018). Thus, when IML methodology is put into action, practitioners need to be aware of the (general or situation-specific) limitations of these methods and how to use them properly (Molnar et al. 2022; König 2023).

[21] We are not aware of any higher-level introductions to this topic in the respective literature on official statistics, survey methodology, and so on. Our overview is a complement to that of Yung et al. (2022) who are more focused on concrete explainability methods.

[22] As ML is a subset of AI, IML should be a subset of Explainable Artificial Intelligence (XAI). Similar to how today much of AI is actually ML, IML is in practice not necessarily distinguished from XAI.

Concept and Background Interpretability can be broadly seen as the degree to which a human can understand how or why an algorithm produces its output (Molnar 2020, Chap. 3.0).[23] Often, this is achieved by demonstrating how inputs and outputs are related in the trained model – globally, locally (i.e., for a specific data point), or somewhere in between (Molnar 2020, Chaps. 3.0, 3.5 and Yung et al. 2022, p. 5f.). While many might agree with this abstract, vague conception, there is no single, universally-accepted definition of interpretability: in particular, a precise or mathematical definition of interpretability, how to measure it, and sharp boundaries are all not obvious (Murdoch et al. 2019, p. 22071; Molnar 2020, Chaps. 3.0, 3.4; Lipton 2018).

We would like to re-emphasize that what IML methods primarily do is explain a trained model. Only secondarily they also allow one to get a glimpse of (relationships and structures in) the training data and, to an even much lesser extent, of the DGP and of the 'true nature of the world' – however, all only through the narrow, often distorting lens of the trained model. Also, IML methods do not change a ML model: they are applied post hoc to facilitate human understanding of a trained model's behavior, but they themselves do not alter the statistical algorithm or its results in any way.[24]

Some model classes have a structure that is both, simple enough and well-understood, so that they are considered *intrinsically interpretable* (Molnar 2020, Chap. 3.2): e.g., the learned beta coefficients (in ML parlance: weights) of a sparse linear regression show directly how a feature's values relate to the model's predictions. Such models come with their own built-in interpretability 'devices' (such as said weights), in contrast to models from the other end of the spectrum: because those exhibit high complexity and low transparency, they are considered a black box and illumination by IML methods is necessary for understanding their behavior.

Model-agnostic IML methods work for any model class (Molnar 2020, Chap. 3.2). Being able to investigate interpretability with the same IML method is key when several trained models are compared, especially when from different classes. *Model-specific* IML methods can only be applied to a small set of model classes, typically because they rely on model internals that only exist for a few model classes.[25]

Scope: interpretability levels

1. *Algorithm transparency* or *mechanical understanding of the algorithm* (Molnar 2020, Chap. 3.3.1 and Yung et al. 2022, p. 6) is about the general, abstract knowl-

[23] We deliberately use the generic term *output* in this definition as the common focus on *predictions* is too centered on supervised ML only, although the latter is certainly the main focus.

[24] Of course, someone who trains an ML model might take the insights gleaned from IML methods and decide to make adaptations to the ML model. However, the IML methods themselves do not directly produce any changes.

[25] The abovementioned inherently interpretable model classes rely on built-in IML 'devices' that are model-specific: e.g., a linear regression's weights.

ships it can learn" (Molnar 2020, Chap. 3.3.1).[26] While such general knowledge can aid with the next two points Yung et al. (2022, p. 5), it is completely decoupled from the specific data and the actually trained model. Consequently, it is typically not considered directly part of IML.

2. *Global, model-level*, or *dataset-level interpretability* (Molnar 2020, Chaps. 3.3.2, 3.3.3 and Murdoch et al. 2019, p. 22076) considers how, in the trained model, inputs are related to outputs (Yung et al. 2022, p. 6). First, on a high level, typical questions include which features were selected, which are the most important ones (by quantifying their respective contributions), and which interactions are incorporated. As holding an understanding of the entire model in one's mind or visualizing it is typically beyond human capabilities, a second, modular approach is crucial (Molnar 2020, Chaps. 3.3.2, 3.3.3): on the feature level, the relationship of a particular feature to the output is elucidated: e.g., positive/negative/zero/non-monotone, linear/U-shaped/cutoffs/etc., moderation by interactions, and so on.
3. *Local, individual-level*, or *prediction-level interpretability* (Molnar 2020, Chap. 3.3.4 and Murdoch et al. 2019, p. 22076) gives *explanations* of how the prediction for a particular instance (statistical unit) comes to be (Molnar 2020, Chap. 3.3.5). Often, this again involves investigating how the features' values relate to the output – but more locally than in model-level interpretations. For instance, in binary classification: Why was the prediction '1' and not '0'? How does the output change when the value of one particular feature is altered but the instance's other feature values are kept constant? In order to receive a desired output: which feature values would need to be changed and how (typically: what is the closest (artificial) data point yielding the desired output)?
   Alternatively, but less frequently, the instance is contrasted with another similar, typical, or otherwise relevant data point or group of data points, whether artificial or actual.[27]

The separation between global, model-level and local, prediction-level interpretability is useful because typically they use different IML methods and they have different goals and target audiences (Murdoch et al. 2019, p. 22076). However, the boundary is not absolute (Molnar 2020, Chap. 3.3.5): First, individual-level explanations can be aggregated to the level of specific groups, enabling across-group comparisons. Second, individual-level explanations can even be aggregated to the feature level. Third, the global methods can be applied to groups of instances (user-specified or formed by the model). This is important for IML methods as a tool for fairness evaluations. While fairness notions can be on the individual level, they often concern groups.

[26] E.g., linear regression fits a line through a cloud of data points so as to minimize the average squared distance from the line to the data points. Some methods such as Deep Learning are not only more complex than traditional statistical techniques but also markedly less well studied (Molnar 2020, Chap. 3.3.1); to overcome their lower inherent transparency, IML is needed even more.

[27] The resulting comparison, however, then typically turns again its focus on the (difference in) feature values.

Outputs, Products, and Tools of IML Methods The types of output produced by the various IML methods are rather heterogeneous. Molnar (2020, Chap. 3.2) organizes them into five partially overlapping groups.

1. Feature summary statistic: e.g., feature importance; pairwise feature interaction strengths; learned beta coefficients in linear models (which are both summary statistics and model internals).
2. Feature summary visualization: e.g., partial dependence plots.
3. Model internals: e.g., the features and thresholds used for the splits in tree-based models; learned beta coefficients in linear models.
4. Data points: e.g., counterfactual data point (similar data point to a specific instance, but with the desired output; see Verma et al. 2022); adversarial example (slightly different X so that $\hat{Y}$ now is wrong); influential instance; prototype.
5. Approximation by a surrogate model from an intrinsically interpretable model class.

Considerations for Official Statistics First, as emphasized above, IML methods provide insights into the trained model. It is tempting to combine results from IML methods with one's own domain knowledge or intuition and believe one has uncovered some insight into the underlying DGP or the true nature of the world. However, one cannot know whether such statements are about the model or about reality. In addition, IML methods typically use only simplifications or approximations of the trained model, and different IML methods, employed to answer (seemingly or actually) the same question, sometimes provide conflicting results (Krishna et al. 2022). Thus, NSOs need to be careful with respect to the nature and stability of conclusions that can be drawn from IML.

Second, many model classes used in the ML paradigm are considered black boxes. IML methods increase the transparency of such systems, increasing credibility and trust directly (Yung et al. 2022, p. 7). As IML is also employed by model developers to improve a model and by auditors to investigate it (Molnar 2020, Chap. 3.1), IML usage may also increase trust in NSO's systems indirectly. Conversely, outside, pretrained models may be harder to probe and understand, let alone fix discovered accuracy, robustness, or fairness problems.

Third, interpretability is a human and social endeavor. Characteristics of the explainer, the recipient, the (social) context, and how explanations are communicated matter and should be considered against the backdrop of human cognitive biases (Miller 2017; Molnar 2020, Chap. 3.6). In particular, stakeholders in different roles (e.g., model developer, model-assisted NSO staff, data user, subject of ADM, or regulator) or with different levels of subject matter or ML expertise may find different IML methods useful (Lakkaraju et al. 2022; Varshney et al. 2022, Chap. 12; Yung et al. 2022, p. 7).

Fourth, interpretability is not equally important for all systems (Molnar 2020, Chap. 3.1): well-understood, well-researched systems or low-stakes settings are different from high-stakes applications (e.g., ADM) or when a system is in widespread use. For foundation models, effects of algorithmic monoculture and homogenization at scale (Bommasani et al. 2021, Chap. 5.6; Kleinberg and Raghavan 2021; Creel

and Hellman 2022) have received attention; within or across NSOs, some systems will also be more important than others.

Finally, interpretability can be of great importance in the context of specific tasks of NSOs: e.g., when the goal is to find and formalize the rules that expert annotators use to identify problematic data (Dumpert 2020, p. 5). As such editing is typically accompanied by imputation, comprehensible rules might also aid in suggesting the replacement values. This application highlights the importance of choosing the set of considered ML methods: inherently interpretable decision trees, especially when combined with appropriate feature engineering, are likely to yield such editing rules, as is the field of rule induction or rule learning (Fürnkranz et al. 2012).

We use IML to illustrate a developing chasm between two types of (ML) data analysis: that of structured data (often with tree-based and traditional model classes) and that of unstructured data (typically with Deep Learning). For the former, we would consider, e.g., which features are important or, in counterfactual examples, which values someone would need to change to get a desired prediction. For the latter, features are of much less consequence, but, for images, we might highlight pixels or regions that the model relies on much or not at all and visualize them akin to heatmaps ('saliency maps'). Some 'Clever Hans effects' – i.e., the model exploits spurious or unreliable signals (Bellamy et al. 2022) – have been discovered that way (see also Sect. 4.4).

4.2 Cost-effectiveness

Cost-effectiveness is about the relationship between the (quality of the) outputs and the incurred costs Yung et al. (2022, p. 5). Costs may be (quasi-)fixed or ongoing. Important categories include: the necessary equipment, data, and skills must be acquired; the data must be processed, a model must be trained and evaluated; equipment, skills, and models must be monitored, maintained, and updated.[28] We refer to Yung et al. (2022, Chap. 6) for more details, but want to highlight some aspects.

Standardization of processes is one important tool to manage cost-effectiveness and other quality dimensions (e.g., Eurostat 2017, Indicator 10.4 and Destatis 2021). Automation, driven by ML (or statistical) models, is a promising avenue in this regard. One anticipated benefit is that automated processes may entail higher (quasi-)fixed costs of setting up – e.g., for equipment, knowledge acquisition, and the training, selecting, and evaluating of models – but once they are implemented, the marginal cost per additional unit – e.g., the cost to generate a prediction for an additional data point – is very low: i.e., such processes are highly scalable.

We consider two cost aspects in more detail. First, the work does not stop with training a model: it must be evaluated on more than its performance – namely its interpretability and its fairness – and it must be continuously monitored after

[28] The CO2 cost of training models, cloud storage, and so on, may not have been at the forefront so far, but will only increase in importance. Organizations, particularly those still building and changing their capacities, might be interested in 'Green AI' (Schwartz et al. 2019; Tornede et al. 2022; Ligozat et al. 2022).

deployment for model drift or decay (declining performance and other changes in behavior) and, when needed, re-trained; this binds manpower, computational resources, and may necessitate further data and data processing work (Choi et al. 2022, Chaps. 1 and 4). Note that these requirements are largely independent of whether the chosen model for automation is ML or statistical: changes in real-world mechanisms or in data collection affect them both. Choosing models that are more stable (see Sect. 5.4) may thus provide financial relief via a decreased need for re-training. Second, data are not only at the core of model performance but also an important consider cost consideration. ML models and the ML paradigm typically exhibit high demands regarding computational resources and data volume,[29] affecting costs, timeliness, and the uncertainty (Sect. 4.6) of the resulting output. The flexibility of ML is one of its advantages, but also increases data requirements: the less 'known' structure and other types of relevant expertise are used, e.g., to create and transform features, the more the method must learn on its own – Deep Learning on unstructured data is the prime example. Also, the training data for supervised editing and imputation models need to be carefully labeled, requiring perhaps more time than the simple editing and imputation itself would, and, if anything changes, existing training data may need to be re-relabeled (Sthamer 2020a, p. 6).

Published cost studies are rare in general and findings for one setting or organization may not translate directly to another (see Groves et al. 2009, Chap. 5.3.6 on survey costs). ML and automated solutions have not shown to be always more cost-effective for NSOs' applications, but there are positive examples (Sthamer 2020b, p. 13). In addition, switching from one process to another is not cost-free. So far, in applications such as editing and imputation, no single ML method clearly dominates and a lot of work may be required, especially to yield more than marginal benefits, and a full range of model classes must be prepared and considered for each application (Dumpert 2020, p. 7).

Unsurprisingly, automation is also being pursued in the ML world. First, automated machine learning (AutoML; see Tornede et al. 2022; Weerts et al. 2023) is concerned with automating the whole pipeline, from data pre-processing and feature engineering to model training, hyperparameter optimization, and model selection. Second, monitoring of a deployed model for drift (Choi et al. 2022) can also be automated. Yet, complete automation is not the goal of NSOs, and expertise and skills are still needed to implement and monitor these even more automated systems. This is also evident in Deep Learning: being able to process (raw) data end-to-end, it may not require (costly) feature engineering, but choosing the proper types of neural networks, optimizing its building blocks, and choosing the best (hyper)parameters does require expertise and some work (e.g., Coronado and Juárez 2020, p. 7 and James et al. 2021, Chap. 10).

[29] There is, however, a notion of 'tinyML' – which can have the additional benefits of being able to run on, e.g., respondents' smartphones so that confidential information may be processed on the device and never leave it.

4.3 Timeliness

Timeliness can be broadly seen as "the time between a need [...] and the release of the information to meet that need"; particularly for information covering a certain point or period of time, this is often conceptualized as the time between that reference point or period and when the information is made available (Yung et al. 2022, Chap. 5).[30] Economic indicators such as GDP and inflation are examples of time-sensitive information: the more delayed they are released, the less relevant and valuable they are to decision-makers. For the work of NSOs, once should consider time for data collection and acquisition, for data processing, and for data analysis. NSOs have processes for these three tasks and Yung et al. (2022, Chap. 5.2) differentiate between the time needed for the development of a process, i.e., from conceptualization to implementation, and the time needed for the process to run. These processes can be sequential so that one cannot begin before the previous one is finished: bottlenecks, such as editing and imputation, may thus particularly benefit from improved, model-based processes (Dumpert 2020, p. 5).

A different aspect of timeliness is the ability of a model to be used in (near) real-time, particularly to assist with data collection. In surveys, models may be employed to predict the likelihood of break-offs or of poor answering behavior and intervene accordingly (e.g., Mittereder 2019, Chap. 6). They may also be used to evaluate data accuracy[31] as interviewers or respondents on the spot should be more able to correct errors than data processing staff can do later on. Sophisticated models that would need constant re-training during ongoing data collection might be too slow to be implemented.

4.4 Robustness

First, in the ML community, *adversarial robustness* is about the ability to withstand attacks (Varshney et al. 2022, Chap. 11): Adversaries may either target the modeling phase, poisoning the training data by injecting additional data or by modifying data in order to change the trained model's behavior (e.g., generally lower accuracy or a different prediction for specific, targeted points in the feature space). Or they may target the deployment phase: e.g., to be able to evade the model's 'intentions', "gaming the system" (Molnar 2020, Chap. 3.1), or they may try to extract the trained model's parameters or use the model's outputs to reverse-engineer valuable information about the training data (perhaps even the whole data set). While 'attacks' might sound overt, documented examples include how slight, imperceptible-to-humans changes to some of an image's pixels can change classification drastically. When NSOs collect their own data, there are typically very few, if any units that have the capacity to modify the training data meaningfully (very large companies

[30] Timeliness is also about the punctuality of outputs. We will not refer to this explicitly other than by mentioning that when the need to re-train a drifted model arises, this might cause delays, particularly when the issue is discovered late and there is no buffer.

[31] Such evaluations may use ML and they may use multiple data sources, including new data sources mentioned above to check, e.g., survey responses or information provided by interviewers.

might be a counterexample). This is different when there is reliance on outside data and data providers. While not used by NSOs themselves for decision-making, governmental institutions and others may base their decisions on data or results provided by NSOs, making the topic of evasion not completely irrelevant.

Second, robustness in the ML community can also be in regard to *data shifts* and *model drift*, i.e., to changes between training and deployment (Quiñonero-Candela et al. 2008; Moreno-Torres et al. 2012, Chap. 7). This an aspect of *transportability*, i.e., the question of whether a model trained on one data set also holds, without bias, for a different set of circumstances: In the supervised ML paradigm, the main concern is whether the target population (deployment data) and the training population (training data) exhibit the same distribution or whether there is a shift. Traditional inferential statistics is mostly worried about whether estimates from the data generalize to a 'general population', i.e., whether results possess external validity.[32] Transportability is a more general concept than these two concerns: any change in the circumstances, environment, or context may pose a threat to the validity of a model outside its training data. Among these, the Total Survey Error (TSE) framework (see Groves et al. 2009 and Sect. 4.6) highlights changes to the data collection protocols and data processing procedures.

Third, traditional statisticians might be inclined to think of *robust statistics*: the "insensitivity to small deviations from the assumptions" of models in actual, finite data (Huber and Ronchetti 2009, Chap. 1.1) – in particular, the robustness of the results to the presence of outliers and otherwise extreme data points (possibly the result of gross errors) in the data (see also Hampel et al. 1986, Chap. 1.1). Ensembles are a ML answer to a lack of robustness of individual learners: this is the advantage of, e.g., random forests over a single tree (James et al. 2021, p. 340). Ensembles can be specified according to fixed rules such as averaging or majority vote. It is also possible to learn how best to weight an ensemble's components. Clustering is another application that is often prone to instability. It can also demonstrate another way to employ ensemble thinking: different cluster models can be compared regarding their conformity and a robust combined model can be created that only contains results that are common to many or even all of the models (Hornik 2005).

Robustness does not only concern the model itself but also the assessment of its accuracy (Sect. 4.6) and model interpretations (Molnar 2020, Chap. 3.5; Dutta et al. 2022). Conversely, IML methods can uncover which features are important in a model. These may then be more closely monitored for signs of drift. Particularly for the adversarial and drift notions of robustness, features that exhibit a causal effect on the outcome are often preferred over mere correlative features (e.g., Molnar 2020, Chap. 3.1): causal relationships may be much more stable over time and much harder to manipulate or game. A model's most important features can be found with IML and investigated in this regard. Similarly, if the most important features are just proxies, we might improve future data collections to come closer to the actual variables of interest, increasing statistical robustness by reducing measurement error: e.g., survey questions may be tweaked or alternative data sources considered.

[32] An exception: the question of whether measurement error models estimated on one gold-standard data set can also be transported to another data set (Carroll et al. 2006, Chap. 2.2.4).

Furthermore, markedly worse than spurious features are shortcut learning and Clever Hans effects: i.e., when the training data contain signals about Y that in deployment will not be present or exhibit a very different relationship (Bellamy et al. 2022). Examples in medical image analysis for disease prediction include the type of imaging device, image timing, watermarks, or, worst, circles or arrows pointing to tumors that were of course not present on the disease-free among the training images (e.g., Chen et al. 2019, p. 104f.).

4.5 Reproducibility

Within and across scientific disciplines, a number of contrasting definitions and terminologies exist in this area (Barba 2018; Plesser et al. 2018). They all describe under which conditions the results from a new data analysis must be the same or qualitatively similar to those from a first data analysis (Goodman et al. 2016): *Methods reproducibility* is defined as obtaining identical results, using the *same data* and the *same methods*. By leaving the era of 'point-and-click adventures' and instead archiving code and data and making them accessible, scientific communities approach what is increasingly seen as the bare minimum for credible empirical work. The importance of metadata and proper documentation has also been emphasized (Choi et al. 2022, Chap. 5). *Results reproducibility* is about finding the same results using *different data* but the *same methods*. We consider *inferential reproducibility* as whether (qualitatively) the same results are produced with the *same data* but *different methods*.[33]

Reproducibility is a big part of Open Science and "Open Science is just good science in a digital age" (Seibold et al. 2023), regardless of the type of data analysis.[34]

[33] In these definitions, 'methods' are to be understood broadly, with analytical choices big (e.g., the considered model classes) and small (e.g., options in an algorithm implementing a model class), but also how to proceed with, e.g., outliers. Also, implied in these definitions is that all data analyses concern the same research question or research object (Goodman et al. 2016; Salwiczek and Rohde 2022).

[34] Here we sketch roughly how reproducibility relates to two classical statistical concepts, *reliability* and *validity*. All three can be described with regard to repetitions in some sense. In the popular archery analogy of repeatedly shooting an arrow at a target disk, *reliability* refers to precision (i.e., how much do the impacts vary around their average impact spot) whereas *validity* refers to accuracy (i.e., how much the just-mentioned average is away from the bull's eye that the archer intends to hit). Unfortunately, neither is there a one-to-one relationship between these three concepts nor can one be characterized as a necessary condition (i.e., superset) or sufficient condition (i.e., subset) of the other. The literature jointly considering all three concepts is also scarce and focused on ML research rather than ML practice (Myrtveit et al. 2005; Raji et al. 2021; Herrmann et al. 2024). The discussion is further complicated by two factors. First, what one exactly means by 'different data' in the definitions of reproducibility above: e.g., a new sample from the same distribution or a sample from a different, perhaps more general population (in the latter case, this is more related to generalizability than to reliability). Second, the above-mentioned reproducibility concepts vary on two dimensions, and even for the first concept, methods reproducibility, there exist subtypes (such as computational reproducibility, meaning running the same code with the same data should yield identical results each time). *Reproducibility* can be seen from a repeatability perspective and the different types of reproducibility then refer to the different conditions under which one investigates repeatability. However, e.g., the aforementioned computational reproducibility would typically be checked *once*, whereas to calculate a statistical validity measure, *many* repetitions would be useful; in addition, one would typically expect *identical* results in computational reproducibility, but one would not think of validity as requiring *zero* variance. Validity, i.e., how close to the 'truth' results or conclusions are, is even harder to

Thus, this important quality dimension is not tied directly to ML. It is, however, true that more sophisticated model classes are more likely to contain stochastic elements; also, data splitting in the supervised ML approach introduces randomness. If uncontrolled, these random elements are a threat to *methods reproducibility*. Yet, this is not restricted to ML: e.g., traditional Gaussian clustering methods may use random initialization values. Versioning, referencing, and archiving of code and data are relatively straightforward for a traditional data analysis world where everything is in-house. However, new data sources may be non-static (e.g., even *past* social media platform data are frequently changed retroactively, see West et al. 2023) or too big. A similar argument pertains to large, pre-trained models (e.g., foundation models) trained and provided by an outside organization and employed by a NSO possibly after some fine-tuning. Any process containing human decision-making is more difficult to archive and to reproduce than an automated one, although very strict, documented guidelines may help.

While we agree with Yung et al. (2022, p. 20) that NSOs often cannot easily collect new data – especially if they are to be collected in precisely the same manner – we, however, do not believe that this makes *results reproducibility* generally unachievable or irrelevant to NSOs. In fact, whether new data sources permit the same results (but more cheaply or timely, see Sect. 2.3) is directly coupled with questions of results reproducibility. Also, recall that the ML paradigm is typically not about just analyzing one data set to answer questions: rather, the rationale is the deployment of the trained model on new data – often more than once or even continuously. How well the model holds up is about results reproducibility and the reason for monitoring for model drift.

Finally, we note that 'results' in the above-mentioned reproducibility definitions are understood to refer to outputs of data analyses. NSOs as producers of data that are used, often in multiple ways, by end users inside and outside the organization, may also consider the reproducibility of data processing (or data production more generally). This is also true for additional information they release: e.g., IML methods may contain stochastic elements.

4.6 Accuracy

The many conceptions and measures of accuracy share a common notion (Yung et al. 2022, Chap. 3): accuracy is about the closeness of 'what one has' to the truth or, when 'truth' is not an adequate concept, to what one intended.[35] Inspired by NSOs' dual role as producers of data and of statistical outputs, we think of three kinds of objects of interest 'which one has': the data, estimates of some population parameters (in traditional statistics), or predictions (in the supervised ML paradigm). In particular for the data and the predictions, one can take an individual view (a specific data point or a local prediction $\hat{y}|x_0$) or an aggregate view (differences in the distribution

measure, except in the rare cases in which the truth is known. In essence, reproducibility measures that vary the date are closer in spirit to reliability, whereas those measures that vary the methods are more related to truth-seeking of validity.

[35] For instance, when subjective opinions of a survey respondent are sought, 'truth' might not be the best concept.

of the data relative to that of the target population or quantifying a model's overall performance with one accuracy number).

Statisticians tend to think of the deviations from the truth/the intention as either systematic (bias) or random (variance). Survey methodologists often employ the Total Survey Error (TSE) framework to conceptualize the different sources for such errors (Groves et al. 2009, Chap. 2): errors of measurement (i.e., deviations of the values in the cells of the data matrix from the truth), errors of representation (i.e., differences in the composition of the analyzed data relative to the target population), and errors occurring during data analysis – although the latter are often not explicitly considered in TSE-based operations. Whether data analysis employs ML or not, the TSE framework remains a powerful tool for planning data collection and considering data quality (Puts et al. 2022). Yet, we suggest that the extensions of the TSE to newer data sources (e.g., Big Data, see Amaya et al. 2020) or to data sources more general than surveys (West et al. 2023) may provide additional value.

Accuracy metrics are at the core of the training, evaluation, and selection of models in the supervised ML paradigm (see Sect. 2.2). For the selection in particular, two types of comparisons exist. First, the relative comparison of models, i.e., against each other, is used for model selection. If NSOs wish to test new procedures involving ML against existing procedures, both must be evaluated on equal ground: ideally, on the same unseen evaluation data and in a manner identical to what the actual implementation in practice would look like. This is also true for ML-assisted procedures, e.g., combining a model's results and human work. Second, there is also an absolute comparison of a (chosen) model's accuracy: how well does it perform? Does it achieve a required minimum standard? On the issue of absolute comparisons, we would like to highlight a common problem in the discussion of (binary) classifiers: often, a high overall accuracy (i.e., the proportion of correct predictions $\hat{y} = y$) such as 0.91 is touted as evidence for a great model, implying that the suitable reference point might be 0.50 or 0. However, the correct reference point is the frequency of the majority class – and this piece of information is often missing from performance discussions. To see why it is crucial, consider an imbalanced classification problem in which the majority class occurs with a frequency of 0.9. The simple model containing only a constant and hence always predicting the majority class thus has an accuracy of 0.9.[36] Suddenly, the added predictive ability of $0.01 = 0.91 - 0.9$ achieved by the selected model and its features is recognized to be only tiny.[37] Alternatively, as in Sect. 5, one may use *balanced accuracy*, i.e., the unweighted average of the true positive rate and the true negative rate, for which the constant classifier always achieves a value of 0.5.

Accurate results hinge on accurate, i.e., high-quality training data (e.g., Coronado and Juárez 2020, p. 8). Deviations might come in (or be remedied) at any level de-

[36] Suppose that the majority class is the 'positive' class ($y = 1$). The constant model then has a true positive rate or sensitivity of 1.0, as it predicts only positive labels, and a true negative rate of specificity of 0. It can be shown that its accuracy is equal to the frequency of the majority class.

[37] Note the contrast to regression. A regression model that only contains the intercept has, by definition, $R^2 = 0$, and for a model containing only irrelevant features one would expect $R^2 \approx 0$ so that 0 is indeed a valid reference point.

scribed by the TSE framework: i.e., on the measurement side, the choice and design of constructs and of proxy variables, measurement instruments, human annotations of the collected raw data, etc., and, on the representation side, the choice of a population frame (coverage), sampling schemes and sample size, strategies about how to combat nonresponse, decisions about which units to exclude during data (pre-)processing, and so on. Their consequences depend on the type of data analysis. If prediction is the ultimate goal, then the error mechanisms in the training data should be as close as possible to those in the deployment data. This is another example of transportability (see Sect. 4.4). If, however, the data are analyzed with traditional statistics, any information about the error components and mechanisms is helpful for deriving unbiased estimates via, mostly, measurement error models or mixed (hierarchical, multi-level) models. In survey data, the contributions at different levels are acknowledged, e.g., via fixed or random effects on the level of respondents, interviewers, and items (e.g., Couper and Kreuter 2013). Yet, similar information about data processing is often not released: e.g., who annotated a particular data point – a ML model or a human (and if so, a pseudo-id for the particular annotator). We must acknowledge that research on how to optimize guidelines, instructions, and other characteristics for annotation tasks is still nascent (Beck et al. 2022; but see, e.g., Fort 2016 on annotating texts).

Accuracy not only concerns the outputs (estimates or predictions) but, especially in the ML paradigm, also the performance evaluations. Adherence to good practices documented in Sect. 2.2 is key, but violations are not necessarily obvious. In particular, any type of initial, exploratory data analysis (influencing feature engineering) and kind of data processing should only be done on the training data, not the whole data including the evaluation data, in order to prevent data leakage and overoptimistic performance evaluations. Target leakage should also be avoided – but this is difficult when data processors do not know the eventual data analysis, i.e., they do not know which variables are outcomes.

Particularly for results of data analysis released by NSOs, uncertainty assessments are required (Yung et al. 2022, p. 13). While traditional statistical methods come with 'self-assessed' uncertainty quantifications, some ML model classes do not (e.g., support vector machines) while others do (e.g., random forests). For classification tasks, the predicted class probabilities are an uncertainty measure; however, ML model classes typically do not quantify how uncertain these probabilities themselves are.[38] One way to express uncertainty is, instead of point predictions, to output prediction sets (for multi-class outcomes, e.g., many image classification tasks) or prediction intervals (for continuous outcomes): conformal prediction (e.g., Angelopoulos and Bates 2022) is a technique to turn predicted probabilities or scores into such sets or intervals – with desirable properties even when the underlying model is not perfect.[39]

[38] Class probabilities express aleatoric uncertainty: that which is caused by the randomness inherent to the non-deterministic relation $y|x = f(x)$ depicted by the 'true model' f. The uncertainty about the predicted probabilities is epistemic: one does not know how the trained model $\hat{f}$ deviates from the truth f. See, e.g., Bengs et al. (2022).

[39] Note that prediction intervals of traditional statistics are based on the assumption of having specified the model correctly.

Note that the ML performance comparisons typically do not involve the uncertainty inherent to having to estimate the models' accuracy. This is unsatisfactory for NSOs: e.g., changing an existing process to a new, ML-based procedure is not cost-free and an organization wants to have some level of confidence about what procedure it should choose.

Finally, we should note that robustness and/or reproducibility without accuracy are typically of little value. (For instance, one could imagine a procedure that outputs the same value every time, regardless of the training data that was used. This is a very robust, precise, and repeatable procedure, but its predictions would not reflect reality in any way.) This re-emphasizes the importance of considering accuracy, as implied by the trade-offs implicit in the bias-variance decomposition (see Sect. 2.2) and in the duality of validity and reliability (see footnote 34).

5 Mapping Fairness to the Quality Dimensions of QF4SA

In the QF4SA, fairness considerations are, at best, discussed as a secondary aspect, e.g., in the context of explainability. As the frameworks' main focus on "intermediate outputs" contrasts with the typical ML use cases that are discussed in the fair ML literature, this missing link may not be surprising. However, as we argue in the following sections, bringing in a fairness perspective, both conceptually and in practice, is critical for a wide range of ML applications, particularly including the uses of ML that are (prospectively) prominent at NSOs (see Sect. 2.4). We connect quality considerations with fairness in two steps: first, in this section, we map each of the existing quality dimensions of the QF4SA to fairness aspects. Second, in Sect. 6, we present how fairness considerations extend beyond the current scope of the QF4SA, identifying neglected aspects in the current framework. The interactions between algorithmic fairness and the QF4SA contribute to the existing quality dimensions by highlighting blind spots and introducing methodology that targets explainability, reproducibility, robustness, and accuracy from a different angle while, at the same time, pointing to fairness as a quality dimension on its own right.

Empirical example We make use of a machine learning application for algorithmic profiling in the public sector (Körtner and Bonoli 2022; Desiere et al. 2019) to illustrate how fairness considerations may be mapped to the QF4SA. All models that are presented are based on data from German administrative labor market records, concretely the *Sample of Integrated Employment Biographies* (SIAB, Antoni et al. 2019) maintained by the Research Data Center of the German Federal Employment Agency at the Institute for Employment Research (IAB). The data include information on (un)employment histories of job seekers for the period between January 1, 2010 and December 31, 2016. The prediction task is to classify, at entry into unemployment, whether an unemployment episode will last longer than one year (long-term unemployment; LTU). For more details see Kern et al. (2021).

5.1 Explainability & Interpretability

Explainability and fairness can be viewed as strongly intertwined processes throughout the ML pipeline. At the development stage, IML methods can help understand whether and how a model inherits societal biases. To this end, initial steps may include investigating the role and importance of (correlates of) protected or sensitive attributes and studying whether 'legitimate' features are utilized in different ways for social subgroups. At the deployment stage, the (perceived) degree of interpretability may shape fairness perceptions of the eventual 'user' of the algorithm, and their reliance on the model's outputs. If IML methods are used in either the production or deployment stage, another consideration is the degree to which the IML methods' *fidelity* varies by group:[40] if the explanations are not able to correctly reflect the models' decisions similarly across the feature space, any conclusions that are drawn about the models' functioning can be differently accurate across subgroups.

In practice, a first step towards merging model interpretation and fairness considerations may include the use of protected attributes as grouping variables to structure the application of IML techniques. Fig. 2, for example, shows two surrogate decision trees based on the same random forest model which predicts long-term unemployment of job-seekers. In Fig. 2a, only predictions for German job seekers are used to build the surrogate tree, whereas the tree in Fig. 2b is based on LTU predictions for non-German job seekers. Note that citizenship was not used as a predictor for the original random forest. In both surrogate trees, the duration of previous unemployment benefit receipt episodes (LHG dur) plays a major role in predicting (future) LTU, with longer receipt histories being associated with higher LTU risk. However, in the surrogate tree for German job seekers older age appears as an additional risk factor. This may indicate that the random forest learned different effect patterns for both subgroups – a finding which seems reasonable from a performance optimization

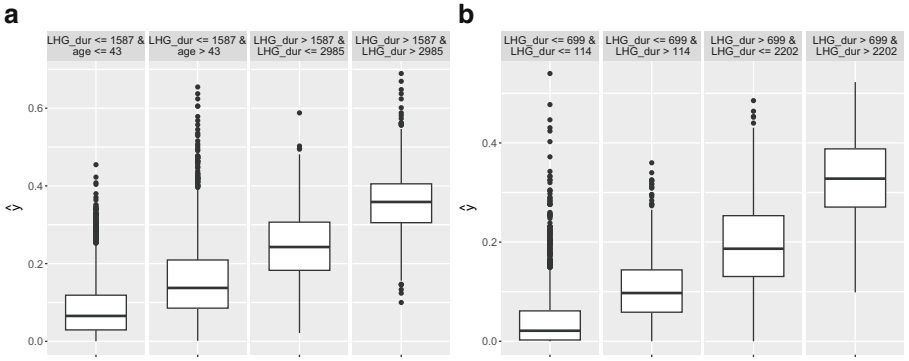

Fig. 2 Surrogate model explanations of a random forest predicting long-term unemployment, computed by protected group membership. **a** Surrogate tree for Germans, **b** Surrogate tree for non-Germans

[40] Recall that IML methods often involve approximations of the actual prediction model. Fidelity is the correctness or accuracy of how an IML method describes the model's behavior (Molnar 2020, Chap. 3.5) – crucial to the method's value.

perspective, but which also points to fairness implications in practice when groups defined by protected attributes are being scored on different grounds: it may be the case that predictions for one group stronger rely on features that are less reliable, exhibit (stronger) measurement error, or are proxy variables for (other) protected attributes such that structural differences map to differences in prediction quality. In any case, putting an emphasis on protected groups in the model interpretation process can be helpful to understand if a model may behave differently for important subgroups in downstream applications.

We note some further connections between explainability and fairness. Adversarial attacks are facilitated by an understanding of the model and of the data. Unfortunately, well-intentioned opportunities to probe a model offer such a gateway for attackers. Unchecked access to a model, particularly with IML tools but also under the guise of fairness evaluations, is more threatening than presenting some aggregate evaluation results. Of course, attacks can also occur on privacy (Pawelczyk et al. 2023). Very large models are able to memorize examples (Belkin et al. 2019; James et al. 2021, Chap. 10.8): allowing unchecked access to such models can then have similar consequences as just releasing the original training data would. Members of minority groups A, because of their smaller size, may be both, easier to re-identify (because of the small group size) and more likely or vulnerable to suffer negative consequences from re-identification. We are not aware of privacy metrics used as fairness metrics (but see Wachter et al. 2017 arguing for privacy-preserving IML).

Individual-level explanations match well with individual-level fairness notions. *Algorithmic recourse* is the notion of giving explanations and recommendations (how to achieve a desired prediction) to the individual, particularly in the form of counterfactual explanations (Verma et al. 2022; Karimi et al. 2021). While giving explanations is often desirable, particularly in the context of ADM, the extent to which there exist legal rights to explanations for individuals or legal mandates for organizations gets commonly overestimated (Doshi-Velez et al. 2019; Wachter et al. 2017). NSOs may also care about *individual* data points: during data analysis, some may be uncovered as influential to the estimates of a statistical model, and during data processing, some may be flagged as outliers; these might cause performance, robustness, and fairness problems. If such data points were produced by a ML model, e.g., in data imputation, IML-based evaluation of these imputations can help to solve the just-mentioned problems.

5.2 Cost effectiveness

In most parts, the costs of adopting ML at NSOs as presented in Yung et al. (2022) are seen to reflect technical needs such as IT infrastructure, maintenance, and staff training. We want to re-emphasize quality assurance and control as a critical component not only as a means to monitor machine learning models with respect to, e.g., fluctuations in (subgroup) performance, but also as a safety measure: Humans may (need to) overwrite the models' output if the uncertainty exceeds a pre-specified threshold (Bhatt et al. 2020). Introducing a 'reject option' in supervised learning models, i.e., forwarding difficult cases to humans for classification, can increase

error-fairness (Kaiser et al. 2022), but by definition comes at the cost of additional manual work. Assessing the need and degree of human oversight thus should be factored into the cost-benefit analysis of high-stakes ML applications at NSOs. Furthermore, fairness cannot be fully automated (Weerts et al. 2023).

Some of the new data sources might be cheaper to acquire than traditional data sources, but the savings might not hold up when the additional work to clean up elevated fairness problems is figured in. This is particularly true for 'found data' (Groves 2011; Japec et al. 2015) over which NSOs and other stakeholders have little to no discretion in design. Even large sample sizes cannot make up for errors of measurement and representation: If a key variable is subject to differential measurement error or if a protected group is missing it does not matter how many observations are in a data set.[41] Likewise, to give an extreme example, if the training data do not contain any women at all, it does not matter how much you increase the sample size – a model trained solely on men will tend to yield poor predictions for women in domains in which the two are markedly different.

A similar argument pertains to data processing: e.g., it might be better to use a medium size survey – perhaps one that combines survey responses with other data types – than to use a record linkage model that introduces fairness problems. In data analysis, simpler models have lower demands for data volume and computational resources, for both training and prediction, implying cost and time savings. In addition, higher interpretability may lead to better discovery and removal of fairness problems. Sophisticated, more flexible models might provide more accuracy and thus could be fairer by being more likely to discover model heterogeneity. Thus, both should always be included among the set of models considered for selection.

5.3 Timeliness

Adding fairness to the discussion and evaluation of quality dimensions should not be perceived as an additional burden to NSOs. As we try to argue and illustrate throughout this section, fairness considerations can be integrated into existing evaluation procedures in practice and can be viewed as an additional safeguard to ensure that the improvement in timeliness that may be achieved through ML-based automation does not come at the cost of disparate impact downstream. At the same time, the quality dimensions of the QF4SA framework each can benefit from a fairness perspective as it enriches the evaluation of algorithms by highlighting the critical role of (social) subgroups.

[41] 'Differential measurement error' describes situation in which the error mechanism is not the same for the whole population. Most salient from a fairness perspective: If you imagine the measured value of a quantitative variable to be the true value plus an error term e, then the expected value or the variance of e might be bigger for some demographic groups than for others. Similarly, for categorical variables, the frequency of measuring the incorrect class may also depend on group membership. For instance, web-scraped data of small enterprises may be more accurate for certain demographic groups than for others. For categorical variables in particular, it is surprisingly unlikely that measurement errors (technically: misclassification errors) will cancel each other out (Gruber et al. 2023, Chap. 4.2.3, 4.2.4). Under which exact conditions even non-differential measurement error can create or exacerbate fairness problems is the subject of ongoing research by the authors, but it is already clear that if groups differ on the base rate of an outcome Y, then even non-differential errors can produce fairness problems.

An interesting development that implicitly links fairness to timeliness is the work on fairness-aware automated machine learning (Weerts et al. 2023). In this line of research, methods are being proposed that can improve timeliness and cost by automating parts of the machine learning pipeline, while the resulting output is also required to fulfill some fairness constraints. While it is important to recognize the limits of such an approach – the authors agree that fairness cannot be fully automated –, fairness-aware AutoML can still expand the methodological toolkit of NSOs.

Timeliness, cost of data collection, and overall sample size are reasons to try to predict, e.g., response propensity in surveys. It has been recognized that focusing recruitment efforts on units with a high predicted propensity to participate is tempting on the aforementioned dimensions but widens the potential for representation bias when response propensity also depends on the outcome variable Y for the eventual data analysis. From a fairness perspective, it is important to note that hard-to-survey or hard-to-reach subpopulations (Tourangeau et al. 2014; Willis et al. 2014) may often coincide with groups for which we worry about discrimination and biased outputs (e.g., Keusch et al. 2021). Note that even if there is no relation between outcome and response propensity in a group, if the group is small in the collected data, the power to detect model heterogeneity is diminished and fairness evaluations become more statistically uncertain.

5.4 Robustness

The lack of robustness and stability can be connected to fairness concerns in multiple ways. On the organizational level, model decay or drift (i.e., deteriorating model accuracy over time) can be a reason for (potentially selective) skepticism towards algorithmic solutions (Choi et al. 2022, p. 2). In downstream applications, model drift can affect different parts of the target population in different ways. That is, differential error (see footnote 41) across subgroups may surface or may be amplified due to shifts in the data to which the model is applied. It may also be harder to detect model drift that occurs mainly or first in (small) protected groups. Also, one type of drift or drift indicator, namely the emergence of new categories in a categorical feature or outcome variable, might itself be directly about the existence and recognition of protected groups. To our knowledge, monitoring of models and decisions about the need to re-train to date consider global performance measures (e.g., Choi et al. 2022). From the fairness perspective, we suggest that (sub)group measures must also be monitored: this will not only inform when error-fairness drops below a pre-specified threshold, but might also inform about the causes and possible countermeasures. We further argue that careful monitoring is also needed even if models are re-trained on a regular basis, as new biases may be picked up along the way.

Furthermore, the robustness of a model within a group and the (epistemic) uncertainty of a model's predictions for a group have, to our knowledge, not been seen as fairness criteria so far. We suggest that these desirable global model properties should be also investigated as individual or (sub)group fairness notions and metrics in the algorithmic fairness literature. This pertains to models that are used in data collection, data processing, and data analysis.

One way to robustify models against drifts (Varshney et al. 2022, Chap. 9.4.3) is to employ causal models: e.g., causal relationships rooted in physics or biology are assumed to be more stable over time than spurious correlations.[42] Causal features and causal fairness notions are also being discussed (e.g, Makhlouf et al. 2022; Plecko and Bareinboim 2022). Yet, while employing causal features may be attractive in some settings, for images or other unstructured data types analyzed with Deep Learning, traditional features (in a potentially causal sense) are hardly involved.

Monitoring fairness metrics can be particularly important in a deployment context that includes data sources that capture complex, natural processes. In Fig. 3 we hold the model design constant, that is, random forest models for predicting long-term unemployment are used with the same hyper-parameter settings and features, but we repeatedly train and test models with data that change over time. Specifically, we use labor market records from 2010–2016 and train one random forest model for each year, and evaluate the respective model with data from the next year. The bold black line shows the difference in overall model performance (balanced accuracy) as we move from one year to the next. From this point of view, we might conclude that we can safely apply our random forest modeling schema over time without any major disruptions. However, assessing fairness metrics points to a different conclusion: a considerable increase in false negative rate (FNR) differences between non-

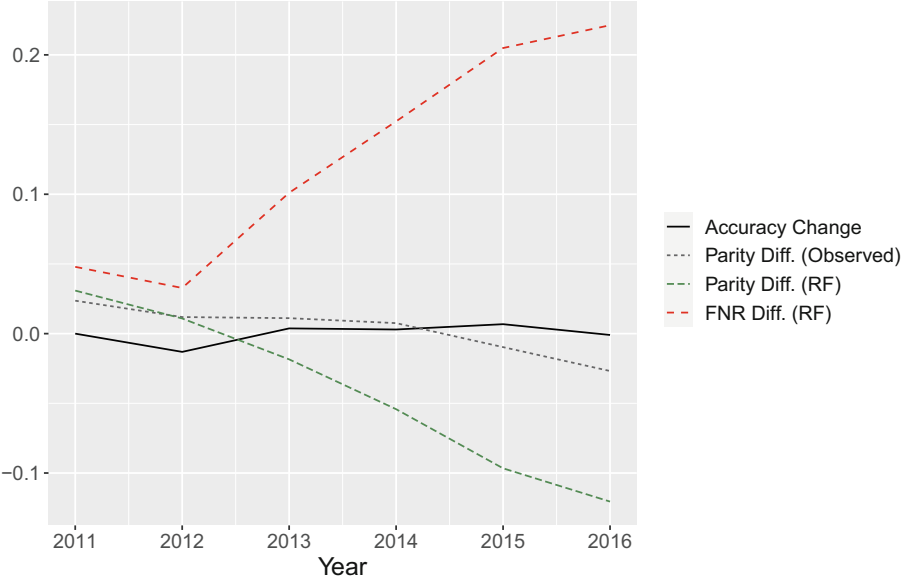

Fig. 3 (Change in) prediction performance and selected fairness metrics for random forest models over time. For each year, a new random forest is trained and evaluated with data from the next year. Parity difference scores show the difference in predicted LTU rates between non-German and German job seekers. FNR difference scores show the difference in false negative rates between non-Germans and Germans

[42] It is also plausible that causal features are harder to game than spurious features (which have no effect on Y and thus may be changed at low 'cost'), making causal models more robust to adversarial attacks.

German and German job seekers (dashed red line) can be observed when training and evaluating models with more recent data. This is accompanied by increasing parity differences in the models' predictions (dashed green line), which over-amplify the true differences in base rates as observed in the data (dotted gray line). As such changes over time can have considerable implications in practice, requiring robustness assessments to also consider subgroup-specific (fairness) metrics appears advisable.

5.5 Reproducibility

From a fairness perspective, (inferential) reproducibility raises questions as to how strongly design decisions in the machine learning pipeline affect outcomes not just overall, but also separately for sensitive subgroups of the target population. Fairness-relevant decision points may not only include the machine learning model itself (e.g., the model type and hyperparameter settings), but also more subtle aspects that include implicit decisions in data pre-processing steps (e.g., NSOs may employ a standard procedure for imputing missing values, while different imputation strategies can affect fairness measures in different ways; Caton et al. 2022). In practice, the implications of non-reproducibility may again be assessed by structuring model evaluations by protected attributes, paired with a grid of design decisions that is centered around the intended deployment setup.

A strong susceptibility to design decisions is of particular concern if the model outputs are further used downstream, either as an input to further analysis or to directly inform actions. Fig. 4 focuses on the effects of different hyperparameter settings on the classifications of random forest models predicting long-term unemployment. Four forests were trained that differ in the number of trees and the minimum size of the trees' terminal nodes and are then used to predict LTU, using the same classification threshold (top 25%). The Jaccard similarities, denoting the overlap (between 0 and 1) between the LTU predictions of the different random forest models are plotted, separately for German (Fig. 4a) and non-German (Fig. 4b) job seekers. Considering the modest changes that were made in the random forests' setup, we observe non-trivial differences between the lists of job seekers that are predicted as being at high risk of LTU by each model. While this generally holds for both German and non-German job seekers, the lowest agreement in predictions is recorded in Fig. 4b (between RF 2 and 4). Assessing the susceptibility of outcomes to small changes in the modeling design with a focus on protected groups thus may allow to identify variation that can challenge both overall reproducibility and the consistency of outcomes for societal subgroups.

As stressed in Sect. 4.5, methods reproducibility is increasingly viewed as a minimum standard. If the root causes of a model's discovered fairness problems are investigated upstream, the respective models that were used, e.g., in data processing must be reproducible or the search for problems and solutions may be futile.

In the three aforementioned definitions of reproducibility (see Sect. 4.5), one criterion was whether the methods are kept the same. As described in Sect. 2.2, for supervised ML problems, typically many supervised ML model classes are trained and only one model is selected, typically based on performance. In that sense,

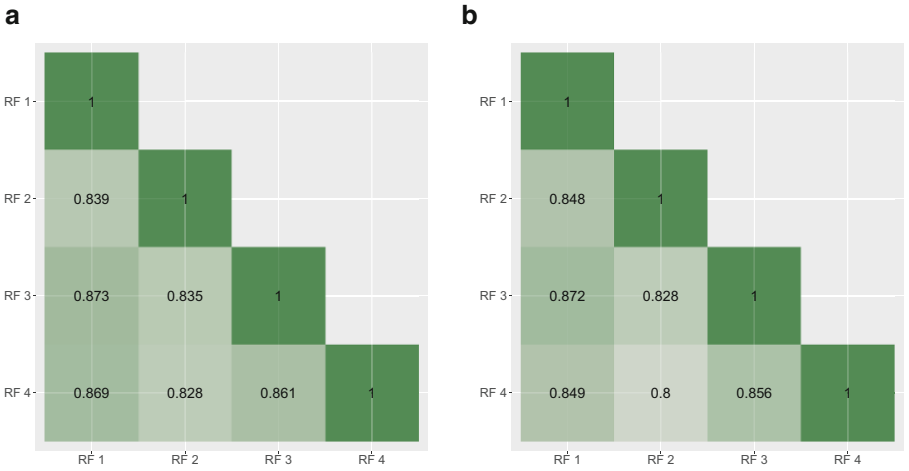

Fig. 4 Jaccard similarities between LTU predictions of random forest models with different hyper-parameter settings (RF 1: ntree = 750, nodesize = 1, RF 2: ntree = 250, nodesize = 1, RF 3: ntree = 500, nodesize = 5, RF 4: ntree = 500, nodesize = 15), computed by protected group membership. **a** Similarities between LTU predictions for Germans, **b** Similarities between LTU predictions for non-Germans

some exploration of the role of methods is more built-in for ML than for traditional statistics – although it would not be accurate to suggest that ML practice is anywhere close to a *multiverse analysis* approach, in which *all* decisions that can be made by an analyst are evaluated (Steegen et al. 2016). Simson et al. (2024) propose to look at all these choices – including those that affect how the raw data are fed into model training – and how they affect fairness.

5.6 Accuracy

Fairness interacts with accuracy (and with the human component) at multiple stages of the production process of NSOs. In the context of data processing and preparation, we note that one of the purported benefits of automation is an increase in consistency: e.g., in annotation tasks, even subject matter experts can disagree (Sthamer 2020b, p. 12) and, over time, an annotator might become tired or less motivated (inter- and intra-annotator reliability, respectively).[43] Meanwhile, a model will 'decide' the same way every time – but it is trained to reproduce the patterns contained in the training data, including those made by human annotators. First, consider a single annotator. She or he is or feels required to provide a label even for difficult-to-decide cases. Absent any other option, the annotator might resort to the marginal distribution of $Y|A$ (or their subjective notion thereof), even when Y is independent of A given X. The trained model will then learn that A (or proxies of A) are predictive of the provided labels (even though it is not predictive of the true Y). Solutions may include letting annotators express uncertainties instead of forcing a choice; this

[43] The same is true for editing and imputation (Dumpert 2020, p. 6; Sthamer 2020a) and other data processing tasks. We will focus on annotation as an example.

may actually fit well with the aforementioned reject-option models (Gruber et al. 2024; Bhatt et al. 2020). Also, information about A may be better hidden from annotators, although it can be difficult to do so in, e.g., image-based classification tasks. A second mechanism of how annotations can induce biased models predictions is a non-random allocation of observations to a set of inherently heterogeneous annotators: e.g., if units from $A = a$ are mostly processed by an annotator with a low general propensity to label $Y = 1$ and, conversely, units from $A \neq a$ are mostly processed by an annotator with a high general propensity for $Y = 1$, again A becomes predictive of the labels even though when it is not related to the true values Y. This is an example of how biases can be introduced during data processing even when there is no overt discrimination. A similar argument pertains to data collection, e.g., the allocation of interviewers might cause measurement or representation errors. We see two options here: NSOs can use stratified randomization to allocate data points to annotators and they can release annotator IDs because, conditional on the annotator, the spurious relationship of A and the labels vanishes.

Aggregation of data is one of the core tasks of NSOs: both in terms of data analysis, which may be simple descriptive statistics, and in terms of producing (aggregate) data for release. For the former, ML may seem hardly helpful for the estimation of population parameters: Other than (short) trees, ML model classes hardly possess parameters that correspond to interpretable, meaningful population characteristics. Also, systematically biased estimation, due to the bias-variance trade-off caused by training supervised ML models to minimize the expected prediction error, is problematic. However, ML can be useful when a parameter of interest is not identical across all subpopulations. For up to a medium number of pre-identified subpopulations, multiple testing correction can be employed to limit the error of falsely claiming heterogeneity (GCSILab 2023, Chap. 4.1). If there are, however, many subpopulations to investigate, as is the case with intersectional fairness, or if there are no pre-specified hypotheses at all, ML can help to discover heterogeneity: data splitting is then the procedure that guards against false discovery (GCSILab 2023, Chap. 4.2). If interpretable heterogeneity is the goal, trees for univariate statistics or, for more complex analyses, causal trees (Athey and Imbens 2016) appear to be most suitable.[44] We suggest that NSOs use such methodology for more fair reporting of the results of data analysis: subgroups for which the parameter deviates more than a pre-specified, meaningful amount from the global average should be identified and reported along with the global average. Aside from fairness, this approach can also be used to determine whether the global parameter value is meaningful and worth reporting at all: from Simpson's paradox, it is well known that the global value may be completely different than the value in all subpopulations, e.g., taking on a different sign.

For error-based fairness notions, the same methodology can be used to find subpopulations that suffer from more errors. Similarly, if there is gold-standard evaluation data for a data production process whether based on human work, ML, traditional statistics, or a combination thereof, such as editing and imputation, NSOs

[44] Such methods may be developed mostly for causal inference, but (finding) heterogeneity is also relevant for more descriptive data analysis.

can find groups for which their process performs more poorly compared to others or to an absolute threshold. If such heterogeneity is found, it may help – but will not replace subject matter and data knowledge – in fixing deficiencies in data collection (e.g., improving survey questions to yield less measurement error for A) and data processing systems. To our knowledge, using heterogeneity-finding ML machinery has not been explicitly suggested in the fair ML literature for either fair reporting of data analysis or finding biases in the data (production process), potentially with the exception of Zahn et al. (2023).

Self-assessed confidence measures by ML models may be used to decide whether something should be labeled by the model or be referred to a human expert (e.g., Text Classification Theme Group 2022 for text classification). However, not every ML model class yields self-assessed uncertainty measures and for those that do, there is no guarantee that they are accurate on average. Moreover, a model's overconfidence may not be the same for every group but could be worse for some (protected) groups or individuals. Some uncertainty measures also do not recognize epistemic uncertainty (i.e., uncertainty related to the correct specification of the structure of a model class and to the estimation of model parameters due to sampling uncertainty; Gruber et al. 2023) which may be greater for small minorities A. Uncertainty evaluations thus must also use actual evaluation data and cannot solely rely on models' self-assessments.

In a supervised learning context, accuracy as a quality dimension can be naturally extended to capture fairness concerns by requiring accurate predictions not just overall, but also for subgroups which may be defined by protected attributes or other features that are viewed as substantively relevant in a given application (Kim et al. 2019; Hebert-Johnson et al. 2018). Based on our long-term unemployment prediction example, Fig. 5 shows balanced accuracy scores of a random forest predicting LTU computed for 48 subgroups in the test set. Specifically, subgroups of job seekers were defined by intersections of the attributes citizenship, gender, age group, and region. While the model achieves an overall balanced accuracy score of 0.667, considerable variation in subgroup performance can be observed. Accuracy ranges between 0.417 and 0.8, indicating that prediction performance is no better, and sometimes worse, than random guessing for some demographic subgroups. The strongest variation in scores can be observed for non-German job seekers (upper half of Fig. 5), i.e., for subsets of the minority group in our example. While this finding may in part be driven by small sample sizes in some cells (although all but three cells include more than 50 observations), it highlights the utility of assessing subgroup accuracy as a means to provide pointers for further model investigation.

6 Fairness Beyond the Quality Dimensions of QF4SA

One mechanism causing fairness problems is what we call unrecognized model heterogeneity: i.e., the true functional relationship between features and outcome for units from some group A is not identical to the relationship in the rest of the population. If there are too few examples from A in the training data, the power to detect the correct model for A is low. There are several upstream causes for this

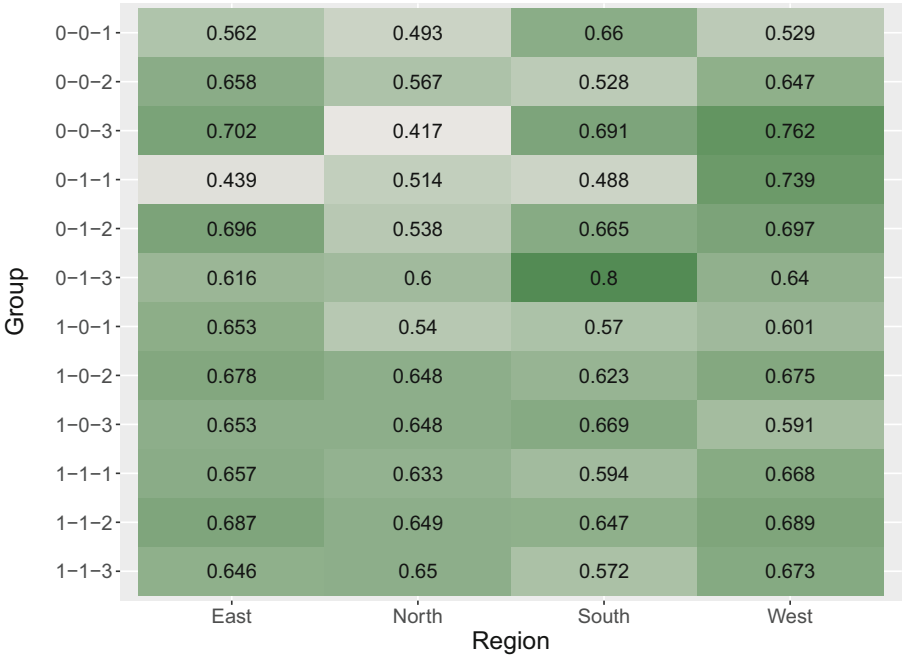

Fig. 5 Subgroup prediction performance (balanced accuracy) of a random forest predicting long-term unemployment. Group coding scheme: Citizenship (0: non-German, 1: German) – Gender (0: Male, 1: Female) – Age group (1: 18–30, 2: 31–50, 3: > 50)

phenomenon. Coverage errors, sampling errors, or unit nonresponse patterns may be such that members of A are underrepresented. Processing may also contribute. Unsupervised outlier detection methods identify unusual data points: thus, units from a small minority A are at risk for being falsely detected and removed – not because of erroneous values, but simply because of their membership in an infrequent group.

Unsupervised identification for record linkage or duplicate removal can also be sensitive to group membership. For instance, name-based distance metrics may be impacted when foreign names have multiple transliterations into the NSO's language or when, e.g., self-chosen Western first names are used in one data set and the original, non-Western first name in the other. Also, the relative frequency of first and last name combinations may be higher or lower for members of some group A than in the general population, hurting or benefiting their record linkage success. Supervised record linkage may be trained on data sources that contained relatively few recognized links (i.e., the label in this case) for members of A. This could be because of differential label error. It could also be that a low base rate (of correct links for members of A) was correct for the original training data sources, but is not for the data sources one currently wants to link.

NSOs as data producers can investigate at which processing steps many units from group A were lost, in relative or absolute terms. Two questions arise. First, which groups should be considered? A general, standard canon of groups to consider plus application-specific groups based on subject matter knowledge are obvious starting

points, and is the focus of algorithmic fairness literature shaped by the notion of protected attributes defined by law. This is also the subject of ethical and legal discussions that methodologists can and should not resolve on their own. We suggest supplementing this with a data-driven approach: groups for which the loss of units, in absolute terms or relative to the rest, is above a certain threshold. The loss of units can be calculated with regard to the previous step in multi-step data processing or, where applicable, with respect to the true distribution of characteristics in the target population (based on large-scale, gold-standard distributional data such as censuses). Second, which criteria should be applied? How to set the threshold for permissible relative loss of units in order to improve internal processes and the data is an organizational decision, based on the available resources. The criterion for absolute loss of units should be tied to the consequence, the loss of statistical power. How many units are needed to detect model heterogeneity beyond a certain level for A? How many units are needed for fairness evaluations with a pre-specified tolerable uncertainty? Unfortunately, the answers are largely application-specific and NSOs cannot anticipate all possible applications by external data users. Still, fairness report cards and metadata for released data should include information on losses of units that exceed thresholds. Users should be put in the position to be able to decide whether a certain product fits their needs and fairness demands – outright, after supplementation with other data sources, or not at all.

7 Discussion

The advent of *automated* decision-making and the rigorous focus on performance inherent to the ML mindset likely both contributed to the rise of fair ML. We argued in this paper that fairness is also a desirable, perhaps even necessary quality dimension of the work of NSOs – similar to how fairness is one dimension of frameworks for Trustworthy ML. This is, more generally, true for all data collection, processing, or analysis processes in official statistics: those that use ML or automation, but also those that employ traditional methods or human work. Nonetheless, the deployment of ML re-amplifies the need for explainable, reproducible, robust, and accurate products and data production processes at NSOs, highlighting quality dimensions that critically interact with fairness considerations as outlined in this article.

We further discussed the importance of the *human component* (Sect. 3.2) in (fair) ML at NSOs. In the pure ML world, some may believe that domain knowledge is unnecessary and that ML models, enough data, and ML knowledge are all that is required. The 'end-to-end' promise of Deep Learning being able to turn (seemingly) raw data into the desired predictions may add to that view.[45] We believe it is unwarranted. For instance, for some unlabeled data, e.g., images, subject matter expertise is required to produce the high-quality annotations on which the model's eventual success depends (Julien 2020, p. 2). For models that assist staff in, e.g., coding,

[45] We caution that many applications, particularly those working with survey data, lack the enormous training data required to render (knowledge-based) feature engineering obsolete.

what makes good suggestions and how humans interact with the model's output can be highly context-dependent.

It is similar with fairness. Our suggestion of data-driven finding and reporting of unfairness (e.g., Sect. 4.6) is a complement, not a replacement for legal knowledge and ethical considerations. For instance, debating which groups might be impacted the most and thus deserve fairness evaluations requires knowledge of the specific context and the general working of society. Once fairness problems have been detected, the work to find the causes and solutions starts. This is especially true for NSOs for whom not publishing data or statistics that are too unfair is often not an option, but who instead must find a way to improve. Finding discrimination in the data or making the annotation process less biased are among the tasks that require e.g., subject matter experts, statisticians, and methodologists.

Another aspect of the importance of the human factor is the willingness to accept systems that involve ML. Beyond the macro level, the different individual stakeholders need to be on board (Julien 2020, Chap. 6f.): e.g., from the (internal and external) users of a system and its output to anyone whose work is affected, such as experts whose roles are shifted. There are two core factors (Julien 2020, Chap. 6): First, such systems must demonstratedly serve the individual and organizational "business needs". Second, a trusted quality framework must form the basis: it guides the workflow (to prevent problems) and the actual performance on its quality dimensions is transparently and credibly evaluated. Fairness, as its own quality dimension, in its interaction with the other dimensions, and as part of legal and ethical considerations plays a big part in this.

Lastly, even if individuals are hesitant to embrace new ML methods outright, it can be still advisable to broaden the toolbox: if in comparison a statistical method performs similarly well, one can, in good conscience, use the more known, and interpretable traditional method; if the statistical method is however vastly outperformed, then this is at least a call to critically assess violations of the assumptions baked into the statistical method. At any rate, there is no need to let these disadvantages keep institutions from profiting from the positives of ML methods.

We further emphasize the critical role of data quality and its interaction with (fair) ML at NSOs. It is no secret that ML applications depend on the quantity and, although sometimes neglected, on the quality of training data. Relative to other data producers, NSOs have a long track record, extensive expertise, and legal obligation to (data) quality principles (e.g., Eurostat 2017, p. 7; Julien 2020, Chap. 2). We believe that NSOs as a whole also have a competitive advantage because of their commitment to collaboration (UNECE 2013, Principles 8-10; Eurostat 2017, Principle 1bis): beyond the sharing of code and knowledge (Julien 2020, Chap. 7), we suggest the different entities can pool training data and share in the expensive, but crucial human annotation tasks. This will increase efficiency and cross-organizational consistency. While the fairness dimension implies further requirements for the metadata and other documentation to be released alongside with NSOs' data and statistics products, the transparent publishing of such valuable, credible documentation can also be seen as a competitive advantage of NSOs over their competitors (Julien 2020, Chap. 2) and of NSOs' products over 'found data'. We thus argue that fairness need not be

seen as an additional burden, but rather caters toward the key objective of NSOs of releasing high-quality data products.

8 Appendix

Table 1 List of abbreviations and acronyms

AI	Artificial Intelligence
ADM	Automated Decision-making (Sect. 3.1)
DGP	Data-generating process
ESPE	Expected squared out-of-sample prediction error (Sect. 2.2)
FNR	False Negative Rate (Sect. 5.4)
IML	Interpretable Machine Learning; same as XAI (Sect. 4.1)
LLM	Large Language Model (Sect. 2.4.4)
LTU	Long-term unemployment (Sect. 5.1)
ML	Machine Learning
NSO	National Statistical Organization
TSE	Total Survey Error (Sect. 4.6)
QF4SA	Yung et al. (2022)'s Quality Framework for Statistical Algorithms (Sect. 4)
XAI	Explainable Artificial Intelligence; same as IML (Sect. 4.1)

Acknowledgements This work has been partially supported by the Federal Statistical Office of Germany.

Funding Open Access funding enabled and organized by Projekt DEAL.

Conflicts of interest We have no competing interests to declare.

References

Agnew W, Bergman AS, Chien J, Díaz M, El-Sayed S, Pittman J, Mohamed S, McKee KR (2024) The illusion of artificial inclusion. In: Proceedings of the CHI Conference on Human Factors in Computing Systems CHI '24. Association for Computing Machinery. https://doi.org/10.1145/3613904.3642703

AlgorithmWatch (2019) Atlas of Automation. Automated decision-making and participation in Germany. https://atlas.algorithmwatch.org/en/. Accessed 29 June 2024

Allhutter D, Cech F, Fischer F, Grill G, Mager A (2020) Algorithmic profiling of job seekers in Austria: how austerity politics are made effective. Front Big Data. https://doi.org/10.3389/fdata.2020.00005

Amaya A, Biemer PP, Kinyon D (2020) Total error in a big data world: adapting the TSE framework to big data. J Surv Stat Methodol 8(1):89–119

Angelopoulos AN, Bates S (2022) A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv: 2107.07511

Angwin J, Mattu S, Kirchner L (2016) *Machine Bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed 13 Mar 2023

Antoni M, Ganzer A, vom Berge P (2019) Sample of integrated labour market biographies regional file (SIAB-R) 1975–2017. FDZ-Datenreport, 04/2019 (en). Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB). https://doi.org/10.5164/IAB.FDZD.1904.en.v1

Argyle LP, Busby EC, Fulda N, Gubler JR, Rytting C, Wingate D (2023) Out of one, many: using language models to simulate human samples. Polit Anal 31(3):337–351. https://doi.org/10.1017/pan.2023.2

Athey S, Imbens G (2016) Recursive partitioning for heterogeneous causal effects. Proc Natl Acad Sci USA 113(27):7353–7360. https://doi.org/10.1073/pnas.1510489113

Bach RL, Kern C, Bonnay D, Kalaora L (2022) Understanding political news media consumption with digital trace data and natural language processing. J Royal Stat Soc Ser A 185(S2):S246–S269. https://doi.org/10.1111/rssa.12846

Baker R (2017) Big data: a survey research perspective. In: Biemer PP, de Leeuw ED, Eckman S, Edwards B, Kreuter F, Lyberg LE, Tucker NC, West. Hoboken BTNJ (eds) Total survey error in practice. John Wiley, pp 47–69

Barba LA (2018) Terminologies for reproducible research. arXiv: 1802.03311

Barocas S, Hardt M, Narayanan A (2023) Fairness and machine learning: limitations and opportunities. MIT Press, Cambridge (www.fairmlbook.org)

Barocas S, Selbst AD (2016) Big data's disparate impact. Calif Law Rev 104(3):671–732

Beck J, Eckman S, Chew R, Kreuter F (2022) Improving labeling through social science insights: results and research agenda. In: Chen JYC, Fragomeni G, Degen H, Ntoa S (eds) HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence. Springer, Cham, pp 245–261

Beck M, Dumpert F, Feuerhake J (2018a) Machine Learning in Official Statistics. arXiv: 1812.10422

Beck M, Dumpert F, Feuerhake J (2018b) Proof of Concept Machine Learning. Abschlussbericht. Federal Statistical Office of Germany (Destatis), Wiesbaden. https://www.statistischebibliothek.de/mir/receive/DEMonografie_mods_00004835

Belkin M, Hsu D, Ma S, Mandal S (2019) Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proc Natl Acad Sci USA 32:15849–15854
Bellamy D, Hernán MA, Beam A (2022) A structural characterization of shortcut features for prediction. Eur J Epidemiol 37(6):563–568
Benedikt L, Joshi C, Nolan L, de Wolf N, Schouten B (2020) Optical character recognition and machine learning classification of shopping receipts. Report. HBS An app-assisted approach for the Household Budget Survey. https://ec.europa.eu/eurostat/documents/54431/11489222/6+Receipt+scan+analysis.pdf
Bengs V, Hüllermeier E, Waegeman W (2022) On the difficulty of epistemic uncertainty quantification in machine learning: the case of direct uncertainty estimation through loss minimisation. arXiv: 2203.06102
Bhatt U, Zhang Y, Antorán J, Liao QV, Sattigeri P, Fogliato R, Melançon GG, Krishnan R, Stanley J, Tickoo O, Nachman L, Chunara R, Weller A, Xiang A (2020) Uncertainty as a form of transparency: measuring, communicating, and using uncertainty. arXiv: 2011.07586
Binns R (2018) Fairness in machine learning: lessons from political philosophy. arXiv: 1712.03586
Bommasani R et al (2021) On the opportunities and risks of foundation models. arXiv: 2108.07258
Bothmann L, Peters K, Bischl B (2022) What is fairness? Implications for fairML. arXiv: 2205.09622
Bothmann L, Wimmer L, Charrakh O, Weber T, Edelhoff H, Peters W, Nguyen H, Benjamin C, Menzel A (2023) Automated wildlife image classification: an active learning tool for ecological applications. arXiv: 2303.15823
Breiman L (2001) Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci 16(3):199–231. https://doi.org/10.1214/ss/1009213726
Buolamwini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In: Friedler SA, Wilson C (eds) Proceedings of the 1st conference on fairness, accountability and transparency Proceedings of Machine Learning Research. vol 81. PMLR, pp 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html
Burton JW, Stein M-K, Jensen TB (2020) A systematic review of algorithm aversion in augmented decision making. J Behav Decis Mak 33(2):220–239
Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM (2006) Measurement error in nonlinear models: a modern perspective, 2nd edn. Chapman and Hall, CRC, Boca Raton
Caton S, Haas C (2024) Fairness in machine learning: a survey. ACM Comput Surv 56(7):1–38. https://doi.org/10.1145/3616865
Caton S, Malisetty S, Haas C (2022) Impact of imputation strategies on fairness in machine learning. J Artif Intell Res. https://doi.org/10.1613/jair.1.13197
Chen J, Beam A, Saria S, Mendonça EA (2019) Potential trade-offs and unintended consequences of artificial intelligence. In: Matheny M, Israni ST, Ahmed M, Whicher D (eds) Artificial intelligence in health care: the hope, the hype, the promise, the peril. National Academy of Medicine, Washington, DC, pp 99–130. https://nam.edu/wp-content/uploads/2019/12/AI-in-Health-Care-PREPUB-FINAL.pdf
Choi I, del Monaco A, Law E, Davies S, Karanka J, Baily A, Piela R, Turpeinen T, Mharzi A, Rastan S, Flak K, Jentoft S (2022) ML model monitoring and re-training in statistical organisations. ONS-UNECE Machine Learning Group 2022, Theme Group – Model Retraining, v2. https://statswiki.unece.org/display/ML/Machine+Learning+Group+2022
Chouldechova A (2016) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. arXiv: 1610.07524
Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B (2019) A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 110:12–22
Clemmensen LH, Kjærsgaard RD (2023) Data representativity for machine learning and AI systems. arXiv: 2203.04706
Coronado A, Juárez J (2020) UNECE – HLG-MOS Machine Learning Project. Imagery Theme Report. v1. https://statswiki.unece.org/display/ML/WP1+-+Theme+3+Imagery+Analysis+Report. Accessed 13 Mar 2023
Couper M, Kreuter F (2013) Using paradata to explore item level response times in surveys. J Royal Stat Soc Ser A 176(1):271–286
Couper MP (2017) New developments in survey data collection. Annu Rev Sociol 43:121–145
Creel K, Hellman D (2022) The algorithmic leviathan: arbitrariness, fairness, and opportunity in algorithmic decision-making systems. Can J of Philosophy 52(1):26–43. https://doi.org/10.1017/can.2022.3

Curtin C, Senanayake P, Clarke C, Lichtenstein I, Jamieson A, Roshanafshar S, Yung W, Piela R, Vaiciulis V, del Monaco A, Palumbo L, Toepoel V, Tingay K, Banks A, Bogdanova B, Sirello O, Zdanowicz K, Museux J-M, Tessitore C, Danforth J, Tebrake J, Choi I, Kipkeeva A (2023) Large language models for official statistics. HLG-MOS white paper. https://unece.org/sites/default/files/2023-12/HLGMOS%20LLM%20Paper_Preprint_1.pdf. Accessed 8 Dec 2023

Desiere S, Langenbucher K, Struyven L (2019) Statistical profiling in public employment services. OECD Social, Employment and Migration Working Papers 224. Organisation for Economic Cooperation and Development (OECD), Paris. https://doi.org/10.1787/b5e5f16e-en

Destatis (2021) Quality Manual of the Statistical Offices of the Federation and the Länder. (Original title: Qualitätshandbuch der Statistischen Ämter des Bundes und der Länder). Statistische Ämter des Bundes und der Länder, Wiesbaden. https://www.destatis.de/DE/Methoden/Qualitaet/qualitaetshandbuch.pdf

Domscheit-Berg A (2024) Press release: Federal government is using more and more AI, ignoring sustainability and failing to establish structures (Original title: Pressemitteilung: Bund nutzt immer mehr KI, ignoriert dabei Nachhaltigkeit und versäumt Aufbau von Strukturen). https://mdb.anke.domscheit-berg.de/2024/07/pm-kleineanfrage-kuenstliche-intelligenz-bund/. Accessed 24 July 2024

Doshi-Velez F, Kim B (2017) Towards A rigorous science of interpretable machine learning. arXiv: 1702.08608

Doshi-Velez F, Kortz M, Budish R, Bavitz C, Gershman S, O'Brien D, Scott K, Schieber S, Waldo J, Weinberger D, Weller A, Wood A (2019) Accountability of AI under the law: the role of explanation. arXiv: 1711.01134

Dumpert F (2020) UNECE – HLG-MOS machine learning project. Edit and imputation theme report. https://statswiki.unece.org/display/ML/WP1+-+Theme+2+Edit+and+Imputation+Report. Accessed 13 Mar 2023

Dutta S, Long J, Mishra S, Tilli C, Magazzeni D (2022) Robust Counterfactual explanations for tree-based ensembles. In: Chaudhuri K, Jegelka S, Song L, Szepesvari C, Niu G, Sabato S (eds) Proceedings of the 39th International Conference on Machine Learning Proceedings of Machine Learning Research. vol 162. PMLR, pp 5742–5756. https://proceedings.mlr.press/v162/dutta22a.html

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on – ITCS '12. ACM Press, Cambridge, pp 214–226. https://doi.org/10.1145/2090236.2090255

Díaz-Rodríguez N, Del Ser M, Coeckelbergh M, López de Prado E, Herrera-Viedma, Herrera F (2023) Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. In: Inf Fusion 99, p 101896. https://doi.org/10.1016/j.inffus.2023.101896

Eckman S (2013) Paradata for coverage research. In: Kreuter F (ed) Improving surveys with Paradata: analytic uses of process information. Wiley, Hoboken, pp 97–116

Engstrom DF, Ho DE, Sharkey CM, Cuéllar M-F (2020) Government by algorithm: artificial intelligence in federal administrative agencies. Public Law Research Paper 20-54. NYU School of Law, New York. https://doi.org/10.2139/ssrn.3551505

EU *AI Watch*. Artificial intelligence website of the European Commission's Joint Research Centre. https://ai-watch.ec.europa.eu/index_en. Accessed 29 June 2024

Eurostat (2017) European Statistics Code of Practice. Revised edition 2017. https://ec.europa.eu/eurostat/web/products-catalogues/-/ks-02-18-142. Accessed 13 Mar 2023

Fort K (2016) Collaborative annotation for reliable natural language processing: technical and sociological aspects. Wiley, Hoboken. https://hal.science/hal-01324322

Fürnkranz J, Gamberger D, Lavrač N (2012) Foundations of rule learning. Springer, Heidelberg

Gajane P, Pechenizkiy M (2018) On formalizing fairness in prediction with machine learning. arXiv: 1710.03184

GCSILab (2023) Machine learning-based causal inference tutorial. https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/. Accessed 4 Aug 2023

Gerdon F, Bach RL, Kern C, Kreuter F (2022) Social impacts of algorithmic decision-making: a research agenda for the social sciences. Big Data Soc 9(1):1–13. https://doi.org/10.1177/20539517221089305

Ghani R, Schierholz M (2020) Machine learning. In: Foster I, Ghani R, Jarmin RS, Kreuter F, Lane J (eds) Big data and social science, 2nd edn. CRC Press, Boca Raton, Chap. 7. https://textbook.coleridgeinitiative.org

Goodman SN, Fanelli D, Ioannidis JPA (2016) What does research reproducibility mean? Sci Transl Med. https://doi.org/10.1126/scitranslmed.aaf5027

Grgic-Hlaca N, Redmiles EM, Gummadi KP, Weller A (2018) Human perceptions of fairness in algorithmic decision making: a case study of criminal risk prediction. In: Proceedings of the 2018 World Wide Web Conference on World Wide Web – WWW '18. ACM Press, pp 903–912. https://doi.org/10.1145/3178876.3186138
Grinsztajn L, Oyallon E, Varoquaux G (2022) Why do tree-based models still outperform deep learning on tabular data? arXiv: 2207.08815
Groves RM (2011) Three eras of survey research. PUBOPQ 75(5):861–871
Groves RM, Fowler FJ Jr, Couper MP, Lepkowski JM, Singer E, Tourangeau R (2009) Survey methodology, 2nd edn. Wiley, Hoboken
Gruber C, Hechinger K, Assenmacher M, Kauermann G, Plank B (2024) More labels or cases? Assessing label variation in natural language inference. In: Pyatkin V, Fried D, Stengel-Eskin E, Liu A, Pezzelle S (eds) Proceedings of the third workshop on understanding implicit and underspecified language. Association for Computational Linguistics, pp 22–32. https://aclanthology.org/2024.unimplicit-1.2
Gruber C, Schenk PO, Schierholz M, Kreuter F, Kauermann G (2023) Sources of uncertainty in machine learning – A statisticians' view. arXiv: 2305.16703
Guts Y (2020) Workshop on target leakage in machine learning. https://github.com/YuriyGuts/odsc-target-leakage-workshop. Accessed 29 June 2023
Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986) Robust statistics: the approach based on influence functions. Wiley, Hoboken
Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York. https://hastie.su.domains/ElemStatLearn/
Hebert-Johnson U, Kim M, Reingold O, Rothblum G (2018) Multicalibration: calibration for the (Computationally-identifiable) masses. In: Dy J, Krause A (eds) Proceedings of the 35th International Conference on Machine Learning Proceedings of Machine Learning Research. PMLR. vol 80, pp 1939–1948
Heidari H, Loi M, Gummadi KP, Krause A (2019) A moral framework for understanding fair ML through economic models of equality of opportunity. In: Proceedings of the conference on fairness, accountability, and transparency. Association for Computing Machinery, pp 181–190. https://doi.org/10.1145/3287560.3287584
Helwegen R, Braaksma B (2020) Fair algorithms in context. Working paper 05–20. Center for Big Data Statistics. https://www.cbs.nl/-/media/_pdf/2020/22/cbds_working_paper_fair_algorithms.pdf
Herrmann M, Lange FJD, Eggensperger K, Casalicchio G, Wever M, Feurer M, Rügamer D, Hüllermeier E, Boulesteix A-L, Bischl B (2024) Position: why we must rethink empirical research in machine learning. In: Forty-first international conference on machine learning. https://openreview.net/forum?id=DprrMz24tk
Herzog TN, Scheuren FJ, Winkler WE (2007) Data quality and record linkage techniques. Springer, New York
von der Heyde L, Haensch A-C, Wenz A (2024) Vox Populi, Vox AI? Using Language Models to Estimate German Public Opinion. arXiv: 2407.08563
Hill CA, Biemer P, Buskirk T, Callegaro M, Córdova Cazar AL, Eck A, Japec L, Kirchner A, Kolenikov S, Lyberg L, Sturgis P (2019) Exploring New Statistical Frontiers at the Intersection of Survey Science and Big Data: Convergence at 'BigSurv18'. Surv Res Methods 13(1):123–135
Hill CA, Biemer PP, Buskirk TD, Japec L, Kirchner A, Kolenikov S, Lyberg LE (2021) Big data meets survey science: a collection of innovative methods. Wiley, Hoboken
Holloway J, Mengersen K (2018) Statistical machine learning methods and remote sensing for sustainable development goals: a review. Remote Sens 10:9. https://doi.org/10.3390/rs10091365
Hornik K (2005) A CLUE for CLUster ensembles. J Stat Soft 14:12. https://doi.org/10.18637/jss.v014.i12
Hou YT-Y, Jung MF (2021) Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making. In: Proceedings of the ACM on Human-Computer Interaction CSCW2. vol 5. Association for Computing Machinery, pp 1–25. https://doi.org/10.1145/3479864
Huber PJ, Ronchetti EM (2009) Robust statistics. Wiley, Hoboken
Ilic G, Lugtig P, Schouten B, Streefkerk M, Mulder J, Kumar P, Höcük S (2022) Pictures instead of survey questions: an experimental investigation of the feasibility of using pictures in a housing survey. J Royal Stat Soc Ser A: Stat Soc 185(Supplement 2):S437–S460. https://doi.org/10.1111/rssa.12960
IPS Observatory IPS-X. The innovative public services explorer. https://ipsoeu.github.io/ips-explorer/. Accessed 29 June 2024
James G, Witten D, Hastie T, Tibshirani R (2021) An Introduction to Statistical Learning. with Applications in R. 2nd edn. Springer, New York (First printing August 4, 2021). https://www.statlearning.com. Accessed 31 August 2021

Japec L, Kreuter F, Berg M, Biemer PP, Decker P, Lampe C, Lane J, O'Neil C, Usher A (2015) Big data in survey research: AAPOR task force report. PUBOPQ 79(4):839–880

Julien C (2020) UNECE – HLG-MOS Machine Learning Project Project report. v2. https://statswiki.unece.org/display/ML/Machine+Learning+Project+Report. Accessed 13 Mar 2023

Jussupow E, Benbasat I, Heinzl A (2020) Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. In: Proceedings of the 28th European Conference on Information Systems (ECIS), An Online AIS Conference. https://aisel.aisnet.org/ecis2020_rp/168

Kaiser P, Kern C, Rügamer D (2022) Uncertainty-aware predictive modeling for fair data-driven decisions. arXiv: 2211.02730

Kapoor S, Narayanan A (2022) Leakage and the reproducibility crisis in ML-based science. arXiv: 2207.07048

Karimi A-H, Barthe G, Schölkopf B, Valera I (2021) A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. arXiv: 2010.04050

Kearns M, Neel S, Roth A, Wu ZS (2018) Preventing fairness gerrymandering: auditing and learning for subgroup fairness. In: Dy J, Krause A (eds) Proceedings of Machine Learning Research, vol 80. PMLR, pp 2564–2572. https://proceedings.mlr.press/v80/kearns18a.html

Kern C, Bach R, Mautner H, Kreuter F (2021) Fairness in algorithmic profiling: a German case study. arXiv: 2108.04134

Kern C, Gerdon F, Bach RL, Keusch F, Kreuter F (2022) Humans versus machines: who is perceived to decide fairer? Experimental evidence on attitudes toward automated decision-making. Patterns 3(10):100591. https://doi.org/10.1016/j.patter.2022.100591

Keusch F, Kreuter F (2021) Digital trace data: modes of data collection, applications, and errors at a glance. In: Engel U, Quan-Haase A, Liu SX, Lyberg L (eds) Handbook of computational social science, vol 1. Routledge, Taylor & Francis, New York, Chap. 7. https://doi.org/10.4324/9781003024583-8

Keusch F, Leonard MM, Sajons C, Steiner S (2021) Using Smartphone technology for research on refugees: evidence from Germany. Sociol Methods Res 50(4):1863–1894. https://doi.org/10.1177/0049124119852377

Keusch F, Struminskaya B, Eckman S, Guyer HM (in preparation) Data collection with wearables, apps, and sensors. CRC Press, Boca Raton

Kilbertus N, Rojas-Carulla M, Parascandolo G, Hardt M, Janzing D, Schölkopf B (2017) Avoiding discrimination through causal reasoning. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates, pp 656–666

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

Kleinberg J, Raghavan M (2021) Algorithmic monoculture and social welfare. Proc Natl Acad Sci USA 118(22):e2018340118. https://doi.org/10.1073/pnas.2018340118

König G (2023) If interpretability is the answer, what is the question? – A causal perspective. Dissertation, Ludwig-Maximilians-Universität München, Munich. https://doi.org/10.5282/edoc.32614 (Dissertation)

Körtner J, Bonoli G (2022) Predictive algorithms in the delivery of public employment services. https://doi.org/10.31235/osf.io/j7r8y

Kreuter F (ed) (2013) Improving surveys with paradata: analytic uses of process information. Wiley, Hoboken

Krishna S, Han T, Gu A, Pombra J, Jabbari S, Wu S, Lakkaraju H (2022) The disagreement problem in explainable machine learning: a practitioner's perspective. arXiv: 2202.01602

Kuppler M, Kern C, Bach R, Kreuter F (2022) From fair predictions to just decisions? Conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making. Front Sociol. https://doi.org/10.3389/fsoc.2022.883999

Kusner MJ, Loftus JR, Russell C, Silva R (2018) Counterfactual Fairness. arXiv: 1703.06856

Lakkaraju H, Slack D, Chen Y, Tan C, Singh S (2022) Rethinking explainability as a dialogue: a practitioner's perspective. arXiv: 2202.01875

Lee, M. S. A., L. Floridi, and J. Singh (2020). *Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics*. Research Paper. Previously titled: From Fairness Metrics to Key Ethics Indicators (KEIs): A Context-Aware Approach to Algorithmic Ethics in an Unequal Society. Centre for Digital Ethics (CEDE). https://doi.org/10.2139/ssrn.3679975.

Ligozat A-L, Lefèvre J, Bugeau A, Combaz J (2022) Unraveling the hidden environmental impacts of AI solutions for environment. arXiv: 2110.11822

Lipton ZC (2018) The mythos of model Interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue 16(3):31–57. https://doi.org/10.1145/3236386.3241340
Little RJ, Rubin DB (2019) Statistical analysis with missing data. Wiley, Hoboken
Loi M, Herlitz A, Heidari H (2021) Fair equality of chances for prediction-based decisions. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society AIES '21. Association for Computing Machinery, p 756. https://doi.org/10.1145/3461702.3462613
Ma B, Wang X, Hu T, Haensch A-C, Hedderich MA, Plank B, Kreuter F (2024) The potential and challenges of evaluating attitudes, opinions, and values in large language models. arXiv: 2406.11096
Makhlouf K, Zhioua S, Palamidessi C (2020) On the applicability of ML fairness notions. arXiv: 2006.16745
Makhlouf K, Zhioua S, Palamidessi C (2022) Survey on causal-based machine learning fairness notions. arXiv: 2010.09553
Measure A (2020) UNECE – HLG-MOS Machine Learning Project. Work Package 3 – Integration. v0.4 final. https://statswiki.unece.org/display/ML/WP3+-+Integration. Accessed 13 Mar 2023
Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. ACM Comput Surv 54:6. https://doi.org/10.1145/3457607
Miller T (2017) Explanation in Artificial Intelligence: Insights from the Social Sciences. arXiv: 1706.07269
Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Spitzer E, Raji ID, Gebru T (2019) Model cards for model reporting. In: Proceedings of the conference on fairness, accountability, and transparency. ACM, https://doi.org/10.1145/3287560.3287596
Mitchell S, Potash E, Barocas S, D'Amour A, Lum K (2021) Algorithmic fairness: choices, assumptions, and definitions. Annu Rev Stat Appl 8(1):141–163. https://doi.org/10.1146/annurev-statistics-042720-125902
Mitra N (ed) (2021) Observational studies 7.1: Special issue: commentaries on Breimen's two cultures paper. https://muse.jhu.edu/issue/45147
Mittereder FK (2019) Predicting and preventing Breakoff in web surveys. Dissertation, University of Michigan, Ann Arbor, MI. https://deepblue.lib.umich.edu/handle/2027.42/149963
Molnar C (2020) Interpretable machine learning. A guide for making black box models explainable. A guide for making black box models explainable, 2nd edn. Leanpub. https://christophm.github.io/interpretable-ml-book
Molnar C (2022) Modeling Mindsets. The Many Cultures of Learning From Data. Independently published at Leanpub. www.modeling-mindsets.com
Molnar C, König G, Herbinger J, Freiesleben T, Dandl S, Scholbeck CA, Casalicchio G, Grosse-Wentrup M, Bischl B (2022) General pitfalls of model-agnostic interpretation methods for machine learning models. In: Holzinger A, Goebel R, Fong R, Moon T, Müller K-R, Samek W (eds) xxAI Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020 Vienna, 18 July 2020 Springer, Cham, pp 39–68. https://doi.org/10.1007/978-3-031-04083-2_4 (Revised and Extended Papers)
Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F (2012) A unifying view on dataset shift in classification. Pattern Recognit 45(1):521–530
Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B (2019) Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci USA 116(44):22071–22080. https://doi.org/10.1073/pnas.1900654116
Myrtveit I, Stensrud E, Shepperd M (2005) Reliability and validity in comparative studies of software prediction models. IEEE Trans Softw Eng 31(5):380–391. https://doi.org/10.1109/TSE.2005.58
Neunhoeffer M, Wu ZS, Dwork C (2021) Private Post-GAN Boosting. arXiv: 2007.11934
Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453. https://doi.org/10.1126/science.aax2342
Ohme J, Araujo T, Boeschoten L, Freelon D, Ram N, Reeves BB, Robinson TN (2024) Digital trace data collection for social media effects research: aPis, data donation, and (screen) tracking. Commun Methods Meas 18(2):124–141. https://doi.org/10.1080/19312458.2023.2181319
Page ET, Antoun C, Gonzalez J, Kantor L, Keusch F, Miller L, Wenz A (2023) Editorial: recent advances in survey methods for collecting food data. In: Survey methods: insights from the field special issue. Food Acquisition Research and Methods, pp 1–8. https://doi.org/10.13094/SMIF-2023-00017
Pawelczyk M, Lakkaraju H, Neel S (2023) On the privacy risks of algorithmic recourse. In: Ruiz F, Dy J, van de Meent J-W (eds) Proceedings of The 26th International Conference on Artificial Intelligence and Statistics, vol 206. PMLR, pp 9680–9696 (Proceedings of Machine Learning Research). https://proceedings.mlr.press/v206/pawelczyk23a.html

Perdomo JC, Zrnic T, Mendler-Dünner C, Hardt M (2020) Performative Prediction. arXiv: 2002.06673
Plecko D, Bareinboim E (2022) Causal Fairness Analysis. arXiv: 2207.11385
Plesser HE (2018) Reproducibility vs. replicability: a brief history of a confused terminology. Front Neuroinform. https://doi.org/10.3389/fninf.2017.00076
Puts MJH, da Silva A, Di Consiglio L, Choi I, Salgado D, Clarke C, Jones S, Baily A (2022) ONS-UNECE machine learning group 2022. Quality of training data. Theme group report. v1. https://statswiki.unece.org/display/ML/Machine+Learning+Group+2022. Accessed 13 Mar 2023
Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND (eds) (2008) Dataset shift in machine learning. MIT Press, Cambridge
Raji ID, Denton E, Bender EM, Hanna A, Paullada A (2021) AI and the everything in the whole wide world benchmark. In: Thirty-fifth Conference on Neural Information Processing Systems. Datasets and Benchmarks Track (Round 2). https://openreview.net/forum?id=j6NxpQbREA1
Raper S (2020) Leo Breiman's "two cultures". Significance 17(1):34–37. https://doi.org/10.1111/j.1740-9713.2020.01357.x
Reusens M, Kurban B, Peszat K, Grancow B, Murawska E (2022) ML2022: Web scraping theme group report. v1. https://statswiki.unece.org/display/ML/Machine+Learning+Group+2022. Accessed 13 Mar 2023
Richards J, Piorkowski D, Hind M, Houde S, Mojsilović A (2020) A methodology for creating AI factsheets. arXiv: 2006.13796
Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, Aviles-Rivero AI, Etmann C, McCague C, Beer L et al (2021) Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell 3(3):199–217
Rodolfa KT, Saleiro P, Ghani R (2020) Bias and fairness. In: Foster I, Ghani R, Jarmin RS, Kreuter F, Lane J (eds) Big data and social science, 2nd edn. CRC Press, Boca Raton, Chap. 11. https://textbook.coleridgeinitiative.org
Saleiro P, Kuester B, Hinkson L, London J, Stevens A, Anisfeld A, Rodolfa KT, Ghani R (2019) Aequitas: a bias and fairness audit Toolkit. arXiv: 1811.05577
Salwiczek C, Rohde J (2022) Dimensions of quality for the use of ML in official statistics. Presented at the Workshop "Quality Aspects of Machine Learning – Official Statistics between Specific Quality Requirements and Methodological Innovation, Munich. Germany. https://ai-watch.ec.europa.eu/index_en
Schenk P, Reuß S (2024) Paradata in surveys. In: Huvila I, Börjesson L, Sköld O (eds) Perspectives to Paradata. Springer, Cham. https://doi.org/10.1007/978-3-031-53946-6_2
Schwanhäuser S, Sakshaug JW, Kosyakova Y (2022) How to Catch a Falsifier: Comparison of Statistical Detection Methods for Interviewer Falsification. PUBOPQ 86(1):51–81
Schwartz R, Dodge J, Smith NA, Etzioni O (2019) Green AI. arXiv: 1907.10597
Seibold H (2023) Bringing open science to formal education. https://heidiseibold.ck.page/posts/bringing-open-science-to-formal-education. Accessed 30 June 2024
Simson J, Pfisterer F, Kern C (2024) One model many scores: using multiverse analysis to prevent fairness hacking and evaluate the influence of model design decisions. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency FAccT '24. Association for Computing Machinery, pp 1305–1320. https://doi.org/10.1145/3630106.3658974
Srivastava M, Heidari H, Krause A (2019) Mathematical notions vs. Human perception of fairness: a descriptive approach to fairness for machine learning. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining KDD '19. Association for Computing Machinery, pp 2459–2468. https://doi.org/10.1145/3292500.3330664
Starke C, Baleis J, Keller B, Marcinkowski F (2022) Fairness perceptions of algorithmic decision-making: a systematic review of the empirical literature. Big Data Soc 9:2. https://doi.org/10.1177/20539517221115189
Statistics Norway (2024). *Adopting artificial intelligence in the production and dissemination of official statistics*. Tech. rep. Geneva, Switzerland: United Nations Economic, Social Council, Economic Commission for Europe, Conference of European Statisticians, Seventy-second plenary session, 20, and 21 June 2024. https://unece.org/statistics/documents/2024/05/working-documents/adopting-artificial-intelligence-production-and. Accessed 30 June 2024
Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W (2016) Increasing transparency through a multiverse analysis. Perspect Psychol Sci 11(5):702–712. https://doi.org/10.1177/1745691616658637

Sthamer C (2020a) Editing of Social Survey Data with Machine Learning – A journey from PoC to Implementation. v2, 2022-10-15. https://statswiki.unece.org/display/ML/Editing+of+Social+Survey+Data+with+Machine+Learning+-+A+journey+from+PoC+to+Implementation. Accessed 13 Mar 2023

Sthamer C (2020b) UNECE – HLG-MOS Machine Learning Project. Classification and Coding Theme Report. v6. https://statswiki.unece.org/display/ML/WP1+-+Theme+1+Coding+and+Classification+Report. Accessed 13 Mar 2023

Struminskaya B, Lugtig P, Toepoel V, Schouten B, Giesen D, Dolmans R (2021) Sharing data collected with Smartphone sensors: willingness, participation, and nonparticipation bias. Public Opinion Quarterly 85(S1):423–462. https://doi.org/10.1093/poq/nfab025

TAG Register Public Law Project Tracking Automated Government (TAG) Register. https://trackautomatedgovernment.shinyapps.io/register/. Accessed 29 June 2024

Text Classification Theme Group (2022) ML 2022 Text Classification Theme Group Report. v1. https://statswiki.unece.org/display/ML/Machine+Learning+Group+2022. Accessed 13 Mar 2023

Tokle J, Bender S (2020) Record linkage. In: Foster I, Ghani R, Jarmin RS, Kreuter F, Lane J (eds) Big data and social science, 2nd edn. CRC Press, Boca Raton, Chap. 3. https://textbook.coleridgeinitiative.org

Tornede T, Tornede A, Hanselle J, Wever M, Mohr F, Hüllermeier E (2022) Towards green automated machine learning: status quo and future directions. arXiv: 2111.05850

Tourangeau R, Edwards B, Johnson TP, Wolter KM, Bates N (eds) (2014) Hard-to-survey populations. Cambridge University Press, Cambridge

TrustML The Trustworthy ML Initiative. https://www.trustworthyml.org/resources

UK Statistics Authority (2021) Ethical considerations in the use of Machine Learning for research and statistics. ONS-UNECE Machine Learning Group 2021 Work Stream 3. https://uksa.statisticsauthority.gov.uk/publication/ethical-considerations-in-the-use-of-machine-learning-for-research-and-statistics/pages/1/. Accessed 13 Mar 2023

UNECE (2013) Fundamental Principles of Official Statistics. Resolution adopted by the Economic and Social Council on 24 July 2013. https://unstats.un.org/unsd/dnss/gp/FP-Rev2013-E.pdf. Accessed 30 June 2024

Varshney KR (2022) Trustworthy machine learning. Independently Published, Chappaqua (http://www.trustworthymachinelearning.com)

Verma S, Boonsanong V, Hoang M, Hines KE, Dickerson JP, Shah C (2022) Counterfactual explanations and algorithmic recourses for machine learning: a review. arXiv: 2010.10596v3

Wachter S, Mittelstadt B, Floridi L (2017) Why a right to explanation of automated decision-making does not exist in the general data protection regulation. Int Data Priv Law 7(2):76–99. https://doi.org/10.1093/idpl/ipx005

Wachter S, Mittelstadt B, Russell C (2017) Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv J Law Technol 31(2):841–887

Wagner JR (2008) Adaptive survey design to reduce nonresponse bias. University of Michigan, Ann Arbor. https://deepblue.lib.umich.edu/handle/2027.42/60831

Weerts H, Pfisterer F, Feurer M, Eggensperger K, Bergman E, Awad N, Vanschoren J, Pechenizkiy M, Bischl B, Hutter F (2023) Can fairness be automated? Guidelines and opportunities for fairness-aware autoML. arXiv: 2303.08485

West BT, Wagner J, Kim J, Buskirk TD (2023) The total data quality framework. https://www.coursera.org/specializations/total-data-quality. Accessed 13 Mar 2023

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva LBS, Bourne PE et al (2016) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3:1–9. https://doi.org/10.1038/sdata.2016.18

Willis GB, Smith TW, Shariff-Marco S, English N (2014) Overview of the special issue on surveying the hard-to-reach. J Off Stat 30(2):171–176

Yung W, Tam S-M, Buelens B, Chipman H, Dumpert F, Ascari G, Rocci F, Burger J, Choi I (2022) A quality framework for statistical algorithms. SJI 38(1):291–308. Page numbers referenced in the main text refer to the preprint available at https://statswiki.unece.org/download/attachments/285216420/QF4SA_2020_Final.pdf

von Zahn M, Hinz O, Feuerriegel S (2023) Locating disparities in machine learning. In: IEEE International Conference on Big Data (BigData). IEEE, pp 1883–1894. https://doi.org/10.1109/BigData59044.2023.10386485

Zenimoto Y, Hasegawa R, Utsuro T, Yoshioka M, Kando N (2024) Coding open-ended responses using pseudo response generation by large language models. In: Cao Y, Papadimitriou I, Ovalle A (eds) Proceedings of the 2024 conference of the north American chapter of the association for computational linguistics: human language technologies Student Research Workshop. vol 4. Association for Computational Linguistics, pp 242–254. https://aclanthology.org/2024.naacl-srw.26