Experience: Bridging Data Measurement and Ethical Challenges with Extended Data Briefs

MARCO RONDINA and ANTONIO VETRÒ, Politecnico di Torino, Italy
ALESSANDRO FABRIS, Max Planck Institute for Security and Privacy, Germany
GIANMARIA SILVELLO and GIAN ANTONIO SUSTO, Università di Padova, Italy
MARCO TORCHIANO and JUAN CARLOS DE MARTIN, Politecnico di Torino, Italy

To promote the responsible development and use of data-driven technologies –such as machine learning and artificial intelligence– principles of trustworthiness, accountability and fairness should be followed. The quality of the dataset on which these applications rely, is crucial to achieve compliance with the required ethical principles. Quantitative approaches to measure data quality are abundant in the literature and among practitioners, however they are not sufficient to cover all the principles and ethical challenges involved.

In this paper, we show that complementing data quality with measurable dimensions of data documentation and of data balance helps to cover a wider range of ethical challenges connected to the use of datasets in algorithms. A synthetic report of the metrics applied (the Extended Data Brief) and a set of Risk Labels for the Ethical Challenges provide a practical overview of the potential ethical harms due to data composition. We believe that the proposed data labelling scheme will enable practitioners to improve the overall quality of datasets and to build more responsible data-driven software systems.

CCS Concepts: • **Information systems** → **Data analytics**; *Decision support systems*; *Information integration*; • **General and reference** → **Measurement**; • **Mathematics of computing** → *Exploratory data analysis*; • **Social and professional topics** → **Socio-technical systems**; • **Software and its engineering** → *Documentation*.

Additional Key Words and Phrases: data quality, data documentation, data ethics

ACM Reference Format:
Marco Rondina, Antonio Vetrò, Alessandro Fabris, Gianmaria Silvello, Gian Antonio Susto, Marco Torchiano, and Juan Carlos De Martin. 2025. Experience: Bridging Data Measurement and Ethical Challenges with Extended Data Briefs. *ACM J. Data Inform. Quality* 1, 1 (March 2025), 26 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Data-driven technologies, particularly artificial intelligence (AI) and machine learning (ML) algorithms, have made significant technical advancements in recent years, impacting countless fields of human activities and, at the same time, raising concerns about their potential harms to society [8, 11, 20, 38]. As a consequence, the demand for a more responsible development and use of these

Authors' addresses: Marco Rondina, marco.rondina@polito.it; Antonio Vetrò, antonio.vetro@polito.it, Politecnico di Torino, Turin, Italy; Alessandro Fabris, alessandro.fabris@mpi-sp.org, Max Planck Institute for Security and Privacy, Bochum, Germany; Gianmaria Silvello, gianmaria.silvello@unipd.it; Gian Antonio Susto, gianantonio.susto@unipd.it, Università di Padova, Padova, Italy; Marco Torchiano, marco.torchiano@polito.it; Juan Carlos De Martin, demartin@polito.it, Politecnico di Torino, Turin, Italy.

1936-1955/2025/3-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

technologies –especially AI– has arisen from many quarters [17, 37, 39]. Several ethical principles are being considered for this goal, specially trustworthiness, accountability, and fairness [12, 16].

Since models learn from and are dependent on data, data quality is a key aspect for AI/ML. The traditional approach to data quality, such as the one defined in the ISO standard [24], involves assessing and ensuring, among other dimensions, accuracy, completeness, consistency, confidentiality and precision of data through various measures. These characteristics play a critical role in improving the reliability of software output, but this approach alone does not address all the ethical concerns associated with AI systems [15]. Our research questions arise from this gap: *Which data measures can help to assess the risk of all the ethical challenges of a data-driven system?* We propose to integrate the traditional approach with other relevant dimensions, namely documentation and balance. They provide a more comprehensive evaluation of the quality of the datasets, able to cover a wider range of ethical concerns. Data balance measures have been proven to be useful to identify the risks of automated discriminations [47, 48], with their long queue of systemic effects in society [6]. Documentation is a key aspect to improve in the development lifecycle of an AI system [41] and quality measures help to make the datasets (and their use) more transparent [44]. We relate each data measure to the possible ethical challenges associated with it by analysing which data characteristics may have an impact on these challenges. By ethical challenge, we refer to the issues raised by algorithms in transforming data into evidence for outcomes, in using those outcomes to motivate further action, and in accounting for the impacts of those actions [36].

In this *Experience* paper[1], we applied a set of selected measures on a sample of wide-known datasets, to produce an *Extended Data Brief* and a set of *Risk Labels for the Ethical Challenges*. We focused on categorical data, because most of the sensitive attributes [13] in datasets are categorical (e.g. gender, marital status, job, etc.). We illustrate both the potential benefits and drawbacks of integrating the approaches, allowing for a more holistic understanding of the quality of a dataset. Overall, this work contributes to the development of effective strategies to create, use and share training datasets in a more trusted, responsible and fair way. In addition, the scripts used are made available[2] to enhance reproducibility and to promote further improvements. The remainder of the paper is organized as follows: the Section 2 summarizes the related work, Section 3 presents the theoretical framework we propose, the Section 4 describes the methodology and measurements related to the application of the framework on a group of datasets. In the Section 5 we show results and discuss them. The Section 6 outlines the main challenges encountered during the research, while the Section 7 identify the main limitations and provide hints for future work. Section 8 recap the main elements and findings of the study.

2 BACKGROUND AND RELATED WORK

Faulty, noisy or inaccurate data easily leads to undesirable results [10, 27], hence the selection, creation and adoption of datasets is a critical but often undervalued step [45]. A growing body of literature has explored how to make the intrinsic characteristic of datasets [4, 7, 21], models [35, 43] or rankings [49, 50] emerge, since knowing the data problems is the very first step to managing them [25]. Different works investigate the different dimensions of data quality [3, 42, 46]: we propose to evaluate accuracy, consistency and completeness using measures from the ISO SQuaRE standards series [23]. Data quality in ISO/IEC 25012:2008 [22] is categorized into 15 characteristics, and each of these characteristics is quantifiable through measures of quality-related properties, defined in ISO/IEC 25024:2015 [24]. The characteristics belong either to the "inherent" point of view if dependent only on the data themselves, such as completeness. Otherwise, they belong to

[1]See https://dl.acm.org/journal/jdiq/call-for-papers#ExperiencePapers
[2]https://github.com/RondinaMR/data-qbd-framework

the "system-dependent" point of view, such as recoverability. They can also belong to both, such as efficiency. In the proposed framework, we rely on characteristics of the inherent point of view because they are the most general and applicable to any dataset.

Balance represents a homogeneous distribution of data between the classes of one or more attributes [19]. Lower levels of balance, especially in protected attributes or their proxies, are related to higher levels of unfairness in the output [47]. Different cases reveal the discriminatory risk associated with highly unbalanced datasets [8], highlighting the need to measure this data dimension. We use measures validated in previous work [32, 48]: the Gini index [9], the Shannon diversity index, the Simpson diversity index, and the Inverse Imbalance Ratio (I.I.R.).

Documentation plays a central role in the discovery of data characteristics. Many issues of fairness, transparency and accountability in ML/AI systems arise from the way data is collected, processed and used [26]: documentation helps to track the adopted procedures (and their implicit beliefs) and thus helps to mitigate risks [5]. Documentation plays an important role in ethical and legal analysis [40], so efforts are made to reduce technical debt as much as possible [1], despite the specificities of documentation in AI development [29]. Sambasivan et al. [45] report that a lack of data documentation hinders the generalization of models thus leading to poor model performance for underserved communities. Gebru et al. [18] proposed a list of questions useful to guide the writing of documentation by dataset creators and, based on these questions, a Documentation Test Sheet (DTS) [44] was created to measure the completeness of documentation. Fabris et al. [14] presented the *data brief* to document the most important properties of a dataset.

Several works provide guidance on the ethical challenges of algorithms. A notable contribution in this area is the work of Mittelstadt et al. [36], who developed a comprehensive map of the ethics of algorithms that provides a framework for understanding and addressing these challenges. The authors examined the gap between the design and implementation of algorithms and the understanding of their ethical implications. This work provides a comprehensive coverage of the different types of ethical challenges, as it also considers actions driven by system outcomes. It is widely recognised for its contribution to the analysis of algorithmic ethics. For these reasons, and given the applicability of this mapping to our research, we decided to use this work to map ethical challenges. The importance of this issue is heightened by the fact that these ethical implications can have profound consequences for individuals, groups and societies as a whole.

3 ETHICAL CHALLENGES AND RELATIONSHIPS WITH DATA DIMENSIONS

In this section, we first present the ethical challenges that we consider and then the data dimensions that aid in assessing datasets. Lastly, we illustrate the specific relationship between the two.

3.1 Ethical Challenges

Mittestald et al. [36] delineate three epistemic and two normative concerns, as well as one overarching challenge, based on how algorithms process data to produce evidence and motivate actions. Here, we briefly recap the six ethical challenges: i) *Inconclusive evidence*: using inferential statistics to draw conclusions from data may result in uncertain knowledge; ii) *Inscrutable evidence*: the link between data and conclusions may be unclear and hence problematic to scrutinise; iii) *Misguided evidence*: if the data is of low reliability or neutrality, the resulting outcomes will also lack reliability and neutrality; iv) *Unfair outcomes*: algorithms have the potential to support actions that do not align with the fairness ethical principle; v) *Transformative effects*: algorithms can affect how we conceptualise the world, and modify its social and political organisation; vi) *Traceability*: challenge related to the difficulties of finding the cause of a harmful outcome. We present the relationships between the ethical challenges and the data dimensions in Section 3.3.

Table 1. Data quality measures (ISO/IEC 25024) adapted to be applicable in the analysis of a general dataset. The arrows indicate the interpretation for each QM (the lower the better: ↓, the higher the better: ↑)

QM	Name	Definition
Acc-I-4 (↓)	Risk of dataset inaccuracy (Accuracy)	X = A/B A = number of data values that are outliers B = number of data values to be considered in a data set
Com-I-1-DevA (↑)	Record completeness (Completeness)	Average of X where X = A/B A = number of not null value in the whole data set B = number of data items considered
Com-I-5 (↑)	Empty record in a data file (Completeness)	X = 1-A/B A = number of records where all data items are empty B = number of records in a data file
Con-I-2-DevB (↑)	Data format consistency (Consistency)	Average of X where X = A/B A = number of data items that have the correct type B = number of data items considered for a single column
Con-I-3-DevC (↓)	Risk of data inconsistency (Consistency)	X = A/B A = Number of data items where exist duplication in value B = Number of the possible duplications
Con-I-4-DevD (↑)	Architecture consistency (Consistency)	X = A/B A = Number of rows that respect the data structure B = Number of rows contained in the data file

3.2 Data Dimensions

3.2.1 Data Quality (DQ). The metrics adopted from the ISO/IEC 25024:2015 standard [24] are shown in Table 1: we include assessments of accuracy, completeness, and consistency. Some measures (with suffix 'Dev') have been slightly adapted to the needs of this work, as described hereafter. The *Acc-I-4* quality measure (QM) was used as defined in the standard, detecting outliers using the Interquartile Range Method with k=1,5. The *Com-I-1-DevA* QM is defined in the standard as *Completeness of data items of a record within a data file*: in the context of this research, it has been adapted as a QM for the whole dataset, dividing the number of null values by the total number of data items. The *Con-I-2-DevB* QM is defined in the standard as *Consistency of data format of the same data item*: since it requires prior knowledge of the data attribute, it has been reformulated as the ratio of the number of data elements that have the correct type in the attribute to the number of data elements considered for a single column. The *Con-I-3-DevC* QM was slightly modified with respect to the definition present in the standard. For each attribute in the column i, there is a possibility of duplication. In addition, duplication can be identified by grouping k attributes together and searching for identical records across all rows. This phenomenon occurs when two or more records have the same values for a given set of k attributes. We looked for duplicates in a single column ($k = 1$) and in a pair of columns ($k = 2$) when applying our framework. Deviating from the standard, we have divided the number of data items where there is a duplication in value by the number of possible duplications. This was done with the aim of obtaining a measure between 0 and 1, even considering a k value of 2. The *Con-I-4-DevD* QM is defined in the standard as the *Degree to which the elements of the architecture have a correspondence in referenced architecture elements*. It was reformulated by specifying the concept of architecture in terms of data structure.

Table 2. Imbalance indexes: m represents the number of classes, f_i is the relative frequency of class i.

Index	Formula (normalized)	Notes
Gini	$G_n = \frac{m}{m-1} \cdot \left(1 - \sum_{i=m}^{m} f_i^2\right)$	Measure of heterogeneity [9]
Shannon	$S = -\left(\frac{1}{lnm}\right) \sum_{i=1}^{m} f_i ln f_i$	Measure of species diversity in a community
Simpson	$D = \frac{1}{m-1} \cdot \left(\frac{1}{\sum_{i=1}^{m} f_i^2} - 1\right)$	Probability that two individuals in a sample belong to the same class
Inverse Imbalance Ratio	$IR = \frac{\{min(f_i,\ldots,m)\}}{\{max(f_i,\ldots,m)\}}$	Ratio between the lowest and the highest frequency

Thus, the ratio became the ratio between the number of rows containing the correct number of values (i.e. columns, attributes) and the total number of rows.

3.2.2 Data Balance (DB). The balance measures adopted from [48] are presented in Table 2. All measures take values between 0 (imbalanced) and 1 (balanced). For each formula, m represents the number of classes, while f_i is the relative frequency of class i. Previous studies [33] have identified fairness implications when each imbalance index falls below a certain threshold: Gini < 40%, Shannon < 50%, Simpson < 30%, I.I.R. < 15%. In the *Extended Data Briefs* we use these thresholds to highlight unbalanced features. The Inverse Imbalance Ratio (I.I.R.) stands out as the most accurate metric for identifying class imbalances within a specific attribute based on selected sample distributions [48]. Yet, it proves to be highly sensitive in cases where classes have close to zero occurrences. Gini and Shannon indexes demonstrate, on average, the least effective performance [47], but they are useful in all cases in which it is desirable to have indexes that are very reactive to imbalance [48]. The Simpson index, instead, represents a very good compromise because it identifies imbalance more clearly [47], without being too sensitive. On the basis of this complementarity, the Simpson index is used to produce the risk labels, but during the discussion different indexes are used in conjunction.

3.2.3 Data Documentation (DD). To perform a quality analysis of the documentation, we used the Documentation Test Sheet (DTS) [44], designed to measure the completeness of the documentation of an ML/AI training dataset. It indicates how much of the relevant information is suitably documented. Its *Documentation Fields* are derived and adapted from different standardization proposals, mainly *Datasheets for Datasets* [4, 18, 21], and they are grouped into *sections* based on the type of information they represent. *1) Motivation* refers to the purpose of the dataset; *2) Composition* describes the characteristics of the data; *3) Collection processes* and *4) Data processing procedures* refer to the procedures adopted to collect and transform the data; *5) Uses* indicates how the dataset should (or should not) be used and *6) Maintenance* brings up all the details related to the evolution of the dataset over time. The individual *Documentation Field* can take on the value 0 (the related information is not available in the documentation under analysis) or 1 (the related information is available). In the Extended Data Briefs, we present the *Section Presence Average* calculated as the average of all the *Documentation Field* values of the specific section. Therefore, all the *Section Presence Averages* take values between 0 (no information is present) and 1 (all information is present).

Table 3. Mapping of ethical challenges with data dimensions. The presence of a bullet in a cell means that the ethical challenge is linked to the data dimension.

	Data quality (DQ)	Data balance (DB)	Data document. (DD)
Inconclusive evidence	•(1)		
Inscrutable evidence			•(8)
Misguided evidence	•(2)		•(9)
Unfair outcomes	•(3)	•(5)	
Transformative effects		•(6)	
Traceability	•(4)	•(7)	•(10)

3.3 Relationships from Ethical Challenges to Data Dimensions

We mapped how each data dimension (Data quality=DQ; Data balance=DB; Data documentation=DD) addresses the six ethical challenges described by Mittestald et al. [36]. Table 3 shows the relationships between the ethical challenges and the data dimensions. They can be explained as follows:

(1) *DQ and Inconclusive evidence.* Data quality affects the statistical properties of a dataset, and the conclusions that can be inferred from it.
(2) *DQ and Misguided evidence.* Conclusions are as reliable as input data, and data quality can be a proxy for the reliability of the evidence drawn from data.
(3) *DQ and Unfair outcomes.* Unfair outcomes can be caused by availability of low quality data for specific population groups.
(4) *DQ and Traceability.* Data quality may be responsible for problematic outcomes (i.e. outcomes vitiated by ethical challenges): in such cases, analysis of data quality measures makes it possible to link the outcome to its cause and the responsibilities associated with it.
(5) *DB and Unfair outcomes.* Imbalanced datasets may lead to imbalanced results, which means harmful differentiation of products, information and services based on personal characteristics. In applications such as wages, insurance, education, etc. such differentiation can lead to unjustified unequal treatment or discrimination based on a sensitive attribute.
(6) *DB and Transformative effects.* As motivated above, imbalanced data can cause polarized classifications in the allocation of resources, benefits, or penalties (e.g. welfare). This has transformative effects on entire segments of the population, amplifying existing inequalities in societies, and reinforcing distances between social classes.
(7) *DB and Traceability.* Data balance may be responsible for problematic outcomes, as described above. In the case of causes that are rooted in the balance of the data itself, analysis of data balance measures enables identification of the root cause of the problematic outcome and the corresponding responsibilities.
(8) *DD and Inscrutable evidence.* Documentation of the data is needed to ground the conclusions to decisions on how data was collected, labeled, which assumptions were made, how measurements were performed.
(9) *DD and Misguided evidence.* Data documentation is useful for clarifying the context in which data are collected, processed and used. Describing and identifying the limits of data validity helps to circumscribe the reliability of results.
(10) *DD and Traceability.* Documenting the characteristics of the data can be useful to clearly and explicitly identify data problems that need to be addressed. In addition, documentation of data collection and processing procedures makes it possible to analyse whether the causes of any problematic outcomes are to be found in these delicate steps. In all these cases, documentation helps to identify responsibility.

There are no explicit and ex-ante strategies for managing trade-offs between the ethical challenges presented: they are highly context dependent and it is up to the final users of the labels to decide which ethical challenges have higher priority in their own context. In such analysis, users might also take into account other aspects not considered in this framework, such as privacy (especially for inscrutable evidence and traceability) or currentness (especially for misguided evidence and unfair outcomes). The possible integration of these aspects will be the object of future investigations.

4 METHODOLOGY AND MEASUREMENTS

The whole framework is intended to be applicable to structured data. While DQ measures can be applied "to any kind of data held in a structured format" [24], and DD measures can be measured on metadata of any kind of data, DB can only be applied to structured, categorical features [33]. We have selected these by identifying the categorical sensitive features through Article 21 *"Non-discrimination"* of the EU Charter of Fundamental Rights [13]. Numerical sensitive features, such as non-discretised age, were excluded from the DB analysis.

We tested the proposed approach on a sample of algorithmic fairness datasets. Firstly, we selected the 10 most popular datasets from the collection[3] organised by Fabris et al. [14]: focusing on popular datasets allowed us to analyse very influential datasets [28]. The 10 selected datasets were: Adult, COMPAS, South German Credit, Communities and Crime, Bank Marketing, Law School, CelebA, MovieLens, Credit Card Default and Toy Dataset 1. We filtered non-textual data, excluding the CelebA dataset, as it is an image dataset: this decision is due to the fact that the DQ measures can only be computed on tabular data and the DB measures can only be calculated for categorical data. We also excluded Toy Dataset 1 because it is synthetic. As a consequence, eight datasets remained. As the records belonging to the *Communities and Crime* dataset refer to communities (not individuals) and are predominantly numerical, we decided to exclude them from the DB measurement. The labels of this dataset were calculated by considering only DQ and DD. In general, if the dataset contained an explicit target variable, this was also included in the DB analysis.

For each dataset, we developed an *Extended Data Brief*, extending the *Data Brief* presented in [14]. We added DQ, DB and DD measures. We completed it with the *Ethical Challenge Risk Labels*: on the basis of the relationships identified in Section 3.3, we related the overall risk of each data dimension to the ethical challenges impacted. For each quality measure (identified by ↑) we transformed the value into a risk measure (1-value); for each risk measure (identified by ↓) we summed the value itself. We then divided this sum by the number of measures in each dimension, to obtain a *data dimension risk ratio*. Finally, we averaged the *data dimension risk ratios* of all the data dimensions that could be attributed to each ethical challenge, to obtain an *ethical challenge risk ratio*. This is the value represented by the *Ethical Challenge Risk Labels*. As a measure for balance, we choose the Simpson index, for the reasons described in Section 3.2.2. The code used is available in the repository mentioned in footnote 2.

From the perspective of the dataset producer, the proposed framework should be used to provide a summary of the context of the dataset, its main qualities and limitations, including a disclaimer (in the form of *Ethical Challenge Risk Labels*) about the main risks embedded in the data. From the perspective of a dataset consumer, the framework is intended to make them aware – at the onset – of the main risks associated with using the dataset. This is similar to the way nutrition labels communicate the characteristics of a commercial food product. In the same intuitive way, users will become aware of these risks and decide for themselves how to proceed in a responsible use of the dataset (as done in the Dataset Nutrition Label framework [21]). Providing a technical mitigation solution is not an objective at this stage, but could be considered in future work.

[3] http://fairnessdata.dei.unipd.it/datasets, popularity was defined by the number of scientific articles that used the dataset.

5 RESULTS AND DISCUSSION

In the following subsections, we present the results on the three most popular datasets of our collection as distinct case studies: Adult (5.1), COMPAS (5.2) and South German Credit (5.3) . The *Ethical Challenge Risk Labels* and the *Extended Data Briefs* of all the eight datasets under analysis are included in the Appendix A. Herein, we provide a short overview over all datasets, using aggregated results [4]. The aggregation is possible for DQ and DD measures, while DB measures are calculated only on sensitive attributes, which are different for each dataset. In terms of documentation, we observe a general lack of information (on average, 65% of the information is missing), leaving key aspects such as data composition, collection and processing unknown. Looking at DQ measures, we observe high values for the completeness measures: this reinforces our hypothesis that measuring DQ is necessary but insufficient on its own to highlight emerging data ethics challenges.

5.1 Adult

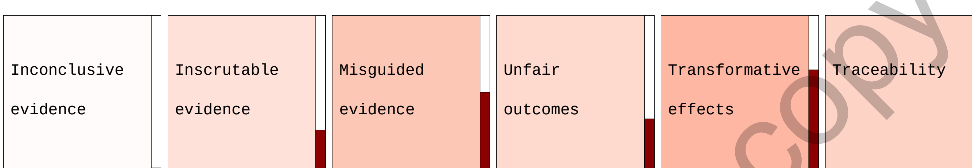


Fig. 1. Ethical Challenge Risk Labels of the Adult dataset.

Adult dataset (Appendix A.1) was constructed to predict an individual's income based on census data. The *Ethical Challenge Risk Labels* reveals that the main risks are related to *transformative effects* (DB) and *misguided evidence* (DQ+DD). The *transformative effects* (DB) ratio risk reveal problems related to DB, since the Simpson (↑) index expose three out of six sensitive features under threshold: *race*, *education*, and *native-country* [5]. This means that these sensitive features deserve special attention when building a model from the data, as their distribution of classes is very imbalanced. On the contrary, *sex* [6] appears to be rather balanced. The second-riskiest challenge is *misguided evidence* (DQ+DD). In terms of DQ, the dataset presents low risks in terms of outliers (*Acc-I-4* (↓)=0,08) and of inconsistency due to duplication of data values (*Con-I-3-DevC* (↓)=0,12). The results of the DD analysis show a lack of relevant information, as only 38% of the requested information is available. The description of the collection processes is very poor, coupled with the data composition. This should alert practitioners to the fact that the data characteristics and processing steps are opaque.

5.2 COMPAS

The COMPAS dataset (Appendix A.2) stems from ProPublica's analysis of the Correctional Offender Management Profiling for Alternative Sanctions commercial tool, used to assess the likelihood that a defendant will reoffend. In this dataset, the risk of *transformative effects* (DB) is over 70%. In fact, the Simpson index exposes three out of five sensitive features as imbalanced: *Language* [7] is the worst one. The rather high value of the *DecileScore* target variable with Gini (↑), Shannon (↑) and Simpson (↑) indices describe a well-balanced situation, although with the least frequent

[4] Figure 10 in the Appendix B integrates what we reported here with two graphs on the summary statistics. These statistics are shown to get an aggregated overview of the datasets included in this research. The empirical study of the fairness datasets, from the perspective of DQ, DD and DB measurements, is beyond the scope of this Experience paper.
[5] Frequencies of classes of *native-country* are:"United-States"=89,59% and other 41 classes below 2%.
[6] Frequencies of classes of *sex* are: "Male"=67%, "Female"=33%.
[7] Frequencies of classes of *Language* are: "English"=99,59%, "Spanish"=0,41%.

class being very rare as pointed by I.I.R. (↑)=0[8]. In order of risk, the second challenge is *misguided evidence* (DQ+DD). In terms of DQ, the dataset has a low risk of containing outliers (*Acc-I-4* (↓)=0,03); there are some null data items (*Com-I-1-DevA* (↑)=0,97) and there are small risks of consistency (*Con-I-2-DevB* (↑)=0,99; *Con-I-3-DevC* (↑)=0,07). In terms of DD[9], there is a general lack of relevant information (*Overall Presence Average* (↑)=0,44), especially in the section on how to (not) use the dataset. This finding echoes wider concerns on misguided use of this dataset [2].

5.3 South German Credit

The South German Credit dataset (Appendix A.3) was constructed with the aim of predicting creditworthiness using 20 variables. In this case, the greater risk is related to *misguided evidence* (DQ+DD), with a risk ratio of 65%. In terms of DQ, the 7% of the numerical data are possible outliers (*Acc-I-4* (↓)=0,07) and the risk of inconsistency due to duplication is moderately low (*Con-I-3-DevC* (↓)=0,10). As far as DD is concerned, this data set is very poorly documented: only a quarter of the relevant information is available. There is very little information on composition, collection processes and uses. The second challenge that presents a higher risk is *transformative effect* (DB), which presents a value similar to *traceability* (DQ+DB+DD). Looking at DB, we can see that *gastarb*[10] (foreign work) is imbalanced, with very low measures in all indexes. On the contrary, *laufkont* (status), *verm* (savings) and*kredit* (credit risk, target variable) are not imbalanced according to any index. *Famges* (marital status and gender) and *beruf* (occupation), are imbalanced only according to the I.I.R. (↑). This indicates a large gap between the most and least frequent classes.

6 CHALLENGES

Herein, we report on the main practical challenges encountered during the research, aiming to bring transparency to this “teaching case”. Since the proposed labels are meant to be informative and not operational, our focus was on the preprocessing part, as the subsequent steps are related to in-process or post-process mitigations. Furthermore, these challenges lay the groundwork for the potential automation, and the consequent integration into the AI pipeline, of the proposed process.

Preprocessing datasets. A significant challenge is the conversion of raw datasets, often in the form of CSV files, into accurately loaded datasets as Pandas dataframes. This data preparation step is a complex and dataset-specific process. A deep understanding of the data structures, formats, encoding and potential issues is essential. This challenge requires tailored strategies, including data cleaning, normalization, and handling of missing values and outliers. The presence of poor documentation often exacerbates the difficulties by leaving critical details unclear.

Adaptability of data balance metrics to different features. Data balance measures are valuable risk indicators for possible unfair outcomes, however their applicability to all attributes is not universal. For example, the analysis of age attributes, produces different results depending on the type of quantisation chosen. This highlights the need to manually identify which columns of a dataset are suitable for the computation of imbalance metrics, challenging scalability and automation.

Finding the complete documentation of the dataset. The process of analysing the completeness of documentation is hampered by the difficulty of obtaining accurate documentation. Sometimes information is scattered across different sources or there is no comprehensive documentation at all. In addition, the lack of a standardized metadata structure, uniformly adopted by repositories,

[8]The least frequent class ("-1") is assigned to 0,07% of the records, compared to the most frequent class ("1"), which is assigned to 30,35% of the records. Moreover, the class "-1" of *DecileScore* target variable corresponds to *RawScore*=1 and *ScoreText*=N/A, i.e. a null value: the coding of the data is anything but clear.

[9]Our analysis focused on the report accompanying the data release [30]: other sources may provide more information.

[10]Frequencies of classes of *gastarb* (Is the debtor a foreign worker?) are: "2"=96% (no), "1"=4% (yes).

makes the task nearly impossible to be automated. Dealing with these discrepancies underlines the complexity of assessing documentation quality, which affects the reliability of subsequent analyses.

7 LIMITATIONS AND FUTURE WORK

We observe some elements of the design and of measurements that could potentially affect the validity of our findings. First, the small number of datasets used in this study may limit the generalizability of our conclusions. However, our primary aim is to prove the feasibility of the proposed approach. Specifically, addressing new data quality challenges with a use case demonstrating the opportunities and limitations of combining different data measurement dimensions to cover a broader range of ethical implications. Applying the proposed approach to synthetic data is a potential avenue for further research to establish its adaptability and scalability.

Secondly, the lack of direct input from domain experts hampers our ability to assess the practical implications of our framework, to validate the proposed schema, and eventually refine it.

The third limitation concerns the lack of exhaustiveness of the measurements dimensions and ethical challenges taken into consideration. The work of Mitchell et al. [34] lays the foundations for extension to numerous data dimensions (adaptable to context and needs) and can be a useful starting point for extending the framework, as well as the very recent ISO standards on data quality for ML, which were just released at the time of finalizing this work. Measuring the dispersion of documentation is also an important avenue to explore. We may investigate a groupwise extensions of quality metrics by slicing the dataset across different categories of protected attributes, potentially making connections between the results within the balance dimension and those within the quality dimension. Moreover, studying the intersection of protected attributes can reveal the unfairness in the outcome [31]. This multifaceted approach could improve our understanding of the data and provide valuable insights into how different sensitive features may affect the overall quality assessment. Future improvements in this direction shall be balanced with the number of measurements to report, to avoid making the reporting sheet difficult to use and to interpret.

8 CONCLUSIONS

The purpose of the study was to expand data quality dimensions to cover a large spectrum of ethical challenges posed by the widespread use of data-driven algorithms in our society. We relied on the knowledge acquired by the authors in their past studies (independently of each other), and combined it in a novel way, to prove the feasibility of the approach and to identify new data quality challenges. We used traditional measures of data quality from the ISO SQuaRE standards in combination with measures of balance and of documentation completeness. We produced an *Extended Data Brief* and a set of *Ethical Challenge Risk Labels* for a selection of popular fairness datasets: the measures identify several detriments to the ethical dimensions under consideration.

The results prove that relying solely on standard quality measures reveals only some faces of the multidimensional ethical implications involved when a dataset is later used as a training source, and that measures of balance and documentation completeness can fill the gap. However, we also observed that their applicability and automatic computation is hampered by a few practical challenges that we reported and discussed. Expansions of the metrics is possible, but a trade-off with ease of use and understandability of the reporting scheme is necessary to preserve the final goal of promoting a more responsible development and distribution of datasets. This will help to make data-driven software applications more trustable, fair and accountable towards the communities of people impacted.

REFERENCES

[1] Jack Bandy and Nicholas Vincent. 2021. Addressing "Documentation Debt" in Machine Learning Research: A Retrospective Datasheet for BookCorpus. https://doi.org/10.48550/arXiv.2105.05241 arXiv:2105.05241 [cs]

[2] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. 2022. It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. https://doi.org/10.48550/arXiv.2106.05498 arXiv:2106.05498 [cs]

[3] Carlo Batini and Monica Scannapieca. 2006. Data Quality Dimensions. In Data Quality. Springer, Berlin, Heidelberg, 19–49. https://link.springer.com/chapter/10.1007/3-540-33173-5_2

[4] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Transactions of the Association for Computational Linguistics 6 (2018), 587–604. https://doi.org/10.1162/tacl_a_00041

[5] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In Proc. of the 2021 ACM Conf. on Fairness, Accountability and Transparency (FAccT '21). ACM, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[6] Elena Beretta, Antonio Santangelo, Bruno Lepri, Antonio Vetrò, and Juan Carlos De Martin. 2019. The Invisible Power of Fairness. How Machine Learning Shapes Democracy. https://doi.org/10.48550/arXiv.1903.09493 arXiv:1903.09493 [cs, stat]

[7] Elena Beretta, Antonio Vetrò, Bruno Lepri, and Juan Carlos De Martin. 2021. Detecting Discriminatory Risk through Data Annotation Based on Bayesian Inferences. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 794–804. https://doi.org/10.1145/3442188.3445940

[8] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency. PMLR, New York, NY, USA, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[9] Stefania Capecchi and Maria Iannario. 2016. Gini Heterogeneity Index for Detecting Uncertainty in Ordinal Data Surveys. METRON 74, 2 (Aug. 2016), 223–232. https://doi.org/10.1007/s40300-016-0088-5

[10] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. 2023. A Comprehensive Empirical Study of Bias Mitigation Methods for Machine Learning Classifiers. https://doi.org/10.48550/arXiv.2207.03277 arXiv:2207.03277 [cs]

[11] Kate Crawford. 2021. The Atlas of AI. Yale University Press, New Haven, CT.

[12] European Commission. 2019. Ethics Guidelines for Trustworthy AI | Shaping Europe's Digital Future. Technical Report. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[13] European Union Agency For Fundamental Rights. 2015. EU Charter of Fundamental Rights - Title III: Quality - Article 21 - Non-discrimination. http://fra.europa.eu/en/eu-charter/article/21-non-discrimination

[14] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Algorithmic Fairness Datasets: The Story so Far. Data Mining and Knowledge Discovery 36, 6 (2022), 2074–2152. https://doi.org/10.1007/s10618-022-00854-z

[15] Donatella Firmani, Letizia Tanca, and Riccardo Torlone. 2019. Ethical Dimensions for Data Quality. Journal of Data and Information Quality 12, 1 (Dec. 2019), 2:1–2:5. https://doi.org/10.1145/3362121

[16] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. https://doi.org/10.2139/ssrn.3518482

[17] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. 2018. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines 28, 4 (Dec. 2018), 689–707. https://doi.org/10.1007/s11023-018-9482-5

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

[19] Haibo He and Edwardo A. Garcia. 2009. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering 21, 9 (Sept. 2009), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

[20] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2022. Gender and Racial Bias in Visual Question Answering Datasets. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 1280–1292. https://doi.org/10.1145/3531146.3533184

[21] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. https://doi.org/10.48550/arXiv.1805.03677 arXiv:1805.03677

[22] ISO. 2008. Software Engineering - Software Product Quality Requirements and Evaluation (SQuaRE) - Data Quality Model (ISO-IEC 25012-2008). https://www.iso.org/standard/35736.html

[23] ISO. 2014. Systems and Software Engineering - Systems and Software Quality Requirements and Evaluation (SQuaRE) - Guide to SQuaRE (ISO-IEC 25000-2014). https://www.iso.org/standard/64764.html

[24] ISO. 2015. Systems and Software Engineering - Systems and Software Quality Requirements and Evaluation (SQuaRE) - Measurement of Data Quality (ISO-IEC 25024-2015). https://www.iso.org/standard/35749.html

[25] Abhinav Jain, Hima Patel, Lokesh Nagalapatti, Nitin Gupta, Sameep Mehta, Shanmukha Guttula, Shashank Mujumdar, Shazia Afzal, Ruhi Sharma Mittal, and Vitobha Munigala. 2020. Overview and Importance of Data Quality for Machine Learning Tasks. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20). Association for Computing Machinery, New York, NY, USA, 3561–3562. https://doi.org/10.1145/3394486.3406477

[26] Eun Seo Jo and Timnit Gebru. 2020. Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 306–316. https://doi.org/10.1145/3351095.3372829

[27] Monique F Kilkenny and Kerin M Robinson. 2018. Data Quality: "Garbage in – Garbage Out". Health Information Management Journal 47, 3 (Sept. 2018), 103–144. https://doi.org/10.1177/1833358318774357

[28] Bernard Koch, Emily Denton, Alex Hanna, and Jacob G. Foster. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. https://doi.org/10.48550/arXiv.2112.01716 arXiv:2112.01716 [cs, stat]

[29] Florian Königstorfer and Stefan Thalmann. 2021. Software Documentation Is Not Enough! Requirements for the Documentation of AI. Digital Policy, Regulation and Governance 23, 5 (2021), 475–488. https://doi.org/10.1108/DPRG-03-2021-0047

[30] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[31] Mariachiara Mecati, Marco Torchiano, Antonio Vetrò, and Juan Carlos de Martin. 2023. Measuring Imbalance on Intersectional Protected Attributes and on Target Variable to Forecast Unfair Classifications. IEEE Access 11 (2023), 26996–27011. https://doi.org/10.1109/ACCESS.2023.3252370

[32] Mariachiara Mecati, Antonio Vetro, and Marco Torchiano. 2021. Detecting Discrimination Risk in Automated Decision-Making Systems with Balance Measures on Input Data. In 2021 IEEE International Conference on Big Data (Big Data). IEEE, Orlando, FL, USA, 4287–4296. https://doi.org/10.1109/BigData52589.2021.9671443

[33] Mariachiara Mecati, Antonio Vetrò, and Marco Torchiano. 2022. Detecting Risk of Biased Output with Balance Measures. Journal of Data and Information Quality 14, 4 (Nov. 2022), 25:1–25:7. https://doi.org/10.1145/3530787

[34] Margaret Mitchell, Alexandra Sasha Luccioni, Nathan Lambert, Marissa Gerchick, Angelina McMillan-Major, Ezinwanne Ozoani, Nazneen Rajani, Tristan Thrush, Yacine Jernite, and Douwe Kiela. 2023. Measuring Data. https://doi.org/10.48550/arXiv.2212.05129 arXiv:2212.05129 [cs]

[35] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In Proc. of the 2019 Conf. on Fairness, Accountability and Transparency (FAT* '19). ACM, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

[36] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The Ethics of Algorithms: Mapping the Debate. Big Data & Society 3, 2 (Dec. 2016), 2053951716679679. https://doi.org/10.1177/2053951716679679

[37] OECD. 2019. Artificial Intelligence in Society. Organisation for Economic Co-operation and Development, Paris. https://www.oecd-ilibrary.org/science-and-technology/artificial-intelligence-in-society_eedfee77-en

[38] Cathy O'Neil. 2016. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group, USA.

[39] Pew Research Center. 2018. Public Attitudes Toward Computer Algorithms. https://www.pewresearch.org/internet/2018/11/16/public-attitudes-toward-computer-algorithms/

[40] Giada Pistilli, Carlos Munoz Ferrandis, Yacine Jernite, and Margaret Mitchell. 2023. Stronger Together: On the Articulation of Ethical Charters, Legal Tools, and Technical Documentation in ML. https://doi.org/10.1145/3593013.3594002 arXiv:2305.18615 [cs]

[41] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 1776–1826. https://doi.org/10.1145/3531146.3533231

[42] Anandhi Ramasamy and Soumitra Chowdhury. 2020. BIG DATA QUALITY DIMENSIONS: A SYSTEMATIC LITERATURE REVIEW. JISTEM - Journal of Information Systems and Technology Management 17 (July 2020), e202017003. https://doi.org/10.4301/S1807-1775202017003

[43] John Richards, David Piorkowski, Michael Hind, Stephanie Houde, and Aleksandra Mojsilović. 2020. A Methodology for Creating AI FactSheets. https://doi.org/10.48550/arXiv.2006.13796 arXiv:2006.13796

[44] Marco Rondina, Antonio Vetrò, and Juan Carlos De Martin. 2023. Completeness of Datasets Documentation on ML/AI Repositories: An Empirical Investigation. In Progress in Artificial Intelligence (Lecture Notes in Computer Science, Vol. 14115), Nuno Moniz, Zita Vale, José Cascalho, Catarina Silva, and Raquel Sebastião (Eds.). Springer Nature Switzerland, Cham, 79–91. https://doi.org/10.1007/978-3-031-49008-8_7
[45] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI. In Proc. of the 2021 CHI Conf. on Hum. Factors in Comput. Syst. (CHI '21). ACM, New York, NY, USA, 1–15. https://doi.org/10.1145/3411764.3445518
[46] Fatimah Sidi, Payam Hassany Shariat Panahy, Lilly Suriani Affendey, Marzanah A. Jabar, Hamidah Ibrahim, and Aida Mustapha. 2012. Data Quality: A Survey of Data Quality Dimensions. In 2012 International Conference on Information Retrieval & Knowledge Management. IEEE, Kuala Lumpur, Malaysia, 300–304. https://doi.org/10.1109/InfRKM.2012.6204995
[47] Antonio Vetrò. 2021. Imbalanced Data as Risk Factor of Discriminating Automated Decisions: A Measurement-Based Approach. JIPITEC 12, 4 (Dec. 2021), 272–288. https://www.jipitec.eu/jipitec/article/view/325
[48] Antonio Vetrò, Marco Torchiano, and Mariachiara Mecati. 2021. A Data Quality Approach to the Identification of Discrimination Risk in Automated Decision Making Systems. Government Information Quarterly 38, 4 (Oct. 2021), 101619. https://doi.org/10.1016/j.giq.2021.101619
[49] Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, HV Jagadish, and Gerome Miklau. 2018. A Nutritional Label for Rankings. In Proceedings of the 2018 International Conference on Management of Data (SIGMOD '18). Association for Computing Machinery, New York, NY, USA, 1773–1776. https://doi.org/10.1145/3183713.3193568
[50] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2021. Fairness in Ranking: A Survey. https://doi.org/10.48550/arXiv.2103.14000 arXiv:2103.14000 [cs]

A EXTENDED DATA BRIEFS

A.1 Adult

Table 4. Application of the framework measures to the Adult dataset. The arrows indicate the best value for each QM (0: ↓, 1: ↑). *: Fields inherited from the Data Brief [14].

Dataset name	Adult		**Date of analysis**	07/28/2023
Description*	This dataset was created as a resource to benchmark the performance of machine learning algorithms on socially relevant data. Each instance is a person who responded to the March 1994 US Current Population Survey, represented along demographic and socio-economic dimensions, with features describing their profession, education, age, sex, race, personal and financial condition. The dataset was extracted from the census database, preprocessed, and donated to UCI Machine Learning Repository in 1996 by Ronny Kohavi and Barry Becker. A binary variable encoding whether respondents' income is above $50,000 was chosen as the target of the prediction task associated with this resource.			
Landing page*	https://archive.ics.uci.edu/ml/datasets/adult			
Sample size*	~50K	**Domain***	economics	
Last update*	1996	**Data specification***	tabular data	
Creator affiliation*	Silicon Graphics Inc.			

Standard data quality (DQ)		**Data documentation** (DD)	
Measure	Value	Measure (Presence Average)	Value (↑)
Acc-I-4 (↓)	0,08	Overall	0,38
Com-I-1-DevA (↑)	1,00	1 Motivation	0,67
Com-I-5 (↑)	1,00	2 Composition	0,21
Con-I-2-DevB (↑)	1,00	3 Collection processes	0,14
Con-I-3-DevC (↓)	0,12	4 Data processing procedures	0,67
Con-I-4-DevD (↑)	1,00	5 Uses	0,80
		6 Maintenance	0,43

Data balance (DB)				
Sensitive Feature	Gini (↑)	Shannon (↑)	Simpson (↑)	I.I.R. (↑)
sex	0,89	0,92	0,79	0,49
race	0,32	0,34	0,09	0,01
education	0,86	0,73	0,28	0,00
marital-status	0,77	0,65	0,32	0,00
native-country	0,20	0,18	0,01	0,00
income	0,73	0,80	0,58	0,32

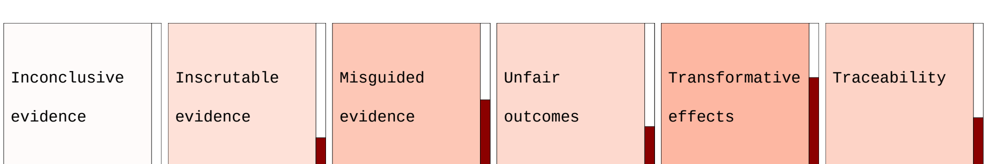


Fig. 2. Ethical Challenge Risk Labels of the Adult dataset.

A.2 COMPAS

Table 5. Application of the framework measures to the COMPAS dataset. The arrows indicate the best value for each QM (0: ↓, 1: ↑). *: Fields inherited from the Data Brief [14].

Dataset name	COMPAS		**Date of analysis**	07/28/2023
Description*	this dataset was created for an external audit of racial biases in the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) risk assessment tool developed by Northpointe (now Equivant), which estimates the likelihood of a defendant becoming a recidivist. Instances represent defendants scored by COMPAS in Broward County, Florida, between 2013–2014, reporting their demographics, criminal record, custody and COMPAS scores. Defendants' public criminal records were obtained from the Broward County Clerk's Office website matching them based on date of birth, first and last names. The dataset was augmented with jail records and COMPAS scores provided by the Broward County Sheriff's Office. Finally, public incarceration records were downloaded from the Florida Department of Corrections website. Instances are associated with two target variables (is recid and is violent recid), indicating whether defendants were booked in jail for a criminal offense (potentially violent) that occurred after their COMPAS screening but within two years.			
Landing page*	https://github.com/propublica/compas-analysis			
Sample size*	~12K	**Domain***	law	
Last update*	2016	**Data specification***	tabular data	
Creator affiliation*	ProPublica			

Standard data quality (DQ)		**Data documentation** (DD)	
Measure	Value	Measure (Presence Average)	Value (↑)
Acc-I-4 (↓)	0,06	Overall	0,44
Com-I-1-DevA (↑)	0,81	1 Motivation	0,67
Com-I-5 (↑)	1,00	2 Composition	0,43
Con-I-2-DevB (↑)	0,99	3 Collection processes	0,29
Con-I-3-DevC (↓)	0,03	4 Data processing procedures	0,67
Con-I-4-DevD (↑)	1,00	5 Uses	0,20
		6 Maintenance	0,57

Data balance (DB)				
Sensitive Feature	Gini (↑)	Shannon (↑)	Simpson (↑)	I.I.R. (↑)
sex	0,62	0,71	0,45	0,24
race	0,73	0,62	0,31	0,00
age cat	0,87	0,89	0,70	0,37
v score text	0,74	0,77	0,49	0,15

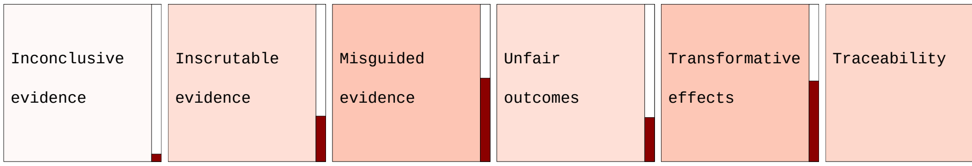


Fig. 3. Ethical Challenge Risk Labels of the COMPAS dataset.

A.3 South German Credit

Table 6. Application of the framework measures to the South German Credit dataset. The arrows indicate the best value for each QM (0: ↓, 1: ↑). *: Fields inherited from the Data Brief [14].

Dataset name	South German Credit	**Date of analysis**	01/23/2023
Description*	The German Credit dataset was created to study the problem of automated credit decisions at a regional Bank in southern Germany. Instances represent loan applicants from 1973 to 1975, who were deemed creditworthy and were granted a loan, bringing about a natural selection bias. The data summarizes their financial situation, credit history and personal situation, including housing and number of liable people. A binary variable encoding whether each loan recipient punctually paid every installment is the target of a classification task. Among covariates, marital status and sex are jointly encoded in a single variable. Many documentation mistakes are present in the UCI entry associated with this resource (UCI Machine Learning Repository, 1994). Due to one of these mistakes, users of this dataset are led to believe that the variable sex can be retrieved from the joint marital status-sex variable, however this is false. A revised version with correct variable encodings, called South German Credit, was donated to UCI Machine Learning Repository (2019) with an accompanying report (Gromping, 2019).		
Landing page*	https://archive.ics.uci.edu/dataset/573/south+german[...]		
Sample size*	~1K	**Domain***	finance
Last update*	2019	**Data specification***	tabular data
Creator affiliation*	Beuth University of Applied Sciences Berlin		

Standard data quality (DQ)		**Data documentation** (DD)	
Measure	Value	Measure (Presence Average)	Value (↑)
Acc-I-4 (↓)	0,07	Overall	0,26
Com-I-1-DevA (↑)	1,00	1 Motivation	1,00
Com-I-5 (↑)	1,00	2 Composition	0,14
Con-I-2-DevB (↑)	1,00	3 Collection processes	0,14
Con-I-3-DevC (↓)	0,10	4 Data processing procedures	0,33
Con-I-4-DevD (↑)	1,00	5 Uses	0,20
		6 Maintenance	0,29

Data balance (DB)				
Sensitive Feature	Gini (↑)	Shannon (↑)	Simpson (↑)	I.I.R. (↑)
gastarb	0,14	0,23	0,08	0,04
laufkont	0,92	0,90	0,75	0,16
famges	0,79	0,77	0,48	0,09
beruf	0,72	0,71	0,39	0,03
verm	0,98	0,97	0,91	0,46
kredit	0,84	0,88	0,72	0,43

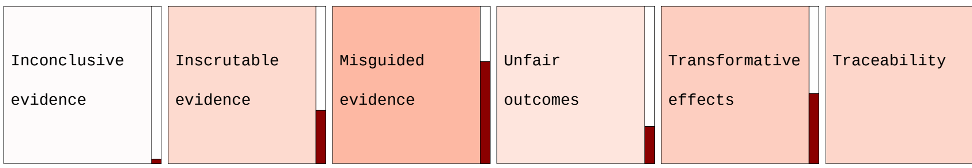


Fig. 4. Ethical Challenge Risk Labels of the South German Credit dataset.

A.4 Communities and Crime

Table 7. Application of the framework measures to the Communities and Crime dataset. The arrows indicate the best value for each QM (0: ↓, 1: ↑). *: Fields inherited from the Data Brief [14].

Dataset name	Communities and Crime	**Date of analysis**	01/23/2023
Description*	This dataset was curated to develop a software tool supporting the work of US police departments. It was especially aimed at identifying similar precincts to exchange best practices and share experiences among departments. The creators were supported by the police departments of Camden (NJ) and Philadelphia (PA). The factors included in the dataset were the ones deemed most important to define similarity of communities from the perspective of law enforcement; they were chosen with the help of law enforcement officials from partner institutions and academics of criminal justice, geography and public policy. The dataset includes socio-economic factors (aggregate data on age, income, immigration, and racial composition) obtained from the 1990 US census, along with information about policing (e.g. number of police cars available) based on the 1990 Law Enforcement Management and Administrative Statistics survey, and crime data derived from the 1995 FBI Uniform Crime Reports. In its released version on UCI, the task associated with the dataset is predicting the total number of violent crimes per 100K population in each community. The most referenced version of this dataset was preprocessed with a normalization step; after receiving multiple requests, the creators also published an unnormalized version.		
Landing page*	https://archive.ics.uci.edu/ml/datasets/communities[...]		
Sample size*	~2K	**Domain***	law
Last update*	2009	**Data specification***	tabular data
Creator affiliation*	La Salle University; Rutgers University		

Standard data quality (DQ)		**Data documentation** (DD)	
Measure	Value	Measure (Presence Average)	Value (↑)
Acc-I-4 (↓)	0,05	Overall	0,33
Com-I-1-DevA (↑)	1,00	1 Motivation	1,00
Com-I-5 (↑)	1,00	2 Composition	0,36
Con-I-2-DevB (↑)	0,97	3 Collection processes	0,00
Con-I-3-DevC (↓)	0,04	4 Data processing procedures	0,33
Con-I-4-DevD (↑)	1,00	5 Uses	0,40
		6 Maintenance	0,29

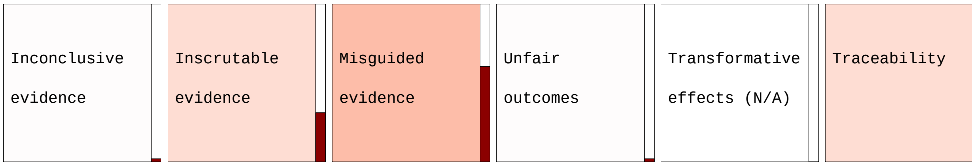


Fig. 5. Ethical Challenge Risk Labels of the Communities and Crime dataset. Since the records of this dataset refer to communities, and not individuals, we decided to exclude the DB measurement. The labels are calculated considering only DQ and DD.

A.5 Bank Marketing

Table 8. Application of the framework measures to the Bank Marketing dataset. The arrows indicate the best value for each QM (0: ↓, 1: ↑). *: Fields inherited from the Data Brief [14].

Dataset name	Bank Marketing	**Date of analysis**	01/18/2023
Description*	Often simply called Bank dataset in the fairness literature, this resource was produced to support a study of success factors in telemarketing of long-term deposits within a Portuguese bank, with data collected over the period 2008–2010. Each data point represents a telemarketing phone call and includes client-specific features (e.g. job, education), features about the marketing phone call (e.g. day of the week and duration) and meaningful environmental features (e.g. euribor). The classification target is a binary variable indicating client subscription to a term deposit.		
Landing page*	https://archive.ics.uci.edu/ml/datasets/Bank+Marketing		
Sample size*	~40K	**Domain***	marketing
Last update*	2012	**Data specification***	tabular data
Creator affiliation*	ISTAR-ISCTE-IUL; University of Minho.		

Standard data quality (DQ)		**Data documentation** (DD)	
Measure	Value	Measure (Presence Average)	Value (↑)
Acc-I-4 (↓)	0,03	Overall	0,26
Com-I-1-DevA (↑)	1,00	1 Motivation	0,67
Com-I-5 (↑)	1,00	2 Composition	0,21
Con-I-2-DevB (↑)	0,98	3 Collection processes	0,29
Con-I-3-DevC (↓)	0,09	4 Data processing procedures	0,33
Con-I-4-DevD (↑)	1,00	5 Uses	0,20
		6 Maintenance	0,14

Data balance (DB)				
Sensitive Feature	Gini (↑)	Shannon (↑)	Simpson (↑)	I.I.R. (↑)
job	0,92	0,85	0,51	0,08
education	0,92	0,86	0,63	0,00
marital	0,81	0,82	0,59	0,19
y	0,40	0,51	0,25	0,13

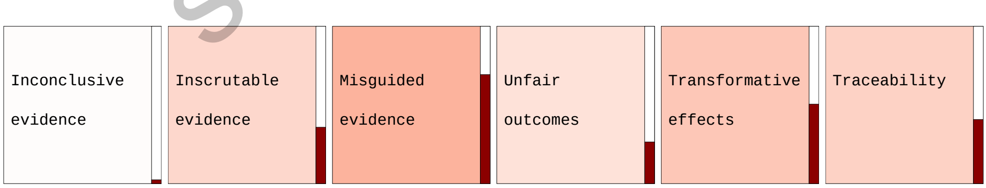


Fig. 6. Ethical Challenge Risk Labels of the Bank Marketing dataset.

A.6 Law School

Table 9. Application of the framework measures to the Law School dataset. The arrows indicate the best value for each QM (0: ↓, 1: ↑). *: Fields inherited from the Data Brief [14].

Dataset name	Law School	**Date of analysis**	03/25/2023
Description*	This dataset was collected to study performance in law school and bar examination of minority examinees in connection with affirmative action programs established after 1967 and subsequent anecdotal reports suggesting low bar passage rates for black examinees. Students, law schools, and state boards of bar examiners contributed to this dataset. The study tracks students who entered law school in fall 1991 through three or more years of law school and up to five administrations of the bar examination. Variables include demographics of candidates (e.g. age, race, sex), their academic performance (undergraduate GPA, law school admission test, and GPA), personal condition (e.g. financial responsibility for others during law school) along with information about law schools and bar exams (e.g. geographical area where it was taken). The associated task in machine learning is prediction of passage of the bar exam.		
Landing page*	https://storage.googleapis.com/lawschool[...]		
Sample size*	~20K	**Domain***	education
Last update*	1998	**Data specification***	tabular data
Creator affiliation*	Law School Admission Council (LSAC)		

Standard data quality (DQ)		**Data documentation** (DD)	
Measure	Value	Measure (Presence Average)	Value (↑)
Acc-I-4 (↓)	0,04	Overall	0,54
Com-I-1-DevA (↑)	0,99	1 Motivation	1,00
Com-I-5 (↑)	1,00	2 Composition	0,79
Con-I-2-DevB (↑)	0,98	3 Collection processes	0,57
Con-I-3-DevC (↓)	0,05	4 Data processing procedures	0,33
Con-I-4-DevD (↑)	1,00	5 Uses	0,20
		6 Maintenance	0,14

Data balance (DB)				
Sensitive Feature	Gini (↑)	Shannon (↑)	Simpson (↑)	I.I.R. (↑)
gender	0,98	0,99	0,97	0,78
race1	0,37	0,41	0,10	0,02
lsat	0,96	0,72	0,18	0,00
ugpa	0,97	0,86	0,55	0,00
pass bar	0,20	0,30	0,11	0,06

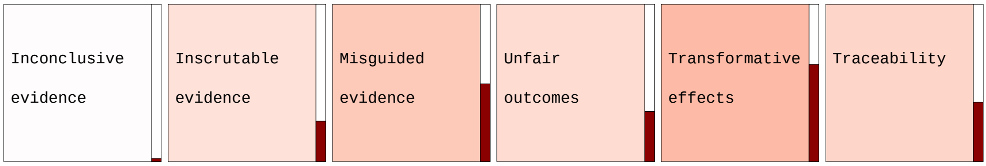


Fig. 7. Ethical Challenge Risk Labels of the Law School dataset.

A.7 MovieLens

Table 10. Application of the framework measures to the MovieLens dataset. The arrows indicate the best value for each QM (0: ↓, 1: ↑). *: Fields inherited from the Data Brief [14].

Dataset name	MovieLens	**Date of analysis**	05/30/2023
Description*	First released in 1998, MovieLens datasets represent user ratings from the movie recommender platform run by the GroupLens research group from the University of Minnesota. While different datasets have been released by GroupLens, in this section we concentrate on MovieLens 1M, the one predominantly used in fairness research. User-system interactions take the form of a quadruple (UserID, MovieID, Rating, Timestamp), with ratings expressed on a 1-5 star scale. The dataset also reports user demographics such as age and gender, which is voluntarily provided by the users.		
Landing page*	https://grouplens.org/datasets/movielens/1m/		
Sample size*	~1M reviews, ~6K users, ~4K movies	**Domain***	information systems,movies
Last update*	2003	**Data specification***	tabular data
Creator affiliation*	University of Minnesota		

Standard data quality (DQ)		**Data documentation** (DD)	
Measure	Value	Measure (Presence Average)	Value (↑)
Acc-I-4 (↓)	0,03	Overall	0,41
Com-I-1-DevA (↑)	1,00	1 Motivation	0,67
Com-I-5 (↑)	1,00	2 Composition	0,29
Con-I-2-DevB (↑)	1,00	3 Collection processes	0,71
Con-I-3-DevC (↓)	0,24	4 Data processing procedures	0,33
Con-I-4-DevD (↑)	1,00	5 Uses	0,20
		6 Maintenance	0,43

Data balance (DB)

Sensitive Feature	Gini (↑)	Shannon (↑)	Simpson (↑)	I.I.R. (↑)
Gender	0,74	0,81	0,59	0,33
Occupation	0,97	0,90	0,60	0,02
Zip-code	1,00	0,93	0,37	0,01

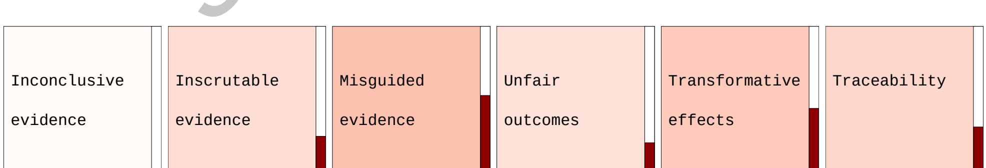


Fig. 8. Ethical Challenge Risk Labels of the MovieLens dataset.

A.8 Credit Card Default

Table 11. Application of the framework measures to the Credit Card Default dataset. The arrows indicate the best value for each QM (0: ↓, 1: ↑). *: Fields inherited from the Data Brief [14].

Dataset name	Credit Card Default		**Date of analysis**	01/18/2023
Description*	This dataset was built to investigate automated mechanisms for credit card default prediction following a wave of defaults in Taiwan connected to patters of card over-issuing and over-usage. The dataset contains payment history of customers of an important Taiwanese bank, from April to October 2005. Demographics, marital status, and education of customers are also provided, along with the amount of credit and a binary variable encoding default on payment, which is the target variable of the associated task.			
Landing page*	https://archive.ics.uci.edu/ml/datasets/default[...]			
Sample size*	~30K credit card holders	**Domain***	finance	
Last update*	2016	**Data specification***	tabular data	
Creator affiliation*	Chung-Hua University;Thompson Rivers University			
Standard data quality (DQ)		**Data documentation** (DD)		
Measure	Value	Measure (Presence Average)		Value (↑)
Acc-I-4 (↓)	0,08	Overall		0,28
Com-I-1-DevA (↑)	1,00	1 Motivation		1,00
Com-I-5 (↑)	1,00	2 Composition		0,14
Con-I-2-DevB (↑)	0,92	3 Collection processes		0,14
Con-I-3-DevC (↓)	0,06	4 Data processing procedures		0,33
Con-I-4-DevD (↑)	1,00	5 Uses		0,40
		6 Maintenance		0,29
Data balance (DB)				
Sensitive Feature	Gini (↑)	Shannon (↑)	Simpson (↑)	I.I.R. (↑)
SEX	0,96	0,97	0,92	0,66
EDUCATION	0,73	0,57	0,28	0,00
MARRIAGE	0,68	0,54	0,35	0,00
default payment next month	0,69	0,76	0,53	0,28

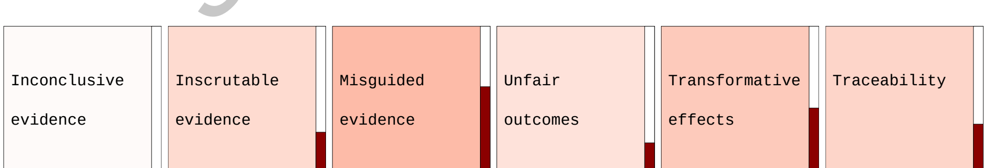
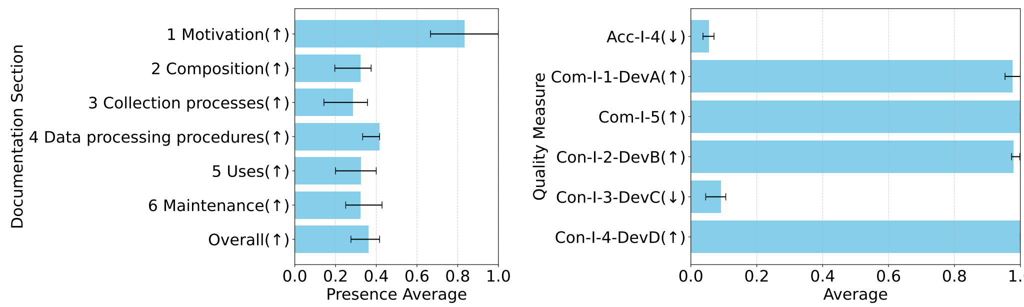


Fig. 9. Ethical Challenge Risk Labels of the Credit Card Default dataset.

B ADDITIONAL FIGURES



(a) Means of the documentation presence averages. (b) Means of the data quality measures.

Fig. 10. Quality Measures results on the selected datasets. The arrows indicate the best value for each QM (0: ↓, 1: ↑). #: This field can have a value greater than 1.

Received 15 November 2023