Data Quality Dimensions for Fair AI*

Camilla Quaresmini[1,*,†], Giuseppe Primiero[2,†]

[1]Department of Electronics, Information and Bioengineering, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milan, Italy

[2]LUCI Lab and PhilTech Research Center, Department of Philosophy, Università degli Studi di Milano, Via Festa del Perdono 7, 20122, Milan, Italy and MIRAI, Srl.

Abstract

Artificial Intelligence (AI) systems are not intrinsically neutral and biases trickle in any type of technological tool. In particular when dealing with people, the impact of AI algorithms' technical errors originating with mislabeled data is undeniable. As they feed wrong and discriminatory classifications, these systems are not systematically guarded against bias. In this article we consider the problem of bias in AI systems from the point of view of data quality dimensions. We highlight the limited model construction of bias mitigation tools based on accuracy strategy, illustrating potential improvements of a specific tool in gender classification errors occurring in two typically difficult contexts: the classification of non-binary individuals, for which the label set becomes incomplete with respect to the dataset; and the classification of transgender individuals, for which the dataset becomes inconsistent with respect to the label set. Using formal methods for reasoning about the behavior of the classification system in presence of a changing world, we propose to reconsider the fairness of the classification task in terms of completeness, consistency, timeliness and reliability, and offer some theoretical results.

Keywords

Bias mitigation, Fairness, Information Quality, Mislabeling, Timeliness

1. Introduction

Machine Learning (ML) models trained on huge amounts of data are intrinsically biased when dealing with people. Common face recognition systems used in surveillance tasks generate false positives labeling innocent people as suspects. Social credit systems link individuals to the state of their social credit, making decisions based on that score. In all of those cases, subjects suffer a credibility deficit due to prejudices related to their social identity [1]: a dark-skinned man could be characterized by a higher risk of recidivism after being arrested; a short-haired skinny young woman – or a long-haired boy with feminine traits – might be the target of transphobic attacks following misgendering. Through the deployment of these technologies, society makes the gap separating rich from poor, cisnormative from non-cisnormative individuals, more constitutive as automatized and standardized.

Already before the explosion of ML algorithms, [2] offered a framework for understanding three categories of bias in computer systems, assuming the absence of bias as necessary to

AEQUITAS 2024: Workshop on Fairness and Bias in AI | co-located with ECAI 2024, Santiago de Compostela, Spain

*Corresponding author.

†These authors contributed equally.

camilla.quaresmini@polimi.it (C. Quaresmini); giuseppe.primiero@unimi.it (G. Primiero)

0000-0002-6474-1284 (C. Quaresmini); 0000-0003-3264-7100 (G. Primiero)

define their quality. Later on, the emergence of contemporary, data-driven AI systems based on learning has significantly worsened the situation, see e.g. [3, 4]. On this basis, the development and deployment of fairer Artificial Intelligence (AI) systems has been increasingly demanded. Such request appears especially relevant in certain application contexts. For example, as examined in [5], face is commonly used as a legitimate mean of gender classification, and this is operazionalized and automatized in technologies such as Automatic Gender Recognition (AGR), which algorithmically derives gender from faces' physical traits to perform classification [6, 7]. This technique relies on the assumption that gender identity can be computationally derived from facial traits. However, a recent study [8] shows that the most famous AGR systems are not able to classify non-binary genders, also performing poorly on transgender individuals. This is due to the fact that AGR incapsulates a binary, cisnormative conception of gender, adopting a male/female scheme which invalidates non-binary identities.

We declare ourselves against the use of gender classification, as considering face as a proxy for detecting gender identity seems to resonate with phrenology and physiognomy, and we believe that the process of automatic gender recognition can easily lead to mismatches between the theoretical understanding of constructs underlying identity and their operationalization [9], especially when it comes to classification of individuals who recognise themselves outside of binarism. However, we note that this kind of classification is already happening [10], spreading with commercial systems offering gender classification as a standard feature, causing a huge impact on the lives of misgendered individuals. Therefore there are contexts in which it is potentially inevitable that classification exists, and in these contexts it must be fairer. This translates into asking whether there is a strategy to ensure that the labels assigned during classification are as less stereotypical and archetypal as possible. While this paper does not investigates the ethical aspects of AGR, we aim at addressing the issues related to the classification strategies to make them fairer, as an initial study to prepare for implementing mitigation strategies.

An important task, common to technology and philosophy, is therefore the identification and verification of criteria that may help developing fairness conditions for AI systems. While a number of techniques are available to mitigate bias, their primary focus on purely statistical analysis to control accuracy across sensitive classes is clearly insufficient to control social discrimination. A different approach is represented by the explicit formulation of ethical principles to be verified across protected attributes, combining statistical measures with logical reasoning, as formally defined in [11, 12, 13, 14, 15] and implemented by the BRIO tool in [16, 17]. In this latter context, an important direction to explore for a refined definition of ethically-laden verification criteria is the study of quality dimensions and associated biases. In the following of this paper, we offer a theoretical contribution in this direction, preparing the ground for a future implementation. We argue that, even if maximizing data quality and fairness simultaneously can be hard as improving one can deteriorate the other [18], the task of bias mitigation tools can be supported by reasoning on quality dimensions that so far have been left ignored. In particular, we offer examples to show how dimensions of consistency, completeness, timeliness and reliability can be used to establish fairer AI classification systems. This research is in line with the quest for integrating useful empirical metrics on fairness in AI with asking key (conceptual) questions, see [19].

The paper is structured as follows. In Section 2 we offer an overview of fainess definitions and bias types relevant for this work. In Section 3 we briefly overview the technical details of

Table 1
Data and Label Bias.

Bias type	Definition	Literature
	Data Bias	
Behavioral bias	User's behavior can be different across contexts	[38]
Exclusion bias	Systematic exclusion of some data	[39]
Historical bias	Cultural prejudices are included into systematic processes	[40]
Time interval bias	Data collection in a too limited time range	[41]
	Label Bias	
Chronological bias	Distortion due to temporal changes in the world which data are supposed to represent	[39]
Historical bias	Cultural prejudices are included into systematic processes	[40]
Misclassification bias	Data points are assigned to incorrect categories	[42]

a particular bias mitigation tool to illustrate what we consider essential limitations of purely statistical analyses. In Section 4 we introduce data quality dimensions arguing for reconsidering their relevance in the task of evaluating the fairness of classification systems, presenting two examples to justify this requirement. In Section 5 we propose a definition of fair AI classification that includes such dimensions and formulate some theoretical results. Section 6 concludes the work illustrating future research lines.

2. Fairness and Bias in ML

Despite a unique definition missing in the literature [2, 3, 20, 21, 22, 23, 24, 25], fairness is often presented as corresponding to the avoidance of bias [26]. This can be formulated at two distinct levels: first, identifying and correcting problems in datasets [27, 28, 29, 30, 31, 32], as a model trained with a mislabeled dataset will provide biased outputs; second, correcting the algorithms [21, 33], as even in the design of algorithms biases can emerge [34]. In the present section we are interested in considering datasets and their labels. Indeed, bias may also affect the label set [35, 36]. Accordingly, we talk about *label quality bias* when errors hit the quality of labels. As shown in [37], the most well-known AI datasets are full of labeling errors. A crucial task is therefore the development of conceptual strategies and technical tools to mitigate bias emergence in both data and label sets.

A variety of approaches and contributions is available in the literature focusing on identifying bias in datasets and labels. Here we list the types of bias which are relevant to the present work, see Table 2. Albeit not exhaustive, these lists of biases represent a good starting point to investigate quality dimensions required to address them. We now analyze a common mitigation strategy used by existing tools addressing the issue of bias in data, showing their limitations. We then study the bias in the classification algorithm (i.e., bias in labels) of the mitigation tool.

Table 2
Symbols used in the present work.

t_n	Time index
$\mathscr{T} := \{t_1, \dots, t_n\}$	Time frame
d	Generic datapoint
i, j, l	Data Labels
y^*	Discrete random variable correctly labeled
$\tilde{y}$	Discrete random variable wrongly labeled
$[m]$	The set of unique class labels
$y* \to \tilde{y}$	A mapping between variables
$p_{\mathscr{T}}[(\tilde{y} = i)_{t_n} \mid (y* = j)_{t_{n-m}}]$	The probability of label i being wrong at time t_n, given that label j was correct at time t_{n-m}
$C_{\tilde{y},y^*}[i, j, \mathscr{T}]$	Temporal confident joint, where the correct label can change from i to j in time frame $\mathscr{T}$
$C_{\tilde{y},y*}[i, \mathscr{T}]$	Temporal confident joint, where the correctness of the same fixed label i can change in time frame $\mathscr{T}$
ε	Change rate
$\hat{p}'(\tilde{y}; x_i; \theta)$	Predicted probability of label $\tilde{y}$ for variable x_i and model parameters θ
L	Label set
X	AI system
$L_{t_1} := \{l_1, \dots, l_n\}$	Partition of the label set
P	Population of interest
p	An element from P
$d(X)_{\mathscr{T}}$	A datapoint in system X over time frame $\mathscr{T}$
$y^*(d)$	A correct label for the datapoint d
π	Threshold variable

3. Mitigating Bias

A *bias mitigation algorithm* is a procedure for reducing unwanted bias in training datasets or models, with the aim to improve the fairness metrics. Those algorithms can be classified into three categories [43]: pre-processing, when the training data is modified; in-processing, when the learning algorithm is modified; post-processing, when the predictions are modified.

Several tools are available to audit and mitigate biases in datasets, thereby attempting to implement diversity and to reach fairness. Among the most common are AIF360 [22], Aequitas [44] and Cleanlab [45]. Recently a post-hoc evaluation model for bias mitigation has been proposed by the tool BRIO [16, 17]. In this article, we consider Cleanlab as a testbed, illustrating below in Section 4 its limitations in view of data quality dimensions. Instead, we propose a theoretical frame for the resolution of such limitation in Section 5, further illustrating the possibility to implement the present analysis in the tool BRIO. For an overview of the symbols used from now on, see Table 2.

Cleanlab is a framework to find label errors in datasets. It uses Confident Learning (CL), an approach which focuses on label quality with the aim to address uncertainty in dataset labels using three principles: counting examples that are likely to belong to another class using the confident joint and probabilistic thresholds to find label errors and to estimate noise;

pruning noisy data; and ranking examples to train with confidence on clean data. The three approaches are combined by an initial assumption of a class-conditional noise process, to directly estimate the joint distribution between noisy given labels and uncorrupted unknown ones. For every class, the algorithm learns the probability of it being mislabeled as any other class. This assumption may have exceptions but it is considered reasonable. For example, a "cat" is more likely to be mislabeled as "tiger" than as "airplane". This assumption is provided by the classification noise process (CNP, [46]), which leads to the conclusion that the label noise only depends on the latent true class, not on the data. CL [45] exactly finds label errors in datasets by estimating the joint distribution of noisy and true labels. The idea is that when the predicted probability of an example is greater than a threshold per class, we confidently consider that example as actually belonging to the class of that threshold, where the thresholds for each class are the average predicted probability of examples in that class. Given $\tilde{y} \in [m]$ takes an observed, noisy label (potentially flipped to an incorrect class); and $y* \in [m]$ takes the unknown (latent), true, uncorrupted label (latent true label), CL assumes that for every example it exists a correct label $y*$ and defines a class-conditional noise process mapping $y* \rightarrow \tilde{y}$, such that every label in class $j \in [m]$ may be independently mislabeled as class $i \in [m]$, with probability $p(\tilde{y} = i \mid y* = j)$. So, maps are associations of data to wrong labels. Then CL estimates $p(\tilde{y} \mid y*)$ and $p(y*)$ jointly, evaluating the joint distribution of label noise $p(\tilde{y}, y*)$ between noisy given labels and uncorrupted unknown labels. CL aims to estimate every $p(\tilde{y}, y*)$ as a matrix $Q_{\tilde{y},y*}$ to find all mislabeled examples x in dataset X, where $y* \neq \tilde{y}$. Given as inputs the out-of-sample predicted probabilities $\hat{P}_{k,i}$ and the vector of noisy labels $\tilde{y}_k$, the procedure is divided into three steps: estimation of $\hat{Q}_{\tilde{y},y*}$ to characterize class-conditional label noise, filtering of noisy examples, training with the errors found.

To estimate $\hat{Q}_{\tilde{y},y*}$ i.e. the joint distribution of noisy labels $\tilde{y}$ and true labels $y*$, CL counts examples that may belong to another class using a statistical data structure named confident joint $C_{\tilde{y},y*}$, formally defined as follows

$$C_{\tilde{y},y*}[i][j] \ :=| \ \hat{X}_{\tilde{y}=i,y*=j} \ | \tag{1}$$

In other words, the confident joint estimates the set $X_{\tilde{y}=i,y*=j}$ of examples with noisy label i which actually have true label j by making a partition of the dataset X into bins $\hat{X}_{\tilde{y}=i,y*=j}$, namely the set of examples labeled $\tilde{y} = i$ with *large enough* expected probability $\hat{p}(\tilde{y} = j; x, \theta)$ to belong to class $y* = j$, determined by a per-class threshold t_j, where θ is the model.

This kind of tools are extremely useful in estimating label error probabilities. However they have some limitations, and it is easy to formulate examples for which their strategy seems unsound. A first problem arises from the initial assumption of the categoricity of data. Take for example the case of gender labeling of facial images, which is typically binary (i.e. with values male, female). For each datapoint, a classification algorithm calculates the projected probability that an image is assigned to the respective label. Consider though two very noisy cases: images of non-binary individuals; images of transgender individuals. In the former case, the label set becomes incomplete with respect to the dataset; in the second case, the dataset is inconsistent with respect to the label set. Hence, there can be datapoints that have either 1) none of the available labels as the correct one, or 2) at different times they can be under different labels. By definition, if we have disjoint labels there can be high accuracy but only

on those datapoints which identify themselves in the disjointed categories. In situations like these, it appears that the dimension of accuracy alone does no longer satisfy the correctness of the classification algorithm. In terms of quality dimensions, the possibility of an uncategorical datapoint or that of a moving datapoint is no longer only an accuracy problem. Hence, the identification of other data quality dimensions to be implemented in tools for bias mitigation may help achieve more fairness in the classification task. In the next section we suggest an improvement of the classification strategy by adding dimensions that should be considered when evaluating the fairness of the classification itself.

4. Extending Data Dimensions for Fair AI

In the literature, data quality dimensions are defined both informally and qualitatively. Metrics can be associated as indicators of the dimension's quality. However, there is no single and objective vision of data quality dimensions, nor a universal definition for each dimension. This is because often dimensions escape or exceed a formal definition. The cause of the large amount of dimensions [47, 48] also lies in the fact that data aim to represent all spatial, temporal and social phenomena of the real world [49]. Furthermore, they are constantly evolving in response to continuous development of new data-driven technologies.

For the purposes of our analysis, we focus on the following basic set of data quality dimensions which is the focus of the majority of authors in the literature [50, 51]:

- *Accuracy*, i.e. the closeness between a value v and a value v', where the latter is the correct representation of the real-life phenomenon that v aims to represent [47];
- *Completeness*, i.e. the level at which data have the sufficient breadth, depth, and scope for their task [48, 52, 47];
- *Consistency*, i.e. the coherence dimension: it amounts to check whether or not the semantic rules defined on a set of data elements have been respected [47];
- *Timeliness*, the data freshness over time for a specific task [53, 54].

We thus indicate them as potential candidates to be implemented in the context of bias mitigation strategies. In particular, we argue that, as data are characterized by evolution over time, the timeliness dimension [47] can be taken as basis for other categories of data quality. We aim at suggesting improvements on errors identification in the classification of datapoints, using the gender attribute as an illustrative case. We thus suggest the extension of classification with dimensions of completeness, consistency and timeliness and then return to Cleanlab to illustrate how this extension could be practically implemented.

4.1. Incomplete Label Set and Inconsistent Labeling

Consider the first example of a datapoint which represents a non-binary individual. This kind of identity is rarely considered in technology [55]. Non-binary identities do not recognize themselves within the binary approach characteristic of classification systems. As such, individual identity is not correctly recognized by the classification system, highlighting the insufficiency of the model which flattens the gender identity umbrella on the two options of male/female.

The conceptual solution would be to simply assume the label set as incomplete. This means that the bias origin is in the pre-processing phase, and a possible strategy is to extend the partition of the labels adding categories as appropriate, e.g. "non binary". The problem is here reduced to the consideration of the completeness of the label set. [8] can be considered a first attempt in this direction.

Consider now a transgender datapoint whose identity shifts over time, being a fluid datapoint by definition. Currently AI systems operationalize gender in a way which is completely trans-exclusive, see e.g. [7, 6]. However, identity is not static: it may move with respect to the labels we have, leading the datapoint to be configured in a label or in a different one during a selected time range. In this case, any extension of the label set is misleading, or at least insufficient. Here we cannot just add more categories, but we have to find a logical solution to changing the label of the same datapoint at different timepoints.

4.2. Enter Time

The two problems above can be formulated adding to completeness and consistency the dimension of temporality. Thus, an important starting point is represented by adding the dimension of timeliness, which concerns the degree to which data represent reality within a certain defined time range for a given population.

We suggest here considering the labeling task within a given time frame, whose length depends on the dataset and the classification task over the pairing of datapoints to labels, to measure a probability of a label-change over time. Intuitively, if the analysis is performed less than a certain number of timestamps away from the last data labeling, then we consider the labeling still valid. Otherwise, a new analysis with respect to both completeness of the dataset and label set must be performed. Technically, this means associating temporal parameters to labels and to compute the probability that a given label might change over the given time frame. The probability of a label being correct (its accuracy) decreases within the respective temporal window. In particular, reasoning on the temporal evolution of the dataset could allow us to model the evolution of the label partitions. Two fundamental theses are suggested for evaluation: the correctness of the task does no longer assume static completeness of the label set, i.e. given the label set is complete at time t_n, it can be incomplete at time t_{n+m}; the labeling does no longer assume static perseverance of the labels, that is, given a label i that is correct at a time t_n for a datapoint d, it could be incorrect at a later time, and conversely if it is incorrect it could become correct.

4.3. Back to Cleanlab

Considering a possible implementation in Cleanlab able to account for such differences implies renouncing the starting assumption on the categoricity of the data. Instead, assume that the probability of assigning a label may change over time. This can be formulated in two distinct ways. First, the probability value of a given label i being wrong, given a label j is correct (their distance) may change over time. The task is now to give a mapping of all the label-variable pairs, i.e. given a mapping $y* \to \tilde{y}$ between variables, where $y*$ is the correct label and $\tilde{y}$ the wrong one, compute the probability over the time frame $\mathcal{T} := \{t_1, \dots, t_n\}$

$$p_{\mathcal{T}}[(\tilde{y}=i)_{t_n} \mid (y*=j)_{t_{n-m}}] \tag{2}$$

such that label i is wrong at time t_n, given that label j was correct at time t_{n-m}. This probability can increase or decrease, depending on the dataset and on the label set. For the definition of the confident joint, this means taking the evaluation of all the elements that have an incorrect label i when their correct label is j, and then associate the wrong label to a time t_n and the correct label to a previous time. This estimate must be made on all time points, so for every $m < n$. Given a timepoint n at which the label is wrong, the estimate on all pairs of probabilities for that point with a previous point in which another label can be correct has to be computed

$$C_{\tilde{y},y*}[i,j,\mathcal{T}] := \sum_{1\leq m<n\in\mathcal{T}}^{n\in\mathcal{T}} \mid \hat{X}_{\tilde{y}=i_{t_n},y*=j_{t_{n-m}}} \mid \tag{3}$$

Second, given a mapping $y* \to \tilde{y}$ between variables, where $y*$ is the correct label and $\tilde{y}$ the wrong one, what is the probability

$$p_{\mathcal{T}}[(\tilde{y}=i)_{t_n} \mid (y*=i)_{t_{n-m}}] \tag{4}$$

such that label i is wrong at time t_n, given that the same label i was correct at time t_{n-m}? In this case, the same label is fixed and the probability that it becomes incorrect can be calculated. The definition of confident joint thus becomes

$$C_{\tilde{y},y*}[i,\mathcal{T}] := \sum_{1\leq m<n\in\mathcal{T}}^{n\in\mathcal{T}} \mid \hat{X}_{\tilde{y}=i_{t_n},y*=i_{t_{n-m}}} \mid \tag{5}$$

To illustrate the point we consider a toy example. Compute

$$p(\tilde{y}=i \mid y*=j) = \frac{p(y*=j \mid \tilde{y}=i)\cdot p(\tilde{y}=i)}{p(y*=j)} = \frac{\frac{[p(y*=j\wedge\tilde{y}=i)]}{p(\tilde{y}=i)}\cdot p(\tilde{y}=i)}{p(y*=j)} \tag{6}$$

i.e. the error rate of $y* = \textit{male}$ has to be determined. First, a confusion matrix is constructed to analyze errors. Suppose to have a dataset of 10 datapoints, see Figure 1. From the matrix, $p(y*=j) = 5/10$ and $p(\tilde{y}=i) = 4/10$. So there are 5 women, of which 2 are incorrectly labeled "male" and 3 are correctly labeled "female", and 5 men of which 1 is incorrectly labeled "female" and 4 are correctly labeled "male". Replacing the values in Equation 6, $p(\tilde{y}=i \mid y*=j) = 0.2$. The obtained value represents the error rate of the "male" label, i.e. the probability of a male datapoint being labeled "female". Looking at the diagonals, the true positive rate TPR = 70% and the false positive rate FPR = 30%.

Consider now the same dataset at a later time t_{n+m}, see Figure 2. The labels might have changed. From the matrix, $p(y*=j) = 5/10$ and that $p(\tilde{y}=i) = 5/10$. Now there are 5 women, of which 3 are incorrectly labeled "male" and 2 are correctly labeled "female", and 5 men of which 3 are incorrectly labeled"female" and 2 are correctly labeled "male". Replacing again the values in 6, $p'(\tilde{y}=i \mid y*=j) = 0.6$. In this case the true positive rate TPR = 40% and the false positive rate FPR = 60%.

		Actual	
		$y* = male$	$y* = female$
Predicted	$\tilde{y} = male$	4	2
	$\tilde{y} = female$	1	3

Figure 1: Confusion matrix at time n.

		Actual	
		$y* = male$	$y* = female$
Predicted	$\tilde{y} = male$	2	3
	$\tilde{y} = female$	3	2

Figure 2: Confusion matrix at time $n + m$.

To understand how the error rate changes, the difference between the two matrices has to be considered. Thus, the change rate can be computed as $\varepsilon = \hat{p'}(\tilde{y}; x_i; \theta) - \hat{p}(\tilde{y}; x_i; \theta) = 0.4$.

Now $p_{\mathcal{T}}[(\tilde{y} = i)_{t_n} \mid (y* = j)_{t_{n-m}}]$ can be written as $p_{\mathcal{T}}[(\tilde{y} = i)_{t_{n+m}} \mid (y* = j)_{t_n}]$. Thus, at a time t_n we have $p_{t_n}(y* = j) = 1 - p(\tilde{y} = i)_{t_n})$. At a subsequent time t_{n+m} we have $p_{t_{n+m}}(y* = j) = 1 - p(\tilde{y} = i)_{t_{n+m}}$. Equation 6 can be computed with respect to time as

$$p_{\mathcal{T}}[(\tilde{y} = i)_{t_{n+m}} \mid (y* = j)_{t_n}] = \frac{p[(y* = j) \mid (\tilde{y} = i)]_{t_{n+m}} \cdot [p(\tilde{y} = i)t_n \pm \varepsilon]}{p(y* = j)_{t_n}} = 0.288 \tag{7}$$

This value represents the (highest) probability that a given label is wrong at a given time, provided it was correct at some previous time. Indirectly, this also expresses the probability that the labeling set is applied to a dataset containing a point for which the labeling becomes inconsistent over time.

5. Temporal-based Fairness in AI

We have argued that a more general discussion on the data dimensions to be adopted in bias mitigation tools is needed, and in particular that the dimension of timeliness is crucial. In this section we summarise our proposal and offer non-exhaustive criteria for fairness in AI based on such temporal approach along with some basic theoretical results.

The first metric that has been addressed in this work is completeness as applied to the label set. In a world where gender classification is actually changing, the present strategy includes the completeness dimension in the quality assessment, verifying that the label set is complete with respect to the ontology of the world at the time this assessment is made. The solution here is to extend the label set as desired adding new labels for the classification task, as already suggested in [8]. Additionally, we suggest an explicit temporal parametrization: completeness can be considered as a relationship between a label set and an individual p belonging to a certain population P, where p is any domain item that enters P at a time t. We must ensure that a correct label l exists for each datapoint in the dataset at each time.

Definition 1 (Completeness of a label set)**.** *A label set L for a classification algorithm in a AI system X is considered complete over a time frame $\mathcal{T} : \{t_1, \ldots, t_n\}$ denoted as $Compl_{\mathcal{T}}(L(X))$ iff given two partitions $L_{t_1} := \{l_1, \ldots, l_n\}$ and $L_{t_n} := \{l'_1, \ldots, l'_n\}$, where possibly $L_{t_1} \cap L_{t_n} \neq \emptyset$ for all $(p \in P)_{\mathcal{T}}$ s.t. $p \in d(X)_{\mathcal{T}}$ there is $l \in L_{t_1} \cup L_{t_n}$ s.t. $y^*(d) = l$.*

In other words, the completeness of a dataset over a time frame is granted if for every datapoint representing an element in the population of interest there exists at any two possibly consecutive points in time a correct label for it.

Next, we considered consistency of the label set with respect to datapoints possibly shifting in categorization. The method here again is to reduce consistency to timeliness. We suggest to compute the probability of an inconsistency arising from a correct label change. Accuracy, albeit the most used metric for evaluating classification models' performances due to its easy calculability and interpretation, is reductive, trivial and incorrect in some contexts. For example, if the distribution of the class is distorted, accuracy is no longer a useful, nor a relevant metric. Even worse, sometimes greater accuracy leads to greater unfairness [56]: some labels like race or gender may allow models to be more predictive, although it seems to be often controversial to use such categories to increase predictive performance. We have suggested to consider temporal accuracy [57] as a function of the error rate over time.

The ability to compute the variance in the error rate across time is functional to determine the reliability of AI systems. This metric is linked to the notion of accuracy, as it is considered as a measure of data correctness, see [47]. In [48] and [57] reliability is even contained in the definition of accuracy itself: data must be reliable to satisfy the accuracy dimension. Overall, it seems that reliability is not actually controlled beyond physical reliability, as in the literature on data quality there is no formal definition to compute it. However, following [58] the previously provided temporal approach is again useful: evaluating reliability is based on the revisions which show how close the initial estimate of accuracy is to the following ones. In this sense, reliability can be reduced to accuracy over time in terms of a threshold on the error rate:

Definition 2 (Reliability of a classification algorithm)**.** *A classification algorithm in a AI system X is considered reliable over a time frame $\mathcal{T} := \{t_1, \dots, t_n\}$ denoted as $Rel_{\mathcal{T}}(X)$ iff $\varepsilon_{\mathcal{T}}(X) < \pi$, for some safe value π.*

The change rate ε we have computed shows how much the system's accuracy deteriorates. If it exceeds a fixed safe value π, the system is no longer accurate. Plain accuracy is the numerical measure at some time $t \in \mathcal{T} := \{t_1, \dots, t_n\}$. If this value does not deteriorate over a certain fixed threshold, the system is considered reliable, and therefore accurate with respect to time.

The two previous definitions offer non-exhaustive criteria for the identification of fair AI systems:

Definition 3 (Fairness for AI classification systems)**.** *$Fair_{\mathcal{T}}(X)$ only if $Rel_{\mathcal{T}}(X)$ and $Compl_{\mathcal{T}}(L(X))$.*

Hence we claim that fairness requires the system's ability to give reliable and correct outcomes over time. While we do not consider these properties sufficient, we believe they are necessary. On this basis, we can formulate two immediate theoretical results:

Theorem 1. *Given a label set L complete at time t, a classification algorithm guarantees a fair classification at time $t' > t$ if and only if the change rate determined with respect to L is $\epsilon < \pi$.*

Proof. Assume $Compl_t(L(X))$, then for $Fair_{t'}(X)$ we need to show $Rel_{t'}(X)$ for $t' > t \in \mathcal{T}$. Assume $\epsilon > \pi$, then by Definition 2 reliability is not satisfied; hence, if $Rel_{\mathcal{T}}(X)$, it must be the case that $\epsilon < \pi$. □

Theorem 2. *Given a fixed change rate $\epsilon < \pi$, a classification algorithm with fair behaviour at time t remains fair at time $t' > t$ if and only if the change to make the label set complete at time t' does not exceed an ϵ' such that $\epsilon + \epsilon' > \pi$.*

Proof. Consider $Fair_t(X)$ with change rate $0 < \pi$ as a base case, then by Definition 3 $Rel_t(X)$ and $Compl_t(L(X))$. Now consider $t' > t$ and a required change ϵ' in $Compl_t(L(X))$ such that $Rel_{t'}(X)$ holds. This obviously holds only if $0 + \epsilon' < \pi$. Generalize for any $\epsilon > 0$. □

Note that in these results the value of ϵ, respectively ϵ', is a proxy for how much the world has changed at t' with respect to $Compl_t(L(X))$.

In the context of an incomplete label set, a detected label bias can originate from an exclusion bias in data, which can also result from a time interval bias. In the case of label-changing datapoints a chronological bias occurs. Then, misclassification bias can be reduced to the two previous types. In the context of use, emergent bias can arise as a result of changes in societies and cultures. It might appear in data as chronological, historical or behavioral bias. Here, a different value bias occurs for example when the users are different from the assumed ones during the system's development. This is the case of ontology switching, to which a label set must adapt. These types of bias can be mitigated by implementing the proposed framework. The tool BRIO [16, 17] works as a post-hoc model evaluation, taking in input the test dataset of the model under investigation and its output. The tool allows to investigate behavioural differences of the model both with respect to an internal analysis on the classes of interest, and externally with respect to chosen reference metrics. Morever, it allows to measure bias amplification comparing the bias present in the dataset and how that manifests itself in the output. While the present work does not aim at offering a full implementation of our theoretical analysis for the BRIO tool, some remarks are appropriate. The time-based analysis of completeness and reliability offered in Definitions 1 and 2, in turn grounding a notion of fairness in Definition 3 are easily implementable in BRIO: both completeness and reliability require the definition of a timeframe to check respectively that any given datapoint of interest is matched against a desirable label and that the overall change rate of error for one or more classes of interest does not surpass a certain threshold. Both features rely on the user for the identification of the desirable label for any datapoint and for the admissible distance.

6. Conclusion

We presented some recommendations for AI systems design, focusing on timeliness as a founding dimension for developing fairer and more inclusive classification tools. Despite the crucial importance of accuracy as shown by significant works such as [4] and [59], the problem of unfairness in AI systems is much broader and more foundational. This can be expressed in terms of data quality: AI systems are limited in that they maximize accuracy, and even if systems become statistically accurate some problems remain unsolved. This is exemplified by the case of binary gender labeling, which leads to inaccurate simplistic classifications [60]. Furthermore, as the work of classification is always a reflection of culture, the completeness of the label set and the (constrained) consistency of labeling have an epistemological value: constructing

AIs requires us to understand society, and society reflects an ontology of individuals. For this reason, misgendering is first of all an ontological error [6].

We suggested that timeliness is a crucial dimension for the definition of gender identity. If we are ready to consider gender as a property that shifts over time [61], and which can also be declined in the plural, as an individual may identify under more than one - not mutually exclusive - labels, then a change of paradigm is required. Design limitations such as binarism and staticity invalidate identities which do not fit into this paradigm. They must be addressed if fairer classifications and more inclusive models of gender are to be designed.

Further work in this direction includes: an implementation and empirical validation of the proposed model through the BRIO tool; and the design of an extension to compute the probability of incorrect labels becoming correct over time, i.e. the dual case of what presently addressed.

Acknowledgments

This research has been partially funded by the Projects: PRIN2020 BRIO (2020SSKZ7R), PRIN2022 SMARTEST (20223E8Y4X), "Departments of Excellence 2023-2027" of the Department of Philosophy "Piero Martinetti" of the University of Milan, all awarded by the Italian Ministry of University and Research (MUR); and MUSA – Multilayered Urban Sustainability Action, funded by the European Union – NextGenerationEU, under the National Recovery and Resilience Plan (NRRP) Mission 4 Component 2 Investment Line 1.5: Strenghtening of research structures and creation of R&D "innovation ecosystems", set up of "territorial leaders in R&D".

References

[1] M. Fricker, Epistemic Injustice: Power and the Ethics of Knowing, New York: Oxford University Press, 2007.

[2] B. Friedman, H. Nissenbaum, Bias in computer systems, ACM Trans. Inf. Syst. 14 (1996) 330–347.

[3] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, CoRR abs/1908.09635 (2019). URL: http://arxiv.org/abs/1908.09635. arXiv:1908.09635.

[4] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA, volume 81, PMLR, 2018, pp. 77–91.

[5] A. Hanna, M. Pape, M. K. Scheuerman, Auto-essentialization: Gender in automated facial analysis as extended colonial project, Big Data and Society 8 (2021). doi:10.1177/20539517211053712.

[6] O. Keyes, The misgendering machines: Trans/HCI implications of automatic gender recognition, Proc. ACM Hum.-Comput. Interact. 2 (2018). URL: https://doi.org/10.1145/3274357. doi:10.1145/3274357.

[7] F. Hamidi, M. K. Scheuerman, S. M. Branham, Gender recognition or gender reductionism? the social implications of embedded gender recognition systems, in: Proceedings of the

2018 CHI Conference on Human Factors in Computing Systems, CHI '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 1–13. URL: https://doi.org/10.1145/3173574.3173582. doi:10.1145/3173574.3173582.

[8] M. K. Scheuerman, J. Paul, J. Brubaker, How computers see gender: An evaluation of gender classification in commercial facial analysis services, Proceedings of the ACM on Human-Computer Interaction 3 (2019) 1–33. doi:10.1145/3359246.

[9] A. Z. Jacobs, H. M. Wallach, Measurement and fairness, CoRR abs/1912.05511 (2019). URL: http://arxiv.org/abs/1912.05511. arXiv:1912.05511.

[10] A. Ramon, G. Olaoye, A. Luz, Machine learning algorithms for gender prediction (2024).

[11] F. A. D'Asaro, G. Primiero, Probabilistic typed natural deduction for trustworthy computations, in: D. Wang, R. Falcone, J. Zhang (Eds.), Proceedings of the 22nd International Workshop on Trust in Agent Societies (TRUST 2021) Co-located with the 20th International Conferences on Autonomous Agents and Multiagent Systems (AAMAS 2021), London, UK, May 3-7, 2021, volume 3022 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-3022/paper3.pdf.

[12] G. Primiero, F. A. D'Asaro, Proof-checking bias in labeling methods, in: G. Boella, F. A. D'Asaro, A. Dyoub, G. Primiero (Eds.), Proceedings of 1st Workshop on Bias, Ethical AI, Explainability and the Role of Logic and Logic Programming (BEWARE 2022) co-located with the 21th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2022), Udine, Italy, December 2, 2022, volume 3319 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 9–19. URL: https://ceur-ws.org/Vol-3319/paper1.pdf.

[13] A. Termine, G. Primiero, F. A. D'Asaro, Modelling accuracy and trustworthiness of explaining agents, in: S. Ghosh, T. Icard (Eds.), Logic, Rationality, and Interaction - 8th International Workshop, LORI 2021, Xi'ian, China, October 16-18, 2021, Proceedings, volume 13039 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 232–245. URL: https://doi.org/10.1007/978-3-030-88708-7_19. doi:10.1007/978-3-030-88708-7_19.

[14] F. A. D'Asaro, F. Genco, G. Primiero, Checking trustworthiness of probabilistic computations in a typed natural deduction system, 2024. arXiv:2206.12934.

[15] E. Kubyshkina, G. Primiero, A possible worlds semantics for trustworthy non-deterministic computations, International Journal of Approximate Reasoning (2024) 109212. URL: https://www.sciencedirect.com/science/article/pii/S0888613X24000999. doi:https://doi.org/10.1016/j.ijar.2024.109212.

[16] G. Coraglia, F. A. D'Asaro, F. A. Genco, D. Giannuzzi, D. Posillipo, G. Primiero, C. Quaggio, Brioxalkemy: a bias detecting tool, in: G. Boella, F. A. D'Asaro, A. Dyoub, L. Gorrieri, F. A. Lisi, C. Manganini, G. Primiero (Eds.), Proceedings of the 2nd Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming co-located with the 22nd International Conference of the Italian Association for Artificial Intelligence (AI*IA 2023), Rome, Italy, November 6, 2023, volume 3615 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 44–60. URL: https://ceur-ws.org/Vol-3615/paper4.pdf.

[17] G. Coraglia, F. A. Genco, P. Piantadosi, E. Bagli, P. Giuffrida, D. Posillipo, G. Primiero, Evaluating ai fairness in credit scoring with the brio tool, 2024. arXiv:2406.03292.

[18] F. Azzalini, C. Cappiello, C. Criscuolo, S. Cuzzucoli, A. Dangelo, C. Sancricca, L. Tanca, Data quality and fairness: Rivals or friends?, in: D. Calvanese, C. Diamantini, G. Faggioli, N. F. 0001, S. M. 0001, G. Silvello, L. Tanca (Eds.), Proceedings of the 31st Symposium of

Advanced Database Systems, Galzingano Terme, Italy, July 2nd to 5th, 2023, volume 3478 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 239–247. URL: https://ceur-ws.org/Vol-3478/paper68.pdf.

[19] T. Scantamburlo, Non-empirical problems in fair machine learning, Ethics Inf. Technol. 23 (2021) 703–712. URL: https://doi.org/10.1007/s10676-021-09608-9. doi:10.1007/s10676-021-09608-9.

[20] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel, Fairness through awareness, CoRR abs/1104.3913 (2011). URL: http://arxiv.org/abs/1104.3913. arXiv:1104.3913.

[21] N. Grgic-Hlaca, M. B. Zafar, K. P. Gummadi, A. Weller, The case for process fairness in learning: Feature selection for fair decision making, 2016.

[22] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, S. Martino, J. and. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, Y. Zhang, Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018.

[23] T. H. Aasheim, K. Hufthammer, S. Ånneland, H. Brynjulfsen, M. Slavkovik, Bias mitigation with aif360: A comparative study, in: Proceedings of the NIK-2020 Conference, 2020. URL: http://www.nik.no/.

[24] M. J. Kusner, J. R. Loftus, C. Russell, R. Silva, Counterfactual fairness, 2018. arXiv:1703.06856.

[25] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, CoRR abs/1610.02413 (2016). URL: http://arxiv.org/abs/1610.02413. arXiv:1610.02413.

[26] S. Feuerriegel, M. Dolata, G. Schwabe, Fair AI: Challenges and opportunities, Business Information Systems Engineering 62 (2020). doi:10.1007/s12599-020-00650-3.

[27] F. Kamiran, T. Calders, Data pre-processing techniques for classification without discrimination, Knowledge and Information Systems 33 (2011). doi:10.1007/s10115-011-0463-8.

[28] F. Azzalini, C. Criscuolo, L. Tanca, E-fair-db: Functional dependencies to discover data bias and enhance data equity, J. Data and Information Quality 14 (2022). URL: https://doi.org/10.1145/3552433. doi:10.1145/3552433.

[29] H. Weerts, M. Dudík, R. Edgar, A. Jalali, R. Lutz, M. Madaio, Fairlearn: Assessing and improving fairness of ai systems, 2023. arXiv:2303.16626.

[30] F. P. Calmon, D. Wei, K. N. Ramamurthy, K. R. Varshney, Optimized data pre-processing for discrimination prevention, 2017. arXiv:1704.03354.

[31] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, Certifying and removing disparate impact, 2015. arXiv:1412.3756.

[32] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork, Learning fair representations, in: S. Dasgupta, D. McAllester (Eds.), Proceedings of the 30th International Conference on Machine Learning, volume 28 of *Proceedings of Machine Learning Research*, PMLR, Atlanta, Georgia, USA, 2013, pp. 325–333. URL: https://proceedings.mlr.press/v28/zemel13.html.

[33] T. Kamishima, S. Akaho, H. Asoh, J. Sakuma, Fairness-aware classifier with prejudice remover regularizer, 2012, pp. 35–50. doi:10.1007/978-3-642-33486-3_3.

[34] S. Hooker, Moving beyond algorithmic bias is a data problem, Patterns 2 (2021).

[35] S. H. Sengamedu, H. Pham, Fairlabel: Correcting bias in labels, 2023. arXiv:2311.00638.

[36] H. Jiang, O. Nachum, Identifying and correcting label bias in machine learning, 2019. arXiv:1901.04966.

[37] C. G. Northcutt, A. Athalye, J. Mueller, Pervasive label errors in test sets destabilize machine learning benchmarks, 2021. Preprint at https://arxiv.org/pdf/2103.14749.pdf.
[38] A. Olteanu, C. Castillo, F. Diaz, E. Kiciman, Social data: Biases, methodological pitfalls, and ethical boundaries, Frontiers in Big Data 2 (2019).
[39] S. Fabbrizzi, S. Papadopoulos, E. Ntoutsi, I. Kompatsiaris, A survey on bias in visual datasets, CoRR abs/2107.07919 (2021).
[40] H. Suresh, J. Guttag, A framework for understanding sources of harm throughout the machine learning life cycle, Equity and Access in Algorithms, Mechanisms, and Optimization (2021).
[41] CertNexus, Promote the ethical use of data-driven technologies, 2021. https://www.coursera.org/learn/promote-ethical-data-driven-technologies/lecture/5Ufbp/data-collection-bias.
[42] U. o. O. Centre for Evidence-Based, Catalogue of bias, 2022. https://catalogofbias.org/biases/.
[43] B. D'Alessandro, C. O'Neil, T. LaGatta, Conscientious classification: A data scientist's guide to discrimination-aware classification, Big Data 5 (2017) 120–134.
[44] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, R. Ghani, Aequitas: A bias and fairness audit toolkit, arXiv preprint arXiv:1811.05577 (2018).
[45] C. G. Northcutt, L. Jiang, I. L. Chuang, Confident learning: Estimating uncertainty in dataset labels, Journal of Artificial Intelligence Research (JAIR) 70 (2021) 1373–1411.
[46] D. Angluin, P. D. Laird, Learning from noisy examples, Mach. Learn. 2 (1987) 343–370.
[47] C. Batini, M. Scannapieco, Data Quality: Concepts, Methodologies and Techniques, Springer, 2006.
[48] R. Y. Wang, D. M. Strong, Beyond accuracy: What data quality means to data consumers, J. Manag. Inf. Syst. 12 (1996) 5–33.
[49] S. Canali, Towards a contextual approach to data quality, Data 5 (2020) 90. doi:10.3390/data5040090.
[50] C. Batini, C. Cappiello, C. Francalanci, A. Maurino, Methodologies for data quality assessment and improvement, ACM computing surveys (CSUR) 41 (2009) 1–52.
[51] M. Scannapieco, T. Catarci, Data quality under a computer science perspective, Journal of The ACM - JACM 2 (2002).
[52] L. Pipino, Y. W. Lee, R. Y. Wang, Data quality assessment, Commun. ACM 45 (2002) 211–218.
[53] A. Rula, et al., Time-related quality dimensions in linked data (2014).
[54] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, S. Auer, Quality assessment for linked data: A survey, Semantic Web 7 (2016) 63–93.
[55] K. Spiel, O. Keyes, P. Barlas, Patching gender: Non-binary utopias in hci, 2019, pp. 1–11. doi:10.1145/3290607.3310425.
[56] A. Nielsen, Practical Fairness, O'Reilly Media, Inc., 2020. URL: http://gen.lib.rus.ec/book/index.php?md5=F9752B2F9693C98855A51504FE224DF6.
[57] C. Batini, A. Rula, M. Scannapieco, G. Viscusi, From data quality to big data quality, Journal of Database Management 26 (2015) 60–82. doi:10.4018/JDM.2015010103.
[58] A. Black, P. Van Nederpelt, Dimensions of Data Quality (DDQ), DAMA NL Foundation, 2020.

[59] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, ProPublica (2016).
[60] E. Edenberg, A. Wood, An epistemic lens on algorithmic fairness, in: Eaamo '23: Proceedings of the 3Rd Acm Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 2023, pp. 1–10.
[61] B. Ruberg, S. Ruelos, Data for queer lives: How LGBTQ gender and sexuality identities challenge norms of demographics, Big Data and Society 7 (2020). doi:10.1177/2053951720933286.