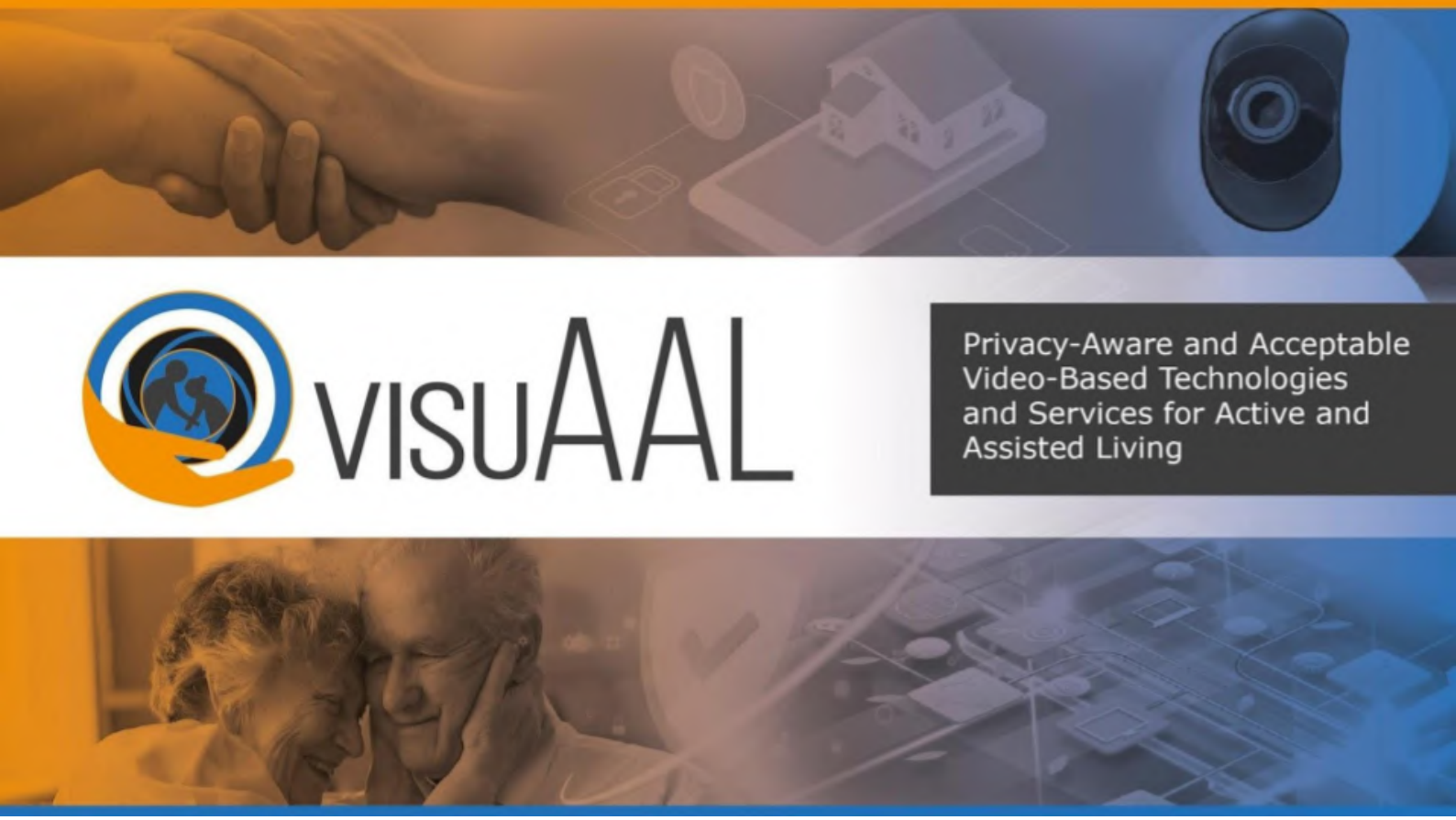


VISUAAL

Privacy-Aware and Acceptable Video-Based Technologies and Services for Active and Assisted Living

Deliverable 3.6

Report on paradigms, policies and metrics for algorithmic fairness

Sophie Noiret and Martin Kampel]

March 3rd 2024

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 861091".

Universitat d'Alacant
Universidad de Alicante
Project Coordinator

RWTH AACHEN UNIVERSITY

Stockholm University

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

TU WIEN

Table Of Contents

1 Terms and Definitions **3**
1.1 Algorithmic Governance . . . 3
1.2 Fairness, Bias and Transparency . . . 3
1.2.1 3 types of bias . . . 3
1.2.2 Different types of discrimination . . . 4

2 Bias Detection and Mitigation **6**
2.1 Competing Definitions of Fairness and Bias . . . 6
2.1.1 Disparate treatment and disparate impact . . . 6
2.1.2 Group Fairness and individual fairness . . . 6
2.1.3 Metrics . . . 6
2.2 Bias Mitigation Techniques . . . 9
2.2.1 Pre-processing . . . 10
2.2.2 In-processing . . . 11
2.2.3 Post-processing . . . 12
2.3 Tools for Bias Detection . . . 13
2.3.1 Tools for Assessing Fairness and Bias . . . 13
2.3.2 Tools for Mitigating Bias . . . 15
2.3.3 Platforms . . . 15

3 Datasets **19**
3.1 Data Acquisition . . . 19
3.1.1 Data Discovery . . . 19
3.1.2 Data Augmentation . . . 19
3.1.3 Data Generation . . . 20
3.2 Data Quality . . . 20
3.3 Data quantity . . . 21
3.4 Guidelines . . . 21
3.4.1 Datasheets . . . 22
3.4.2 Framework for dataset development transparency . . . 22
3.4.3 Dataset Nutrition Label . . . 22

4 Model Interpretation **23**
4.1 Explanation of black-box models . . . 23
4.1.1 Local Interpretable Model-agnostic Explanations (LIME) . . . 23
4.1.2 Shapley values . . . 24
4.1.3 Counterfactual explanations . . . 24
4.1.4 Explainability for CNN . . . 24
4.2 Evaluation of explanations . . . 24

5 Policies **26**
5.1 EU Regulations . . . 26
5.1.1 GDPR . . . 26
5.1.2 AI Act . . . 27

5.2 Standards 29
5.2.1 ISO/IEC TR 24028 29
5.2.2 ISO/IEC 27001:2013 29
5.2.3 IEEE P7003 30
5.3 Non-Binding Instruments 30
5.3.1 EU Artificial Intelligence Ethics Checklist 30
5.3.2 Guidelines on Automated individual decision-making and Profiling . 30

1 Terms and Definitions

1.1 Algorithmic Governance

Algorithmic governance is a mode of governance that is machine-like in nature, and founded on computer-based procedures and rules [11, 18]. While it is a mode of management by and based on automation, it is not intrinsically linked to any particular technology. However, the availability of large amount of data and the advances in machine learning have made algorithmic governance quicker, more efficient, and consequently more ubiquitous [18]

1.2 Fairness, Bias and Transparency

Bias and discrimination are closely related and are often used interchangeably in the literature, even though they have slightly separate meanings. Bias is often considered as an overarching term, defined as a preconceived, internal opinion about individuals or groups, that can be positive or negative. Discrimination however, is the actual negative treatment of or actions against groups and individuals based on one or multiple bias(es). In this regard, discrimination can be the result of intentional as well as unintentional actions, also in their manifestation through algorithms. The term discrimination often finds use in a legal context, where multiple regulations are in force to prevent intentional discrimination and also increasingly unintentional discrimination, or disparate impact [20].
A rare and thorough analysis of what bias means in the context of algorithms and automated decision making was provided by Friedman and Nissenbaum [22], who consider algorithms as biased, in that they have the potential to "systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others". The authors furthermore explain what "unfairly discriminates" means in the context of computer systems and algorithms, namely "if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate".
This definition is insofar useful, as it encompasses both the terms of bias and of discrimination and shows how discrimination should be considered as the biased action, the act of treating someone unfairly, bridging to the term of fairness. Fairness has also become a much used notion in the computer sciences, particularly in relation to fair algorithms and fair AI [67]. The understanding of having fair algorithms, fair automated decision making, and of fairness in general is that all individuals and groups are treated equally. Referring again to the earlier definition of Friedmann and Nissenbaum [22], this shows that fairness stands in direct opposition of discrimination.

1.2.1 3 types of bias

Classifications and categorisations that are made through automated decision making have particularly been shown to disfavour minorities unfairly, automatically discriminating them based on biased algorithms [48, 26]. Having seen that an algorithmic system is biased if it "unfairly discriminates against certain individuals or groups", it is worth considering the origin(s) of the bias. Generally, there are three different types of algorithmic biases

that can be differentiated: (1) a pre-existing bias; (2) a technical bias; (3) an emergent bias [22].

A pre-existing bias is a bias that is prevalent within our societies, and is the preconceived, internal opinion about individuals or groups. A pre-existing bias stems from the social institutions, from our norms and attitudes and generally manifests itself as discriminative through our practices. This also means that the pre-existing bias can reflect solely the personal opinion and prejudice of one individual - in this case of someone who is responsible for the design of the algorithmic system. But the bias can also be systemic within our society and thus way more difficult to address. There are many examples of pre-existing biases in algorithms, as it are the ones that appear to be the most logical, mirroring the biases that are existing in our societies. A common example from policing is the use of predictive policing algorithms. The technology is intended to direct police forces into areas in which most a criminal activity is most likely to happen. However, as many studies have shown [44][9][57], the prediction algorithm is very often trained on biased data, collected through years of police work that predominantly targets minorities and people of colour. The algorithm is simply repeating a pre-existing bias. However, as it becomes a defining characteristic of the system, it risks to further reinforce and materialize this bias, creating a feedback-loop with a severe impact on the targeted groups in our society.

The second type of algorithmic bias is inherent of the technology itself. Technical biases usually emerge within the design process of the algorithms and are the results of limitations of computer tools such as hard- and software. They can be the result of errors in coding, and in the construction and design of the algorithm [37]. But they also can emerge when system designers attempt to digitalise human qualities, when they are trying to make fundamental human aspects machine readable.

The third type of algorithmic biases are emergent biases. These do not exist (per se) in the technology straight “out of the box.” Instead, they emerge over time as new knowledge is created that can’t be integrated into the algorithm. Or through the interaction between the technology and the users, producing outcomes that were not intended or considered by the system designers. As automated decision making technologies are increasingly integrated in everyday societal practices, individuals have to adapt to incorporate these technologies into their routines. In most cases, this means that they need to make themselves ”machine-readable” [40]. If such technologies are applied in areas with a low technological knowledge, amongst individuals which have difficulties reading, hearing, seeing, etc., biases quickly emerge.

1.2.2 Different types of discrimination

As with bias, there are also different aspects of discrimination, that need to be considered. Although the concept of discrimination as an unfair treatment of individuals or groups remains the same, this treatment can be based on different attributes. Direct discrimination describes the unfair treatment that is based on protected grounds, such as age, gender, ethnicity, disability, creed, sexual orientation, etc. However, unfair treatment can also be the outcome of indirect discrimination, through proxies[46]. It is worth noting, for

instance, that COMPAS allegedly does not explicitly use race as an input. However, the 173-questions long questionnaire includes sections about the neighbourhood and family history of the defendant.

Besides direct and indirect discrimination, there are also two other forms of discrimination that need to be mentioned here, as they increasingly emerge through the use of algorithms and automated decision making. These are intersectional discrimination and emergent discrimination.

Intersectional discrimination addresses the more complex situations of discrimination that occur through a combination of discriminating characteristics. The idea behind intersectional discrimination is that, for example, the combined discrimination based on gender and ethnicity for women of colour is experienced differently than single entities of discrimination based on ethnicity and based on gender [46]. Particularly through algorithmic profiling and individualised and personalised decision-making, these intersection of discriminating identities risk to occur more often, because much more individual characteristics are used as a method to assess individuals. Which also brings us to the aspect of emergent discrimination. As with emergent bias seen before, also discrimination can occur over time, without having accounted for the potential of future discrimination situations. This is even more the case when intersectional discrimination is taken into account, where discrimination might emerge due to a combination of potentially discriminating characteristics.

2 Bias Detection and Mitigation

2.1 Competing Definitions of Fairness and Bias

In the broadest of terms, it can be defined as impartial and just treatment without favoritism or discrimination, moving the burden of definition on the terms favoritism and discrimination. This looseness in definition carries over to the domain of machine learning, in which fairness can be defined as the absence of algorithmic bias. However, some definitions have been broadly used in the scientific community.

2.1.1 Disparate treatment and disparate impact

The first distinction that must be made is between disparate treatment and disparate impact. While the terms come from US law[1], they have been used outside this context. Disparate treatment refers to intentional discrimination or D, while disparate impact occurs when a policy or an outwardly neutral criteria for decision ends up affecting a group more than the other. Crucially, disparate impact can occur even without an explicit intention to discriminate. For instance, basing the decision to grant a loan on the ZIP code of the applicant can lead to disparate impact in heavily segregated areas. Assessing whether or not an algorithm exhibits either of these practices is problematic when little is known about the inner workings of the system, or about the predominant factors influencing a prediction.

2.1.2 Group Fairness and individual fairness

The second distinction is between group fairness and individual fairness [14, 47]. In individual fairness, the goal is to have similar predictions for similar individuals. Counterfactual fairness is an example of such definition, in which a model is considered fair if the prediction for an individual is the same as it would be for a counter-individual whose attributes are identical except for the sensitive attribute. Group fairness, on the other hand, requires that different groups (with regard to the sensitive attributes) are treated equally. This can translate into having a measure of performance (precision, recall, f1-score, etc.) be equal across groups. In other circumstances, group fairness can be measured through, for instance, demographic parity.

2.1.3 Metrics

An appropriate choice of a metric can change our vision of the fairness of an algorithm. As a large number of these metrics rely on confusion matrix, which presents a summary of the predictions made by a classifier against the actual labels, an example of a confusion matrix with the associated terms can be seen in Fig. 1:

[1] EEOC v. Sambo's of Georgia, Inc., 530 F. Supp. 86, 92 (N.D. Ga. 1981)

Figure 1: Confusion Matrix and Associated Metrics

		Prediction			
	Total Population	Predicted Positive	Predicted Negative	$Prevalence = \frac{\sum TP}{\sum Total\ Population}$	
Ground Truth	Ground Truth Positive (GTP)	True Positive (TP)	False Negative (FN)	True Positive Rate , Sensitivity , Recall $TPR = \frac{\sum True\ Positive}{\sum GTP}$	False Negative Rate , Miss Rate $FNR = \frac{\sum False\ Negative}{\sum GTP}$
	Ground Truth Negative (GTN)	False Positive (FP)	True Negative (TN)	False Positive Rate , Fallout $FPR = \frac{\sum False\ Positive}{\sum GTN}$	True Negative Rate , Specificity $TNR = \frac{\sum True\ Negative}{\sum GTN}$
	$Accuracy = \frac{\sum TP + \sum TN}{\sum Total\ Population}$	Positive Predictive Value , Precision $PPV = \frac{\sum TP}{\sum Predicted\ Positive}$	False Omission Rate $FOR = \frac{\sum FN}{\sum Predicted\ Negative}$	Positive Likelihood Ratio $LR+ = \frac{TPR}{FPR}$	Diagnostic Odds Ratio $DOR = \frac{LR+}{LR-}$
		False Discovery Rate $FDR = \frac{\sum FP}{\sum Predicted\ Positive}$	Negative Predicitive Value $NPV = \frac{\sum FN}{\sum Predicted\ Negative}$	Negative Likelihood Ratio $LR- = \frac{FNR}{TNR}$	

Accuracy Parity

The accuracy is equal across groups.

Demographic Parity and Proportional or Statistical Parity

These terms are often used synonymously ([14, 70], in which case they are if the likelihood of a positive prediction is the same for protected and unprotected groups. It should be noted, however, that some tools use other definitions. Fairness (R)[2], for instance, considers that demographic parity is achieved when the absolute number of positive predictions in the subgroups are close to each other.

Predictive Rate Parity

The positive predictive values are equal across groups.

Conditional Demographic Parity

Also called conditional non-discrimination [13]. Controlling for a set of legitimate factors, the probability of being predicted positive is equal across groups.

Equal opportunity

True Positive Rate is equal across groups.

Equalized Odds

True Positive Rate and False Positive Rate are equal across groups.

Conditional use accuracy

Negative Predictive Value and Positive Predictive Value are equal across groups. When the result is not a binary classification but a risks-core, this metrics becomes "calibration".

[2]see Section 2.3.3

Calibration

Given a particular score, the probability of being ground-truth positive (resp. negative) is equal across groups.

False Positive (resp. Negative) Rate Parity

False positive (resp. negative) rate parity is achieved if the false positive (resp. negative) rates in the subgroups are equals.

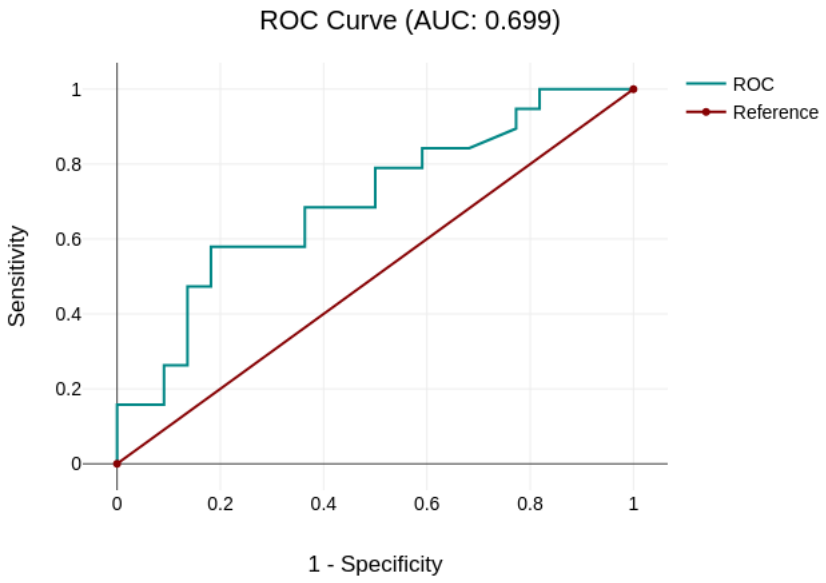
Positive (resp. Negative) Predictive Value Parity

Positive (resp. negative) Predictive Value parity is achieved if the Positive (resp. negative) Predictive Value in the subgroups are equals.

Specificity Parity

Specificity parity is achieved if the specificity in the subgroups are close to each other. This function can be considered the 'inverse' of the equalized odds.

Figure 2: An example of a ROC curve, made in `https://datatab.net`



ROC AUC Parity

The ROC (Receiver Operating Characteristic) curve is a plot that illustrates the performance of a binary classification model. It displays the trade-off between the true positive rate (sensitivity) and the false positive rate (specificity). An example of a ROC curve can be seen in Fig. 2. The AUC (Area Under the Curve) varies between 0 and 1. An AUC of 0.0 signifies a model with predictions that are entirely incorrect, while

an AUC of 1.0 indicates a model with predictions that are entirely accurate. ROC AUC parity is achieved if the ROC AUC is equal for all subgroups.

Generalized Entropy Index

The Generalized Entropy Index [62] is a statistical measure used to assess the level of inequality or diversity within predictions. It considers differences in an individual's prediction (b_i) to the average prediction accuracy (μ), where n is the number of predictions, $b_i = \hat{y}_i - y_i + 1$ and $\mu = \frac{\sum_i b_i}{n}$

$$GEI = \frac{1}{n\alpha(1-\alpha)} \sum_{i=1}^{n} \left((\frac{b_i}{\mu})^{\alpha} - 1 \right)$$

Counterfactual Fairness

It aims to ensure that decisions made by a model would remain the same even if sensitive attributes of an individual were different. The formal definition given in [41] is that a predictor $\hat{Y}$ is counterfactually fair if under any context $X = x$ and $A = a$

$$P(\hat{y}_{A \leftarrow a}(U) = y | X = x, A = a) = P(\hat{y}_{A \leftarrow a'}(U) = y | X = x, A = a)$$

where A, X and Y represent the protected attributes, remaining attributes, and output of interest respectively.

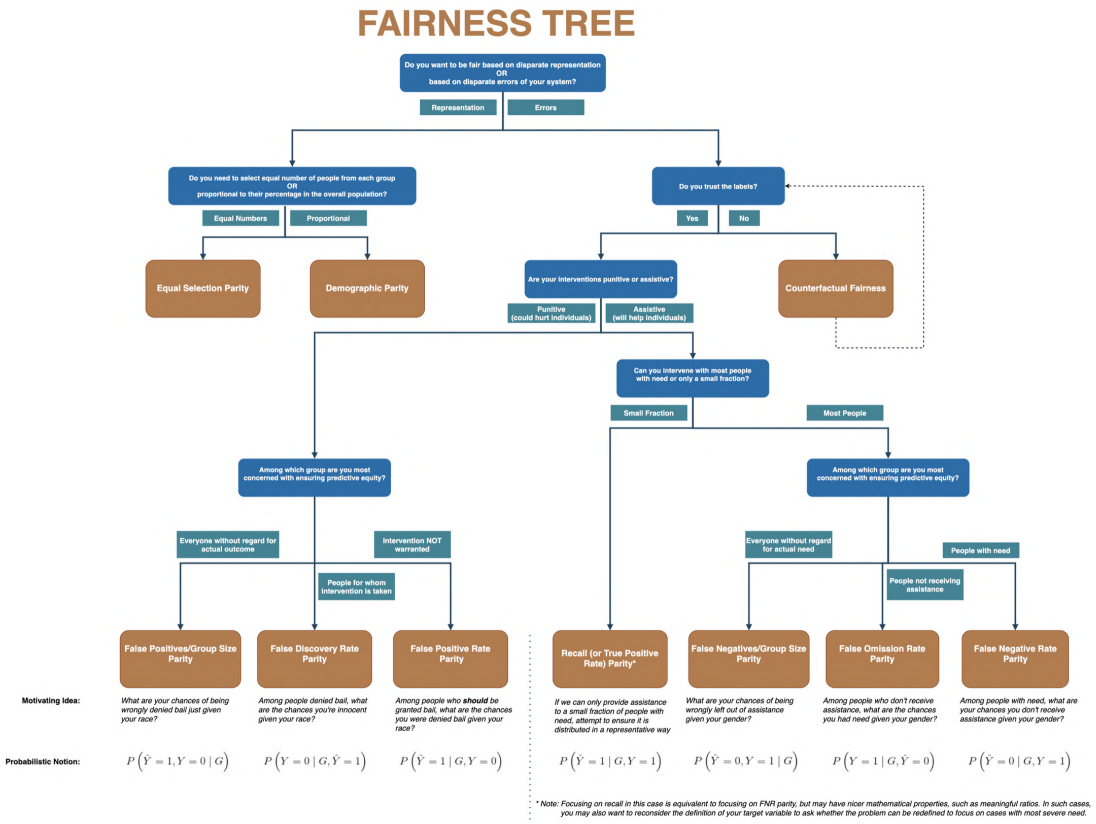
As we can see, fairness metrics are numerous and do not have universal definitions. Moreover, some of them are mathematically incompatible: equalized odds and conditional use accuracy can only be achieved at the same time if the prevalence is equal across group, or in the case of a perfect classifier [7], which means that practionners must choose which metrics they want to prioritize in their system. Aequitas[3] proposes a "Fairness Tree" (see Figure 3) to help make this choice.

2.2 Bias Mitigation Techniques

Selecting the appropriate definitions of fairness is only the first step; subsequently, practitioners must determine how to achieve it. Both [14] and [47] categorize these approaches into pre-processing, in-processing, and post-processing. As [47] highlights, the choice of approach depends on the type of bias being addressed and the resources available to the practitioner. For example, if the goal is to enhance fairness in an already trained system without access to training data or the algorithm, neither pre-processing nor in-processing would be feasible. [14] mentions that bias mitigation methods like adversarial learning and constraint optimization can fall into multiple categories. We do not give an exhaustive list of all bias mitigation techniques, but a curated selection of relevant approaches.

[3] `http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/`, last accessed 27/02/2024

Figure 3: Fairness Tree, from http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/



2.2.1 Pre-processing

Pre-processing methods focus on altering the training data.

Optimized Pre-processing

Optimized pre-processing [12] learns a probabilistic transformation that edits the features and labels in the data with group fairness, individual distortion, and data fidelity constraints and objectives. This technique is not applied during the training of the system whose bias we want to mitigate, but to the potentially biased dataset on which this system will be used.
The optimization problem is to minimize the utility loss, which means that the distribution of transformed labels and features must be statistically close to the distribution of original labels and features.The first additional constraint is to limit the dependence of the transformed outcome on the protected features, meaning that in the transformed distribution, the conditional distribution of the outcome on the sensitive features is "close" to a target distribution, with the authors of [12] remarking that the meaning of "close" and the target distribution should be informed by societal aspects. The second additional constraint is distortion control, which means that the transformation should avoid very large changes when mapping the original labels to transformed labels, in order to preserve

the predictive power of the algorithm trained on this data.

Disparate Impact Remover

As the title indicate, Disparate Impact Removal[21] seeks to remove disparate impact. This method, using the US Equal Employment Opportunity Commission (EEOC) guidelines (sometimes known as the "80% rule") defines disparate impact as a positive likelihood ratio (see Figure 1) over 1.25.
In order to remove this disparate impact, [21] proposes an algorithm to "repair" the dataset(i.e. remove disparate impact) by only changing the non-sensitive attributes. In order to preserve the predictive power of the system that will be trained on that data, they add a constraint that the transformation must preserve the rank of the non-sensitive attribute.

Reweighing

In reweighing, instead of changing the labels, the labels and features in the training dataset are assigned weights. By carefully choosing the weights, the training dataset can be made discrimination-free with regards to the sensitive attributes without having to change any of the labels.
Reweighing presents an approach that spans both pre-processing and in-processing methodologies. For instance, in [36], the aim is to assign weights considering the probability of an instance with a particular class and sensitive attribute combination, representing a pre-processing technique. Conversely, [38] initially constructs an unweighted classifier, then proceeds to learn sample weights, and finally retrains the classifier using these weights, illustrating an in-processing strategy. Similarly, [33] identifies sensitive training instances as a pre-processing step, but subsequently learns weights for these instances during the optimization process for the chosen fairness metric, representing an in-processing technique.

2.2.2 In-processing

In-processing methods intervene during training.

Learning Fair Representations

In [72], the authors aim to transform each individual, depicted as a data point within a given input space, into a probability distribution within a novel representation space. The goal of this transformation is to eliminate any discernible information regarding whether the individual belongs to a protected subgroup, while preserving as much other pertinent information as feasible. This novel representation is expressed as a probabilistic mapping to a collection of prototypes, although it's important to note that this is just one potential form of intermediate representation among many. Furthermore, these representations are optimized to ensure that any classification tasks utilizing them achieve maximum accuracy.
This method can be understood as both an in-processing approach, and be compared, as

the authors do, to other machine learning models, or as a pre-processing method which modify the data to transform it into representation on which different models can be trained. This method also do not require the practioner to chose beforehand which fairness criteria they want to optimize: [72] uses both group fairness (in the form of statistical parity) and individual fairness (by comparing a model's classification prediction of a given point to its nearest neighbor)

Adversarial Debiasing

Adversarial Debiasing is a technique that uses adversarial training to mitigate bias. It involves simultaneous training of a predictor and a discriminator. In [73], fairness measures are explored within the framework of adversarial debiasing, focusing on supervised deep learning tasks where the objective is to predict an output variable Y based on an input variable X, while ensuring impartiality regarding a variable Z, referred to as the protected variable. In these learning systems, the predictor $\hat{Y} = f(X)$ is trained using a dataset comprising tuples of input, output, and protected variables (X, Y, Z). The predictor f typically has access to the protected variable Z, although it's not strictly necessary. This setup enables the selection of which biases are deemed undesirable for a specific application by specifying the protected variable.
The predictor is trained to predict Y given X.It is assumed in [73]that the model is trained by adjusting weights W to minimize a loss function $L_P(\hat{y}, y)$ using a gradient-based method like stochastic gradient descent. Subsequently, the output layer of the predictor feeds into another network termed the adversary, which aims to predict Z. This component of the network corresponds to the discriminator in a typical GAN. The adversary incorporates a loss term $L_A(\hat{z}, z)$ and weights U, with potential additional inputs depending on the fairness definition being targeted. For Demographic Parity, the adversary receives the predicted label $\hat{Y}$, enabling it to predict the protected variable using solely the predicted label, against which the predictor seeks to defend. For Equality of Odds, the adversary receives both $\hat{Y}$ and the true label Y. For Equality of Opportunity on a specific class y, the training set of the adversary can be restricted to instances where $Y = y$.

2.2.3 Post-processing

Post-processing occurs after modeling.

Equalized Odds Postprocessing

After a model has made predictions, Equalized Odds Post-processing adjusts these predictions to satisfy the equalized odds criterion. Unlike previously discussed bias mitigation technique, this (and other post-processing techniques) do not require to be able to modify the training data or the training procedure. [29] presents the process of constructing an equalized odds or equal opportunity predictor $(\tilde{Y})$ from a potentially discriminatory binary predictor $(\hat{Y})$ or score (R). The approach involves creating $\tilde{Y}$ based solely on the random variables (R, A), where A represents a protected attribute such as race or gender. This derived predictor $\tilde{Y}$ should be independent of the features (X) conditional on (R, A).

While constructing $\tilde{Y}$ relies on information about the joint distribution of (R, A, Y), prediction only requires knowledge of (R, A). The training process remains unchanged, with the focus on a post-learning step. However, it's essential to minimize loss by designing derived predictors $(\tilde{Y})$ that minimize the expected loss while satisfying equalized odds. The authors present an optimization problem to derive the optimal equalized odds predictor from $\hat{Y}$ and A, describing it as a linear program whose solution provides the optimal predictor.

Counterfactual Correction

Path-Specific Counterfactual Fairness (PSCF), introduced in [17], examines fairness in various decision-making paths that individuals may follow based on their characteristics. It transform the output of a classifier in accordance to the identification of unfair causal pathways through counterfactual correction. This method assumes a graphical causal model (GCM) between the variables. The latent inference-projection technique ensures fairness specific to different decision pathways by adjusting variables descended from the sensitive attribute during testing, while maintaining the original data-generation process intact during training.

2.3 Tools for Bias Detection

This summary is based on [14], which provides an entry-level overview of the state of the art and a list of current platforms. Here, we examine each platform is examined with regard to functionality, useability (e.g., license, maintenance, installation, source), and its place in the ML pipeline. As the tools vary considerably there is no uniform test scenario. Therefore, the investigations are centered around examples provided by the authors and in some cases additional tests. Nevertheless, the majority of the frameworks for measuring bias in binary classification or regression tasks use either the COMPAS[4] or German credit[5] dataset in their tutorials. For a quick overview, Table 1 lists all the tools and contains information about the license, source code availability, installation, maintenance, the year the project started, and the organization.

2.3.1 Tools for Assessing Fairness and Bias

The majority of tools fall into the category for assessing fairness and bias and predominantly consider binary classification or regression tasks. The exceptions are: Revise, which is the only platform focused on detecting bias in visual datasets, and Responsibly, which provides metrics for bias in word embeddings. Except for the Fairness R package, all the tools are implemented in Python. Fairlearn, Aequitas, and AI Fairness 360 additionally offer GUIs to guide users through the assessment process.

1. **Fairlearn**: Grouped and un-grouped metrics for binary classification and regression tasks. Provides additional functionality to perform aggregations over multiple

[4] `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`,(last accessed on 27/02/2024)

[5] `https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data`, last accessed on 27/02/2024

Table 1: Overview of the tools.

Tool	License	Source	Install	Maintained	Year	Organisation
Fairlearn	MIT	✓		✓	2020	Microsoft
Fairness (R)	MIT	✓	✓	✓	2020	independent
TFCO	Apache 2.0	✓	✓		2018	Google
Audit-AI	MIT	✓	✓		2018	pymetrics
Fairness-Measures	GNU v3.0	✓	✓		2018	independent
ML-Fairness-Gym	Apache 2.0	✓	✓		2020	Google
AIF360	Apache 2.0	✓	✓	✓	2018	IBM
Aequitas	MIT	✓	✓	✓	2018	University of Chicago
Fairness-Measures	GNU v3.0	✓	✓		2018	independent
Responsibly		✓	✓		2018	independent
FairTest	Apache 2.0	✓	✓	✓	2015	Columbia University
Revise	MIT	✓	✓	✓	2020	Princeton Visual AI Lab

measurements to obtain scalar results, which can be used, for example, for hyperparametertuning.

2. **Audit-AI**: Bias testing tool for demographic differences, especially in employee selection procedures.

3. **Fairness**: R package containing metrics for binary classification problems.

4. **Revise**: A Tool for measuring bias in visual datasets. Currently focuses on object-, gender-, and geography-based discrimination.

5. **Fairness Measures**: Python implementation for the metrics discussed in [76] including absolute measures and statistical tests. Compared to the other frameworks its functionality is rather limited and it uses a peculiar workflow.

6. **Aequitas**: Offers a set of measurements for binary classification and regression tasks. Additionally, it provides a web application to interactively audit datasets.

7. **Responsibly**: Responsibly is the only tool, which includes methods for NLP and especially word embeddings.

8. **AI Fairness 360**: Offers an extensive list of over 70 metrics.

9. **Fair Test**: Testing for unwarranted associations between the output and subgroups defined by protected variables.

2.3.2 Tools for Mitigating Bias

The tools for mitigating bias can further be subdivided concerning their position in the ML pipeline, namely pre-processing, in-processing, and post-processing.

Pre-Processing

1. **Revise**: A tool for revealing biases in visual datasets but additionally, it provides example actions (e.g., "Query images of baseball glove in different scenes like a sidewalk") to decrease said biases. In other words, the tool acts as a guide for pre-processing, which then still has to be done manually.
2. **AI Fairness 360**: Contains a reweighing approach that modifies the weight of different training examples described in [36], and a data transformation algorithm (optimized preprocessing) from [12].
3. **Fairlearn**: Provides a linear transformation to remove the correlation of the non-sensitive features and the sensitive features, while retaining as much information as possible.
4. **Responsibly**: Debiasing for word-embeddings as proposed by [8, 27, 29].

In-Processing

1. **Fairlearn**:Failearn's in-processing approach is based on constraint optimization using Lagrange multipliers and works for binary classification and regression. It can wrap any base learning algorithm with a fit and predict method.
2. **TFCO:**:TensorFlow Constraint Optimization is a similar approach Fairlearns constraint optimization but with an additional "shrinking" process to further enhance the performance. As opposed to Fairlearn, TFCO only works with TensorFlow.
3. **AI Fairness 360**: Contains multiple mitigation techniques listed in Section 2.2

Post-Processing

1. **Fairlearn**: Fairlearn contains a Threshold Optimizer that takes a classifier and transforms its output to enforce certain parity constraints. It is based on [21].
2. **AI Fairness 360**: Contains multiple mitigation techniques listed in Section 2.2
3. **Responsibly**: Multiple threshold definitions for binary classification tasks..

2.3.3 Platforms

Fairlearn

Fairlearn is an open-source, and community-driven python package containing metrics to assess the fairness of a given system and algorithms to mitigate observed fairness issues. It is currently actively developed and licensed under the MIT License. Fairlearn has grown from a project at Microsoft Research in New York City[6] .

[6] `https://fairlearn.org/`(last accessed on 27/02/2024

Assessment The metrics for fairness assessment include grouped and ungrouped metrics for binary classification and regression tasks. The tool provides additional functionality to perform aggregations over multiple measurements to obtain scalar results, which can be used for e.g., hyperparametert uning. The metrics module is accompanied by a Jupyter notebook widget provided by Microsoft included in responsible-ai-widgets[7], which offers an interactive experience to assess a model's fairness and performance.

Mitigation The methods for bias mitigation cover pre-, in-, and post-processing methods. The algorithms are not tied to a specific ML framework and are implemented as a "wrapper" for any model with a fit and predict method. The pre-processing revolves around a linear transformation to remove the correlation of the non-sensitive features and the sensitive features while retaining as much information as possible. The in-processing approach builds on the idea of constraint optimization using Lagrange multipliers as described in [23, 24]. For post-processing, Fairlearn contains a Threshold Optimizer that takes a classifier and transforms its output to enforce certain parity constraints. It is based on [21].

Fairness (R)

Fairness is an R package providing fairness metrics for binary classification problems and tools to visualize and compare the results. It is actively developed and maintained by Nikita Kozodoi and Tibor V. Varga, licensed under the MIT License, and available as a CRAN package. The following lists the metrics provided in the latest version (1.2.2) including a short description taken from the tutorial[8] .

1. **Predictive rate parity**: Predictive rate parity is achieved if the precisions (or positive predictive values) in the subgroups are close to each other. The precision stands for the number of the true positives divided by the total number of examples predicted positive within a group.

2. **Demographic parity**: Demographic parity is achieved if the absolute number of positive predictions in the subgroups are close to each other. This measure does not take true class into consideration and only depends on the model predictions.

TensorFlow Constrained Optimization

TensorFlow Constrained Optimization (TFCO) is a library for optimizing inequality-constrained problems in TensorFlow8. It is developed by Google-Research, is licensed under Apache 2.0, and classifies as an in-processing approach for mitigating bias. While the library can handle arbitrary objective functions and constraints it proofs to be most useful when used with the built-in helper functions. They allow defining constraints

[7] `https://github.com/microsoft/responsible-ai-widgets`, (last accessed on 27/02/2024)

[8] `https://kozodoi.me/blog/20200501/fairness-tutorial#6.-Computing-fairness-metrics`, last accessed on 27/02/2024

with rates (e.g. the error rate, true positive rate, recall, etc). Because rates are non-differentiable they are approximated with so-called proxy constraints so they can be used with gradient-based algorithms.
Constrained problems can lead to oscillation instead of convergence and for that case, the library offers a procedure called shrinking: Multiple snapshots of the model are collected during training time and then post-processed to a stochastic model. The idea is that even if none of the models performs particularly well, multiple models combined usually yield better results. For more details, and theoretical results refer to [25, 26].

Audit-AI

Audit-AI is an open-source Python package developed by pymetrics for bias testing in classification and regression tasks. The latest version (0.1.1) was released on 29th July 2020 under the MIT License. It is built on top of pandas and sklearn and with a focus on employee selection procedures.

Fairness-Measures

Fairness Measures[9] is a code repository containing implementations from [76] for quantifying discrimination. The code is written in Python and released under the GPL-3.0 License. The program expects the input to be a dataset where each row represents a person. One of the attributes has to be the target (predicted by a model) and can be either binary or numeric. Furthermore, protected attributes can be declared with the prefix protected.

ML-Fairness-Gym

ML-Fairness-Gym is a tool for simulating the impacts of deploying ML-based decision systems in social environments. It is developed by Google and built upon OpenAI's Gym API. The current environments can replicate the dynamic studies proposed in [27] (lending), [28, 29] (attention allocation), as well as [30, 31] (strategic manipulation).

REVISE

REVISE (REvealing VIsual biaSEs) is a tool for measuring and mitigating bias in visual datasets developed at Princeton University. It aims to address biases early in the ML pipeline and currently offers metrics in three categories:

1. **Object-based:** The object-based analysis considers statistics about object frequency, scale, context, or diversity of representation. It is based on instance labels and utilizes bounding boxes and object category if available. Additional semantic labels are inferred by automated computer vision techniques.

[9] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.fisherexact.html`, last accessed on 27/02/2024

2. **Gender-based:** The gender-based analysis gives insight into contextual representation, interactions, appearance differences, gender label inference. It requires gender labels and is limited to a binarized socially-perceived gender expression.

3. **Geography-based:**The geography-based analysis relies on country- or subregion-label annotations, ideally accompanied by information about who took the picture, and additional tags. The metrics are country distribution, local language analysis, tag counts, and appearances.

For each of the three categories, the tool provides actionable insights based on the metrics described above. The actions vary from concrete instructions (e.g. collect more images of airplanes) to more nuanced observations.

AI Fairness 360

AI Fairness 360 is an open-source Python and R toolkit developed by IBM Research. It offers metrics to measure individual and group fairness and an extensive collection of algorithms for mitigating bias. The library is designed to be extendable and released under the Apache-2.0 License.

Aequitas

Aequitas [43] is an open-source toolkit to audit ML models for bias and discrimination developed at Carnegie Mellon University. The python library is actively maintained and is released under the MIT License. The tools are also available as a web application[10].

Responsibility

Responsibly[11] is an open-source toolkit containing metrics and algorithmic interventions for binary classification tasks as well as metrics and debiasing methods for word embeddings. The python package is released under the MIT license.

Fair Test

Fair Test is a tool for discovering and testing for unwarranted associations between an algorithm's outputs and certain user subpopulations identified by protected features. The tool has been developed in 2015 at the Columbia University and is written in Python 2.7. The metrics cover binary classification and regression tasks and can be extended with custom ones.

[10] `http://aequitas.dssg.io/upload.html`, last accessed on 27/02/2024
[11] `http://docs.responsibly.ai/`, last accessed on 27/02/2024

3 Datasets

Data availability, collection, cleaning and management is the number one challenge faced by Machine Learning (ML) practitioners [3]. As illustrated by the saying "Garbage in, garbage out", the quality of a machine learning model is directly linked to the quality of the data that it is trained on [68]. The training data is the input that the model will learn from. It can take multiple forms: images, text, tabular data (like an excel sheet), sound, etc. It can also contain additional information, such as bounding boxes or annotations. In supervised learning, the training data will be augmented with labels, i.e with the information that your model is trying to predict.

3.1 Data Acquisition

Data acquisition can be separated in data discovery, data augmentation and data generation [58].

3.1.1 Data Discovery

Data discovery consists of using existing datasets. These datasets can be found on websites such as Kaggle[12], DataHub[13] or Google Dataset Search[14]. Companies can also use collections of internal datasets referred to as "data lakes". The term was coined by James Dixon in 2011[15] to to describe a repository of raw data in which one needs to "fish" for useful datasets [54].

3.1.2 Data Augmentation

Data augmentation can be used to fill out missing information or to extend existing datasets [58]. This is the approach chosen to create the datasets used to train the pre-trained models available in the dlib library[16], for which the VGGFace dataset [17] and the face scrub dataset [18] were supplemented with images scraped from the internet. Datasets of images can also be augmented by modified images from the dataset, using geometric transformations (rotation, flips, cropping), colorspace transformations [61] or noise injection [59].

[12] `https://www.kaggle.com/`, last accessed on 27/02/2024
[13] https://datahub.io/, last accessed on 27/02/2024
[14] `https://datasetsearch.research.google.com/`, last accessed on 27/02/2024
[15] `https://www.forbes.com/sites/ciocentral/2011/07/21/big-data-requires-a-big-new-architecture/`, last accessed on 27/02/2024
[16] `https://github.com/davisking/dlib-models`, last accessed on 27/02/2024
[17] `http://www.robots.ox.ac.uk/~vgg/data/vgg_face/`, last accessed on 27/02/2024
[18] `http://vintage.winklerbros.net/facescrub.html`, last accessed on 27/02/2024

3.1.3 Data Generation

If there is no existing data, datasets can be generated manually or automatically. Crowdsourcing is an approach to data collection in which web users contribute, gather or preprocess data [4, 24]. It is done through platforms such as like Amazon Mechanical Turk (AMT), in which human workers complete tasks. Typical tasks include natural language description of images [69] and generate questions for question answering datasets [64]. Synthetic data generation is used in cases where restricted access to data, as well as privacy concerns, limit the possibility of outside help [49]. It can be generated through deep learning techniques such as Generative Adversarial Networks (GANs) or Variational AutoEncoders (VAEs), or using software such as Blender or Cloud compare to create synthetic visual or 3D data [28].

3.2 Data Quality

There are several frameworks [6, 53, 35, 39, 51] to evaluate data quality, without any clear consensus on which is the best. However, there are recurring criteria:

- Accuracy [6, 53, 35]: The labels used in the training data must correspond to the reality or the model runs the risk of 'learning' incorrect information. This can also be referred to label errors or label noise.
- Uniformity/Consistencyy [6, 53]: In order to be usable, each data point must have the same structure. For instance, if the dataset is an image dataset, all data points have to be images. If the dataset consists of sensor reading at certain times, the readings must be represented the same way and the time stamp must follow the same format. When this is not the case, a preprocessing phase can be applied to the data prior to the machine learning.
- Usability [6]: In addition to being uniform, the data needs to be in a form that is readable by the program. For instance, a black and white image might be represented by a two-dimension array, each cells corresponding to a pixel.
- Representativeness [53]: Representativeness is a measure of the similarity between the data and the reality . It evaluates the ability of the sampled information to replicate the larger population it was sampled from.
- Currentness [6, 53, 35]: Data need to be as up-to-date as possible. Data will necessarily be from the past, but efforts must be made to not have too avoid data-drift.
- Balance [39]: In classification problems, the distribution of the samples between classes might not be equal. While it can be representative of the reality, using a severely imbalanced dataset can lead to poor performances, especially for the minority class. Class imbalance can be intrinsic to the problem (for instance, fraud detection or rare event prediction) or be the result of biased sampling or measurement errors.
- Fairness [35, 51]: The data must be as free of bias as possible. This criteria can be hard to define and to achieve. It can relate to balance (for instance, face datasets

tend to be heavily skewed toward Caucasian faces, leading to face analysis and recognition algorithms performing worse on other groups [42]). It can also come into conflict with other criteria: if a group is a minority in the general population does not mean it should be a minority in the dataset, even if that would be 'representative'. Historical bias must also be taken into account: records of arrest, for instance, might be a poor substitute for actual criminality considering racial bias in policing [1].

- Quantity [6, 35]: The amount of data must be appropriate to complete the task. However, there is no hard consensus on how much data you need.

3.3 Data quantity

Depending on the source, the recommendation for adequate dataset size can be:

- 100 to 500 examples per label
- At least 1000 for each plausible case [19]
- Ten times more data than degrees of freedom. To put it differently, 10 times more examples than features [20]
- For image classification, 1000 images per class[21]
- Etc.

In the academic literature, there are conflicting answers to what the optimal amount of data is. In tweet sentiment classification, the authors of [52] study the influence of dataset size for tweet sentiment classification. All models tested improve with the dataset size, but the effect becomes less noticeable as the dataset grows: tripling the dataset size from 1000 to 3000 improves the Area Under the Curve (AUC) up to 5%, but tripling it from 81000 to 243000 only improves the AUC by 1%. For 2 labels (positive or negative) and 1000 features, the optimal dataset size for this task is 81000.
In the medical domain, the performances stop improving over 490 datapoints for skin segmentation (2 labels, 4 features) and for hospital readmission prediction (3 labels, 55 features) [2].
According to [34] the representativeness of the dataset and the complexity of the model used have more influence than the size of the dataset.

3.4 Guidelines

In high-stakes Artificial Intelligence (AI), such as health, finance or public safety, specialized datasets are required, leading to AI practitioners undertaking data collection from scratch [59]. Unfortunately, the industry lacks standards for data collection and documentation [31, 59, 25]. Practitioners have reported not being able to discard poor quality

[19] https://www.v7labs.com/blog/quality-training-data-for-machine-learning-guide, last accessed on 27/02/2024

[20] https://towardsdatascience.com/machine-learning-rules-of-thumb-b50232b4b2f8, last accessed on 27/02/2024

[21] https://petewarden.com/2017/12/14/how-many-images-do-you-need-to-train-a-neural-network/, last accessed on 27/02/2024

datapoints because of the limited amount of data [59], with software engineers describing data collection as "almost like the Wild West" [31]. The lack of transparency with regards to the collection and annotation process have caused reproducibility issues [50] in datasets as widely used as Imagenet and CIFAR [55].
In order to combat these issues, guidelines for dataset documentation have been put forward.

3.4.1 Datasheets

Datasheets for datasets [25] makes an analogy with the electronics industry, in which every component is accompanied by a datasheet. It proposes a document that records the motivation, composition, collection process, pre-processing, cleaning, labeling, uses, and maintenance. By answering the questions laid down in each section, data scientists are encouraged to reflect on their practices and to emphasize transparency. Although this documentation process needs not be done at the collection phase (the creators of these datasheets provide an example for a dataset they did not create), it is most effective when dataset creators consider the questions before collection.

3.4.2 Framework for dataset development transparency

[32] emphasizes the critical role that datasets play in machine learning and proposes a framework for dataset development transparency that describes critical documents that should be produced during the dataset lifecycle.These include Requirements Specification (in order to make explicit why the data is being collected and to what use it will be put), Design Document (which lays down how the requirements are to be met and justify design decision), and Testing Report (in order to trace what requirement have been tested and what flaws have been discovered).

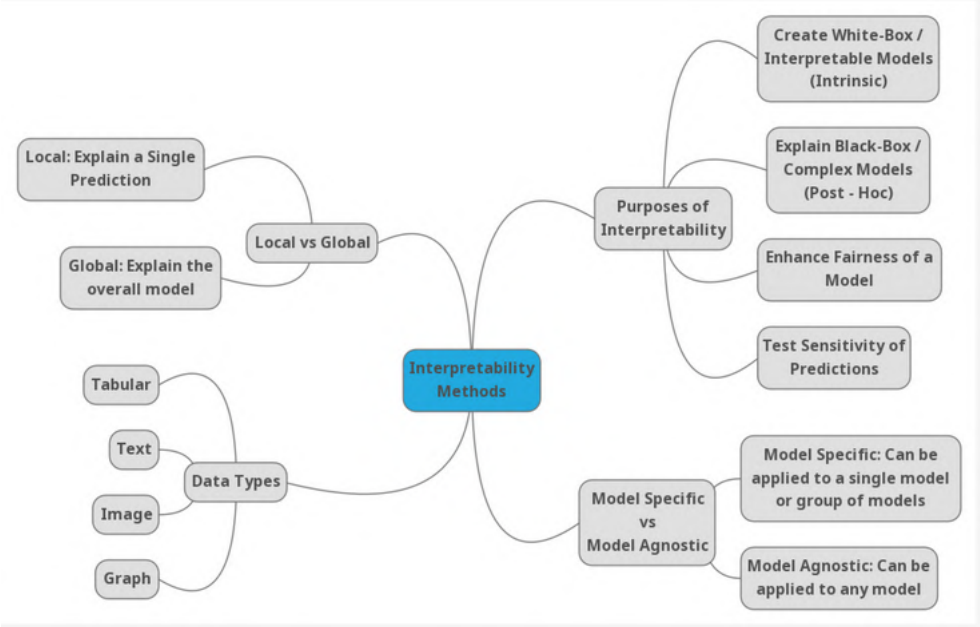
3.4.3 Dataset Nutrition Label

The Dataset Nutrition Label [30] is a tool to enhance the context, content and legibility of datasets. It takes a form similar to nutrition labels on food with labels, use cases and alerts, as well as information on the composition, provenance, collection and management of the dataset. However, this label is only intended for tabular data and is therefore focused on technical information.While it is designed to be modular, as to accommodate the needs of different types of datasets, the list of modules provided by the authors only contains two modules that improve the transparency of the collection process: the Metadata module (which contains keywords related to the dataset as well as a description) and the Provenance module (which contains information about the authors of the dataset).

4 Model Interpretation

The opacity of algorithmic systems might come from intentional trade or state secrecy, technical illiteracy or the inherent black-box nature of some algorithms [10]. Regardless of the cause, it complicates the task of assessing the fairness of the system, leading to the need for explainability techniques to explain either a single output (local explanation) or the overall model (global explanation). Explainability seeks to provide rationales for the decision made by a black-box model, either through global or local explanation. By contrast, a model is considered interpretable if it intrinsically can be understood by a human. Such models (Linear Regression, Logistic regression, Decision Tree, Rule-Based, Generalized Additive Models...) can be paired with natural language explanations and visualizations to improve clarity, but they are generally outperformed by deep-learning techniques..

Figure 4: Taxonomy mind-map of Machine Learning Interpretability Techniques, from [43]



4.1 Explanation of black-box models

This section explores explainability techniques that are model-agnostic can be used without assumptions about the architecture of the model.

4.1.1 Local Interpretable Model-agnostic Explanations (LIME)

LIME [56] creates a local explanation for an output, i.e. it only explains one prediction. The term 'Model Agnostic' refers to the fact that it can be used without assumptions about the model which made the prediction.First, new data points are generated in the neighborhood of the input we wish to explain: These new samples are then weighted to

give more importance to the closer datapoints. Then, a new interpretable model (such as linear regression) is trained on the new dataset.

4.1.2 Shapley values

Shapley values [45] explain a single prediction by measuring the feature importance among participating features. The concept of Shapley value is based on game theory, in which it is used to calculate the contribution of a player in a payoff. In our case, the "player" is a feature of interest. For one prediction and for one feature of interest, the value is computed by making a prediction with every combination of feature, with and without the feature of interest, and taking the average of difference from all combinations. The importance of individual features can also be averaged over several predictions to create a global explanation of the model [23, 5]

4.1.3 Counterfactual explanations

Counterfactual explanations(CFEs) are an emerging technique within the field of interpretable machine learning (ML) models. They provide "what-if" feedback in the form of "if an input data point were x' instead of x, then an ML model's output would be y' instead of y." Counterfactual explainability for ML models has not yet seen widespread adoption in the industry [65, 66]. In the context of interpretable machine learning, counterfactual explanations can be employed to elucidate predictions for individual instances. The "event" refers to the predicted outcome of an instance, while the "causes" are the specific feature values of that instance that were input into the model, thereby "causing" a certain prediction. When represented as a graph, the relationship between the inputs and the prediction is straightforward: The feature values cause the prediction

4.1.4 Explainability for CNN

As deep learning has led to both better performances in machine learning and a drop in the inherent interpretability of the models, a significant part of explainability has been dedicated to it. Class Activation Maps (CAM) [74] and theirsubsequent improvements [60, 15] highlight which part of the input was most relevant in the final classification in a Convolutional Neural Network (CNN). Deconvolution [71] and guided back-propagation [63] are techniques for visualizing the features learned by a CNN. ProtoPNet [16] uses a neural network architecture to learn features and prototypes which are then used to explain classification by comparing them to images in the training set.

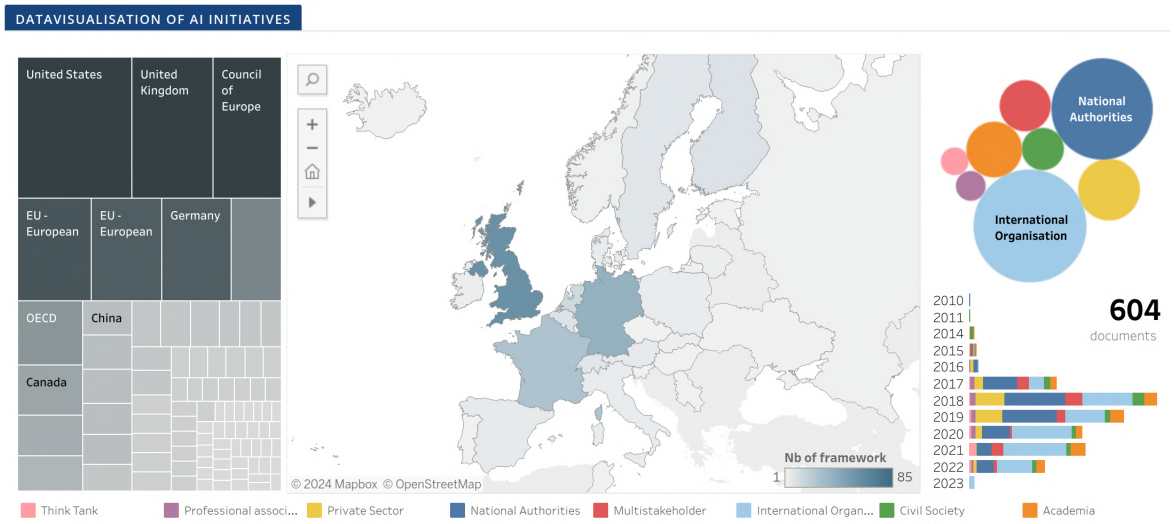
4.2 Evaluation of explanations

As deep learning has led to both better performances in machine learning and a drop in the inherent interpretability of the models, a significant part of explainability has been dedicated to it. Class Activation Maps (CAM) [74] and theirsubsequent improvements [60, 15] highlight which part of the input was most relevant in the final classification in a Convolutional Neural Network (CNN). Deconvolution [71] and guided back-propagation [63] are techniques for visualizing the features learned by a CNN. ProtoPNet [16] uses

a neural network architecture to learn features and prototypes which are then used to explain classification by comparing them to images in the training set. .3.2. Evaluation of explanations There is no widespread consensus on what makes a good explanation [75], but the following goals must be achieved: the explanation must be accurate (it must faithfully represent the actual working of the system), understandable (the person to which the explanation is presented can comprehend the information) and efficient (the explanation is quickly understandable) [13]. In order to measure these properties, [19] proposes three types of evaluations: application-grounded evaluations, human-grounded evaluations, and functionally-grounded evaluations. Application-grounded evaluations measure the quality of an explanation when it is presented to a domain-expert performing a real task, which means that they require experiments in the real world with the end-user. Human-grounded evaluations relax these constraints by having a layperson perform simplified tasks. Lastly, functionally-grounded evaluations require no human experiments and use a mathematical definition of interpretability as a proxy.

5 Policies

Figure 5: Data Visualization of AI Initiatives, from https://www.coe.int/en/web/artificial-intelligence/national-initiatives



As illustrated in Figure 5, the European Union (EU) has witnessed a surge in AI initiatives aimed at regulation to ensure the responsible and ethical development and deployment of artificial intelligence technologies. From the release of the European Commission's AI White Paper in 2020 to the proposed AI Act in April 2021, the EU has been actively shaping its regulatory landscape to address the challenges posed by AI while maximizing its potential benefits. In addition to EU-level initiatives, various national governments within the European Union have also launched their own AI regulatory efforts. Countries such as France, Germany, and the UK have introduced national AI strategies outlining priorities for AI development and governance. These strategies often include measures to promote innovation, address ethical concerns, and ensure the responsible use of AI technologies. Furthermore, non-binding instruments such as guidelines, codes of conduct, and best practices have been developed by both governmental and non-governmental entities

5.1 EU Regulations

5.1.1 GDPR

The General Data Protection Regulation (GDPR) addresses the issue of algorithmic fairness indirectly by providing a framework for the protection of individuals' personal data and establishing principles for responsible data processing. While the GDPR primarily focuses on data protection and privacy, its provisions can indirectly contribute to ensuring fairness in algorithmic decision-making processes.

Lawfulness, fairness, and transparency

One of the fundamental principles of the GDPR is that personal data must be processed lawfully, fairly, and transparently. This principle implies that algorithms and decision-making processes must not result in unfair discrimination against individuals based on protected characteristics such as race, gender, or religion.

Purpose limitation and data minimization

The GDPR requires that personal data be collected for specified, explicit, and legitimate purposes and not further processed in a manner that is incompatible with those purposes. By limiting the purposes for which data can be used, the GDPR helps prevent algorithms from being applied in ways that might result in unfair outcomes.

Data accuracy and accountability

The GDPR mandates that organizations processing personal data must take reasonable steps to ensure that the data is accurate and up-to-date. Additionally, organizations are required to implement appropriate technical and organizational measures to ensure accountability and demonstrate compliance with GDPR principles. This includes ensuring that algorithms used in decision-making processes are regularly audited and monitored for fairness and accuracy.

Right to explanation

Article 22 of the GDPR provides individuals with the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning them or similarly significantly affects them. Furthermore, individuals have the right to obtain meaningful information about the logic involved in automated decision-making processes and to challenge decisions made solely by algorithms.

Data protection impact assessments (DPIAs)

Under the GDPR, organizations are required to conduct DPIAs for processing activities that are likely to result in a high risk to the rights and freedoms of individuals. This includes assessing the potential risks associated with algorithmic decision-making, such as the risk of bias or discrimination, and implementing measures to mitigate these risks.

5.1.2 AI Act

The AI Act[22], proposed by the European Commission in April 2021, has undergone several revisions and amendments throughout its legislative process[23]

[22] https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence, last accessed 27/02/2024

[23] https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-regulation-on-artificial-intelligence, last accessed 27/02/2024

Introduced in April 2021, the Act aims to establish rules and standards for the development, deployment, and use of AI systems across various sectors. Key provisions include defining high-risk AI applications, imposing obligations on providers and users of such systems, ensuring transparency and accountability in AI processes, and establishing mechanisms for conformity assessment and enforcement.

Risk-based approach

The AI Act classifies AI according to its risk:

- Unacceptable risk, which is prohibited (e.g. social scoring systems and manipulative AI).
- High risk, the regulation of which isthe subject of most of the text:
- Limited risk, which is ubject to lighter transparency obligations: developers and deployers must ensure that end-users are aware that they are interacting with AI (chatbots and deepfakes).
- Minimal risk, which is unregulated (such as spam filters)

High-Risk Systems

The majority of obligations fall on providers (developers) of high-risk AI systems. High-risk AI systems encompass AI technologies utilized in various domains:

- Critical infrastructures, such as transportation, where malfunctions could jeopardize citizen safety.
- Educational or vocational training, influencing access to education and career paths, like exam scoring systems.
- Safety components of products, like AI in robot-assisted surgery, which impact patient well-being.
- Employment management, including CV-sorting software affecting recruitment fairness.
- Essential private and public services, such as credit scoring determining loan eligibility.
- Law enforcement, where AI may affect fundamental rights through evidence evaluation.
- Migration, asylum, and border control management, including document authenticity verification.
- Administration of justice and democratic processes, where AI aids in applying laws to specific cases.

In addition to these domains, a system is considered high risk if it is used as a safety component or a product covered by EU laws in Annex II of the Act and required to undergo a third-party conformity assessment, or if they engage in profiling individuals, i.e. if they automatically analyse personal data to evaluate different facets of a person's life, including their professional performance, financial status, health, preferences, interests, reliability, behavior, location, or movements.

Obligations

While algorithmic fairness is not explicitly mentioned as a standalone concept, the Act addresses related issues such as bias and transparency, which are key factors in mitigating discriminatory outcomes. Specifically, the Act requires developers of high-risk AI systems to adhere to transparency obligations, ensuring that users are aware of the AI's functionality, limitations, and potential biases. Moreover, developers are required to document their AI systems' design and training data, enabling authorities to assess the system's potential for bias or discriminatory outcomes.

5.2 Standards

The International Organization for Standardization[24] (ISO) is an independent, non-governmental international organization that develops and publishes international standards. The IEEE Standard Association [25] outlines best practices for identifying and addressing bias in AI systems and AI-aided decision-making, covering topics such as bias and fairness overview, sources of unwanted bias, bias assessment metrics, and strategies for bias treatment. These guidelines aim to enhance the accountability and transparency of AI systems across all phases of their lifecycle, from data collection and training to continual learning, design, testing, evaluation, and deployment. The document's scope encompasses a comprehensive examination of bias within AI systems, particularly concerning AI-aided decision-making, offering measurement techniques and methods to mitigate bias-related vulnerabilities effectively.

[24] `https://www.iso.org/home.html`, last accessed 27/02/2024

[25] https://standards.ieee.org/ is an organization within the Institute of Electrical and Electronics Engineers (IEEE) that develops global standards in various fields related to electrical engineering, electronics, and information technology. The International Electrotechnical Commission[26] (IEC) is a worldwide entity responsible for the development and publication of international standards encompassing electrical, electronic, and associated technologies. While ISO and IEC are recognized international standards organizations, IEEE SA is not a body formally authorized by any government, but rather a community.

5.2.1 ISO/IEC TR 24028

The aim of document ISO/IEC 24028:2020[27] is to analyze the factors influencing the trustworthiness of systems utilizing AI. The document provides an overview of existing approaches that can support trustworthiness in technical systems and discusses their potential application in AI systems. It also addresses possible approaches to addressing vulnerabilities in AI systems related to trustworthiness.

5.2.2 ISO/IEC 27001:2013

The document ISO/IEC 27001:2013[28], last accessed 27/02/2024

5.2.3 IEEE P7003

The IEEE P7003 Standard for Algorithmic Bias Considerations3[29] constitutes one of eleven IEEE ethics-oriented standards presently in progress as a component of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Its primary objective is to furnish individuals or organizations engaged in the development of algorithmic systems with a structured framework to prevent unintended, unjustified, and discriminatory outcomes for users.

5.3 Non-Binding Instruments

A non-binding instrument is a document or agreement that does not impose legally enforceable obligations on the parties involved.

5.3.1 EU Artificial Intelligence Ethics Checklist

On April 8, 2019, the High-Level Expert Group on Artificial Intelligence (AI HLEG) published a document titled *ETHICS GUIDELINES FOR TRUSTWORTHY AI*[30]. This document is not a framework directive or regulation of the EU and is therefore not legally binding. However, it provides principles and requirements for AI systems and serves as the basis for the aforementioned OECD agreement among states. The guidelines contain 7 Key Requirements that AI systems should meet to be considered trustworthy:

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination, and fairness
- Societal and environmental well-being
- Accountability

The document includes an assessment list covering the key requirements of ethical AI and provides guidance for their practical implementation.

5.3.2 Guidelines on Automated individual decision-making and Profiling

This document[31]addresses the implications of profiling and automated decision-making under the GDPR, recognizing their increasing prevalence across various sectors such as banking, healthcare, and marketing. While these practices offer benefits like increased efficiencies and tailored services, they also present risks to individuals' rights and freedoms,

[29] `https://ieeexplore.ieee.org/document/8452919`, last accessed 27/02/2024

[30] `https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai`, last accessed 27/02/2024

[31] `https://ec.europa.eu/newsroom/article29/items/612053`, last accessed 27/02/2024

including privacy infringement and unjust discrimination. The GDPR introduces provisions to mitigate these risks, emphasizing transparency and accountability. The guidelines clarify definitions, general provisions, and specific regulations regarding automated decision-making, highlighting the importance of data protection impact assessments and the role of data protection officers. it concludes with best practice recommendations and a commitment from the Article 29 Data Protection Working Party (WP29) to monitor implementation and potentially provide further guidance.
Specifically, the Annex lists good practices including the following recommendations, which while not exhaustive, offer valuable guidance for controllers aiming to ensure algorithmic fairness in solely automated decisions, including profiling as defined in Article 22(1):

- Conduct regular quality assurance checks on systems to guarantee equitable treatment of individuals, irrespective of special categories of personal data or other factors.
- Implement algorithmic auditing procedures, involving rigorous testing of machine learning algorithms to verify their intended functionality and prevent the generation of discriminatory, erroneous, or unjustified outcomes.
- In cases where decision-making based on profiling significantly impacts individuals, facilitate independent third-party audits by providing auditors with comprehensive insights into the workings of the algorithm or machine learning system. These measures aim to enhance transparency, accountability, and the elimination of biases in automated decision-making processes.

References

[1] Kiana Alikhademi, Emma Drobina, Diandra Prioleau, Brianna Richardson, Duncan Purves, and Juan E Gilbert. 2021. A review of predictive policing from the perspective of fairness. *Artificial Intelligence and Law* (2021), 1–17.

[2] Alhanoof Althnian, Duaa AlSaeed, Heyam Al-Baity, Amani Samha, Alanoud Bin Dris, Najla Alzakari, Afnan Abou Elwafa, and Heba Kurdi. 2021. Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain. *Applied Sciences* 11, 2 (2021).

[3] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software engineering for machine learning: A case study. *IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)* (2019), 291–300.

[4] Yael Amsterdamer and Tova Milo. 2015. Foundations of Crowd Data Sourcing. *ACM SIGMOD Record* 43, 4 (feb 2015), 5–14.

[5] Daniel W Apley and Jingyu Zhu. 2020. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82, 4 (2020), 1059–1086.

[6] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. 2009. Methodologies for Data Quality Assessment and Improvement. *Comput. Surveys* 41, 3, Article 16 (jul 2009), 52 pages.

[7] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* 50, 1 (Feb. 2021), 3–44. https://doi.org/10.1177/0049124118782533 Publisher: SAGE Publications Inc.

[8] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).

[9] Jeffrey P. Brantingham, Matthew Valasik, and George O. Mohler. 2018. Does Predictive Policing Lead to Biased Arrests? Results from a Randomized Controlled Trial. *Statistics and Public Policy* 0, ja (2018), 0. https://doi.org/10.1080/2330443X.2018.1438940

[10] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 2053951715622512.

[11] Katzenbach C. and Ulbricht L. 2019. Algorithmic governance. *Internet Policy Review* 8, 4 (2019).

[12] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems* 30 (2017).

[13] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.

[14] Simon Caton and C. Haas. 2020. Fairness in Machine Learning: A Survey. *ArXiv* abs/2010.04053 (October 2020).

[15] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 839–847.

[16] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. 2019. *This Looks like That: Deep Learning for Interpretable Image Recognition*. Curran Associates Inc., Red Hook, NY, USA.

[17] Silvia Chiappa. 2019. Path-specific counterfactual fairness. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 7801–7808.

[18] John Danaher, Michael J Hogan, Chris Noone, Rónán Kennedy, Anthony Behan, Aisling De Paor, Heike Felzmann, Muki Haklay, Su-Ming Khoo, John Morison, Maria Helen Murphy, Niall O'Brolchain, Burkhard Schafer, and Kalpana Shankar. 2017. Algorithmic governance: Developing a research agenda through the power of collective intelligence. *Big Data & Society* 4, 2 (2017).

[19] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[20] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 259–268. `https://doi.org/10.1145/2783258.2783311`

[21] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.

[22] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems* 14, 3 (1996), 330–347. `https://doi.org/10.1145/230538.230561`

[23] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[24] Hector Garcia-Molina, Manas Joglekar, Adam Marcus, Aditya Parameswaran, and Vasilis Verroios. 2016. Challenges in data crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering* 28, 4 (2016), 901–911.

[25] 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.

[26] Tarleton Gillespie. 2014. The Relevance of Algorithms. In *Media Technologies*, Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot (Eds.). MIT Press Scholarship Online, 167–194.

[27] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862* (2019).

[28] Ankur Handa, Viorica Pătrăucean, Simon Stent, and Roberto Cipolla. 2016. Scenenet: An annotated model generator for indoor scene understanding. *2016 IEEE International Conference on Robotics and Automation (ICRA)* (2016), 5737–5743.

[29] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[30] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2020. The dataset nutrition label. *Data Protection and Privacy, Volume 12: Data Protection and Democracy* 12 (2020), 1.

[31] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need? *Proceedings of the 2019 CHI conference on human factors in computing systems* (2019), 1–16.

[32] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021), 560–575.

[33] Heinrich Jiang and Ofir Nachum. 2020. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 702–712.

[34] Brendan Juba and Hai S Le. 2019. Precision-recall versus accuracy and the role of large data sets. 33, 01 (2019), 4039–4048.

[35] Michael G Kahn, Marsha A Raebel, Jason M Glanz, Karen Riedlinger, and John F Steiner. 2012. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Medical care* 50 (2012).

[36] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems* 33, 1 (2012), 1–33.

[37] Rob Kitchin. 2016. Thinking critically about and researching algorithms. *Information, Communication & Society* 20, 1 (2016), 14–29. https://doi.org/10.1080/1369118X.2016.1154087

[38] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 world wide web conference*. 853–862.

[39] Bartosz Krawczyk. 2016. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence* 5, 4 (2016), 221–232.

[40] Reinhard Kreissl. 2014. Assessing security technology's impact: Old tools for new problems. *Science and Engineering Ethics* 20, 3 (2014), 659–673. https://doi.org/10.1007/s11948-014-9529-9

[41] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).

[42] Kimmo Kärkkäinen and Jungseock Joo. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2021), 1547–1557.

[43] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2020. Explainable ai: A review of machine learning interpretability methods. *Entropy* 23, 1 (2020), 18.

[44] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x arXiv:https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1740-9713.2016.00960.x

[45] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[46] Monique Mann and Tobias Matzner. 2019. Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to emergent discrimination. *Big Data & Society* 6, 2 (2019). https://doi.org/10.1177/2053951719895805

[47] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54, 6 (2021), 1–35.

[48] Frank Pasquale. 2015. *The black box society: The secret algorithms that control money and information*. Harvard University Press, Cambridge.

[49] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. 2016. The synthetic data vault. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (2016), 399–410.

[50] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (2021), 100336.

[51] Evaggelia Pitoura. 2020. Social-Minded Measures of Data Quality: Fairness, Diversity, and Lack of Bias. *J. Data and Information Quality* 12, 3, Article 12 (jul 2020), 8 pages.

[52] Joseph Prusa, Taghi M. Khoshgoftaar, and Naeem Seliya. 2015. The Effect of Dataset Size on Training Tweet Sentiment Classifiers. (2015), 96–102.

[53] Naresh Sundar Rajan, Ramkiran Gouripeddi, Peter Mo, Randy K. Madsen, and Julio C. Facelli. 2019. Towards a content agnostic computable knowledge repository for data quality assessment. *Computer Methods and Programs in Biomedicine* 177 (2019), 193–201.

[54] Franck Ravat and Yan Zhao. 2019. Data Lakes: Trends and Perspectives. *Database and Expert Systems Applications* (2019), 304–313.

[55] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? *International Conference on Machine Learning* (2019), 5389–5400.

[56] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[57] Rashida Richardson, Jason Schultz, and Kate Crawford. 2019. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. *94 N.Y.U. L. REV.* Online 192 (2019), 30.

[58] Yuji Roh, Geon Heo, and Steven Euijong Whang. 2021. A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective. *IEEE Transactions on Knowledge and Data Engineering* 33, 4 (2021), 1328–1347.

[59] 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. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), 1–15.

[60] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[61] Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data* 6, 1 (2019), 1–48.

[62] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2239–2248.

[63] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. 2015. Striving for Simplicity: The All Convolutional Net. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.).

[64] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 1 (2019), 4149–4158.

[65] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E Hines, John P Dickerson, and Chirag Shah. 2020. Counterfactual explanations and algorithmic recourses for machine learning: A review. *arXiv preprint arXiv:2010.10596* (2020).

[66] Sahil Verma, John Dickerson, and Keegan Hines. 2021. Counterfactual explanations for machine learning: Challenges revisited. *arXiv preprint arXiv:2106.07756* (2021).

[67] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. 1–7. `https://doi.org/10.1145/3194770.3194776`

[68] Bertie Vidgen and Leon Derczynski. 2021. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS ONE* 15, 12 December (2021).

[69] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. *Caltech-UCSD Birds-200-2011*. Technical Report CNS-TR-2011-001. California Institute of Technology.

[70] Sirui Yao and Bert Huang. 2017. Beyond Parity: Fairness Objectives for Collaborative Filtering. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. `https://proceedings.neurips.cc/paper_files/paper/2017/file/e6384711491713d29bc63fc5eeb5ba4f-Paper.pdf`

[71] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833.

[72] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning*. PMLR, 325–333.

[73] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.

[74] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2921–2929.

[75] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics* 10, 5 (2021), 593.

[76] Indrė Žliobaitė. 2017. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery* 31, 4 (2017), 1060–1089.

Disclaimer

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.861091. This document reflects the views only of the authors, and the European Union cannot be held responsible for any use which may be made of the information contained therein.