Fair AI: State-of-the-art overview of the literature

TNO Publiek ⟩ TNO 2023 R10060
December 2022

Artificial Intelligence
www.tno.nl
+31 88 866 00 00
info@tno.nl

TNO 2023 R10060 – December 2022

Fair AI: State-of-the-art overview of the literature

Author(s)	Wouter Korteling, Romy van Drie, Cor Veenman
Classification report	TNO Publiek
Report text	TNO Publiek
Number of copies	1
Number of pages	29 (excl. front and back cover)
Number of appendices	0
Project name	AI Innovatie Oversight Lab
Project number	060.49629

Contents

1 Introduction

Advances in artificial intelligence (AI) and machine learning (ML) have resulted in the development of increasingly powerful algorithms that learn to identify complex patterns in data that relate to specific outcomes. AI models that are derived using such algorithms are widely being applied in various layers of society, such as finance, healthcare and education, in business as well as governmental bodies. These algorithms have sometimes been falsely appraised with the quality of providing “neutral” decisions that are void of human bias. The presumption is known as the “neutrality fallacy” (Sandvig 2014). AI algorithms are vulnerable to various forms of bias that may sometimes lead to unfair decision-making. A well-known example of an AI model that was deemed unfair is the COMPAS tool (Angwin, et al. 2016). COMPAS is used by courts in the US to make parole decisions. It calculates risk scores indicating whether a person is likely to re-offend. The algorithm was criticized by ProPublica of assigning a higher recidivism risk score to African-Americans than to Caucasians with a similar profile. Since the report on COMPAS came out in 2016, there has been a lot of attention and debate on the topic of fairness in AI models. As a result, various guidelines, algorithms, metrics and toolkits have been developed that can help data scientists in the field to increase the fairness of their AI applications. In order to help data science practitioners navigate their way through the literature, the current report provides an overview and brief description of several of these state-of-the-art documents and toolkits that exist in the field of fair AI. Hence, this report is not intended as a guideline or handbook to be used for implementing fair AI solutions, but aims to provide an overview of the literature. Since this overview is not exhaustive, we attempt to direct the reader towards other sources that provide additional information on specific topics.

The report is organized as follows. The *Fairness in AI* chapter describes some of the notions of fairness used within AI literature, as well as related concepts such as prejudice and bias. The *Fairness Metrics* chapter explains some of the metrics that can be used by data scientists to provide an indication of the fairness level in the AI model. Here, we also provide some of the insights from literature on fair AI about possible considerations that can be made when choosing a suitable fairness metric in an AI project. The *Fairness Algorithms* chapter provides some examples of algorithms that exist to mitigate bias in an AI model or in training data. The final two sections describe several sources of software programs and documents that are intended to facilitate in the actual implementation of fair AI models. Here, we follow the distinction as made in (Madaio, et al. 2020) between technical software tools and guidelines/checklists. The *Fair AI: Tools* chapter describes some of the software tools that contain “off-the-shelf” implementations of fairness algorithms and metrics that can be used in AI projects. The *Fair AI: Handbooks and Checklists* chapter describes handbooks and checklists that provide process-based and more high-level suggestions and considerations that can be made when developing fair AI models.

2 Fairness in AI

2.1 Fairness

The predictions made by AI models that are used for decision support are likely to have direct or indirect implications for people's lives. Whenever humans make decisions that affect the lives of others, we want these decisions to be "fair". This is especially the case in situations where the consequences of a decision have a high impact. Likewise, most AI models or tools that are used for decision support, such as in governmental oversight bodies, will directly or indirectly influence the lives of others. We therefore want these AI models to be fair as well. In this section we describe some of the notions of fairness and "bias" as used by researchers working on fairness in AI.

To illustrate what fairness might mean in everyday life, consider the following hypothetical scenario of a governmental body that performs oversight on companies operating containerships that may or may not violate the law.

Containership scenario:

In order to detect violations by companies operating containerships, an oversight body has to perform inspections on these containerships. Given that the inspections are labor and time intensive, the body uses an AI model that categorizes containerships into high and low risk profiles and this categorization is subsequently used as input to the decision of whether a containership will be inspected. Since the oversight body does not want to discriminate based on the place of incorporation of the company, the body seeks to treat companies that are located in Western countries similarly to companies located in non-Western countries. Given that the decisions by the oversight body will be affected by the AI model, the oversight body requires that their AI model treats these companies similarly as well.

In this example, the oversight body makes an initial judgement about potential discrimination between Western and non-Western companies. Namely, the body identifies two groups that should be treated similarly, otherwise the body deems their inspection strategy unfair. For example, if the oversight body finds that their AI model assigns risk profiles based on the company's place of incorporation, the body may conclude that their AI model treats companies unfairly.

To create clarity on the issue of fairness in AI, researchers have posed definitions of fairness in AI. For example, (Mehrabi, et al. 2021) defines **fairness** in decision making as:

"The absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics".

Fairness is therefore related to concepts of prejudice and favoritism, where prejudice can be regarded as an unfavorable opinion that is formed beforehand or without sufficient knowledge, reason or thought, and favoritism as the opposite.

Within literature on fairness in AI, a distinction is often made between *individual* fairness and *group* fairness. For example, (Dwork, et al. 2012) and (Mehrabi, et al. 2021) define **individual fairness** as "*treating similar individuals similarly"*. Here, the similarity of individuals is based on characteristics that are regarded as relevant for a specific task context. This means that individuals may be regarded as similar in one context, but not in another. Consequently, according to this individual fairness definition, these individuals should be treated similarly in the one context, but not necessarily in the other. **Group fairness**, on the other hand, is often considered as *"treating different groups similarly"* (Mehrabi, et al. 2021). For example, demographic characteristics of individuals, such as gender, may be regarded as irrelevant (or even illegal) to use as a basis of distinguishing between individuals in a certain task context. As (Binns 2020) outlines, both conceptualizations of fairness rely on similar high-level principles and are therefore fundamentally aligned with each other. For example, both fairness concepts are concerned with separating relevant (or fair) characteristics from irrelevant (or unfair) characteristics to distinguish between individuals. However, whereas individual fairness starts from the perspective of asking "which characteristics are fair/relevant for making a distinction?", group fairness starts from the perspective of asking "which characteristics are *not* fair/relevant for making a distinction?". For example, for a specific task, features x1, x2, and x3 may be regarded as *relevant* and individuals that are similar within these features, should be treated similarly in order to satisfy individual fairness. To satisfy group fairness, the approach is usually to specify a feature, such as a demographical variable d1, that is considered *irrelevant* (or illegal) in the current context as a basis to distinguish between individuals, and treat individuals in the different demographical groups similarly. As a result, the paradigms and methodologies to measure fairness of AI models can roughly be divided along this dichotomy. As (Binns 2020) explains, the "blunt" application of both types of fairness criteria often leads to apparent contradictions between both metric types. We refer the reader to (Binns 2020) for a philosophical analysis of individual and group fairness in the context of machine learning. In section Fairness Metrics of the current document, we describe some of the group fairness metrics that have been defined.

2.2 Bias

Another concept that is linked to fairness in AI is *bias*, which, in literature on AI fairness, is sometimes considered to be the *source* of unfairness (e.g. Mehrabi et al, 2019), and sometimes as the skewness in the outcomes/predictions of the AI model with respect to irrelevant characteristics of individuals (i.e. a *form* of unfairness in an AI model, or perhaps even a synonym) (e.g. in (Muhammad 2022)). In the current document, we use bias to reflect the former conceptualization, namely:
Any structural tendency in the training data, the model or use of the model that leads to unfair decisions.

As this definition implies, there exist multiple types of bias that may introduce unfairness in an AI model at different levels, from the data that is used to optimize the parameters of the AI model, to how the model is used in practice when making decisions. An example of bias is when an algorithm is supposed to model some aspect in the current world, but the data that is used to train the model still reflects historical societal tendencies that are currently not applicable anymore. For example, if non-

Western companies operating containerships have more often violated regulations than Western companies operating containerships in the past, an AI model may learn this historic tendency, even if it is currently not applicable anymore. This source of unfairness is often referred to as “historical bias”. For an overview of various types of bias in machine learning, we refer the reader to (Mehrabi, et al. 2021).

3 Measuring Bias: Fairness Metrics

In order to assess whether an AI model is biased, one needs to perform some type of evaluation. A straightforward way to infer whether an AI model uses “irrelevant” characteristics of individuals in its decisions is to check whether such characteristics are included as input features to the model. These features may then simply be left out of the feature space of the model. One of the reasons why this may not be sufficient, however, is due to the possibility of input features being used as “proxies” to other features. Proxies are (combinations of) features that statistically relate to other features. A famous example is the city postal codes of an individual’s residency that may correlate with the individual’s ethnicity. Another reason why it may be hard to assess whether an AI model bases its decisions on relevant information, is due to the complexity of many of the modern AI models. The internal “reasoning” underlying an AI decision may remain vague to a human interpreter. For this purpose, data scientists have defined metrics that attempt to compute the level of fairness based on the *output* of the model, as opposed to the reasonings that are internal to the model. In this chapter, we first describe some of these fairness metrics and subsequently discuss some of the considerations suggested in literature that can be made on how to choose an appropriate fairness metric.

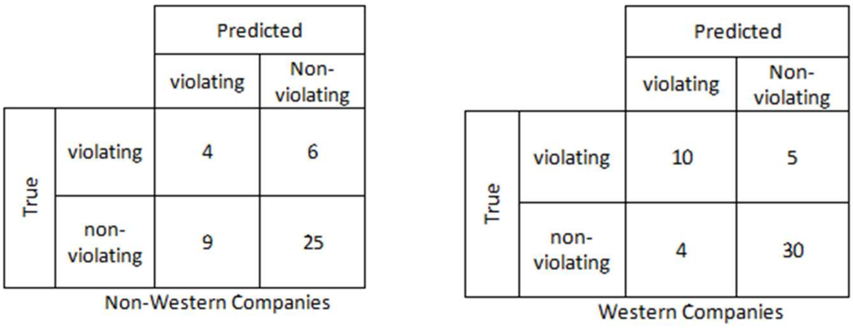
3.1 Fairness Metrics

Most of the attention in fairness research has been devoted to measuring group fairness. One of the reasons is that it is hard to specify the properties to use for measuring similarity. Another possible reason is that metrics that are designed to measure individual fairness usually rely on distance measures that are often hard to interpret. In addition, it is more straightforward to use group fairness metrics to prevent discriminating between different demographic groups, which is often an ethical and legal concern in practice. In this section, we therefore focus on metrics designed to measure group fairness.

		Predicted	
		$\hat{y} = 1$	$\hat{y} = 0$
True	$y = 1$	True Positives (TP)	False Negatives (FN)
	$y = 0$	False Positives (FP)	True Negatives (TN)

Table 1. Confusion Matrix. y is the actual (true) class to which an individual belongs. $\hat{y}$ is the predicted class to which the individual belongs.

Group fairness metrics commonly rely on measurements derived from a *confusion matrix*. A confusion matrix shows the level of “confusion” of a model between two classes in a binary classification task. Table 1 depicts the metrics in a confusion matrix, where True Positives (TPs) are correctly classified instances belonging to the positive class, False Negatives (FNs) are incorrectly classified instances belonging to the positive class, False Positives (FPs) are incorrectly classified instances belonging to the negative class and True Negatives (TNs) are correctly classified instances belonging to the negative class.



		Predicted	
		violating	Non-violating
True	violating	4	6
	non-violating	9	25

Non-Western Companies

		Predicted	
		violating	Non-violating
True	violating	10	5
	non-violating	4	30

Western Companies

Table 3. Hypothetical confusion matrices for an AI model predicting law violation for Western and non-western companies operating containerships.

When interested in fairness with respect to two sensitive subgroups, such as Western and non-Western companies, a separate confusion matrix can be defined for the two subgroups (e.g. see Table 3). Most group fairness metrics compare performance using measures derived from the resulting two confusion matrices. Metrics that are commonly described in literature are *demographic parity*, *equality of opportunity, equalized odds, calibration and predictive parity*. These are briefly described below.

Demographic parity (or statistical parity) is satisfied if the favorable and unfavorable predictions are assigned to the sensitive subgroups at similar rates. That is, the Positive Rates (PRs) and Negative Rates (NRs)[1] are equal between all sensitive subgroups. For example, ship owners from Western countries are categorized as “high risk” at an equal rate as shipowners from non-Western countries. Hence, this is irrespective of the percentage of actual violators in both groups. As such, the observed predictions are statistically independent of the protected attribute. Figure 1 gives a visual example of demographic parity. Out of all predictions given, members of group A receive a positive prediction as often as members of group B (4 out of 8 times), satisfying demographic parity.

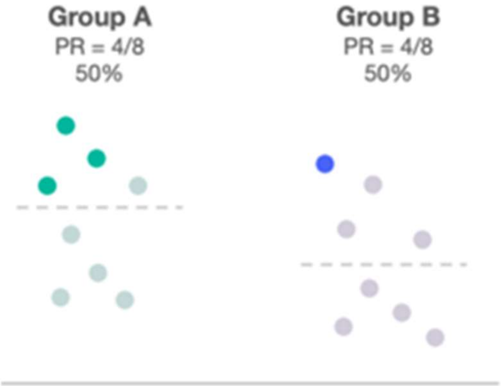


Figure 1. Demographic parity: an equal positive rate (PR) across groups. The dotted line represents the decision boundary. Data points above the dotted line have received a positive prediction. The color of the data points represents the actual labels. Bright colors indicate a positive, while muted colors indicate a negative. Figure adapted from (Cortez 2019).

Equality of opportunity is satisfied if the positive outcome is assigned at an equal rate between both groups for all instances for which the positive outcome is correct. Hence, it conditions on the true label, as it is about having equal True Positive Rates (TPRs) for the separate subgroups. For example, if a law violator from a Western country is equally likely to be checked as a violator from a non-Western country, then equality of opportunity is satisfied. Figure 2 shows an example of equality of opportunity. Here, the ratio of true positives is similar across the groups: half of all actual positives are indeed predicted as positives in both groups.

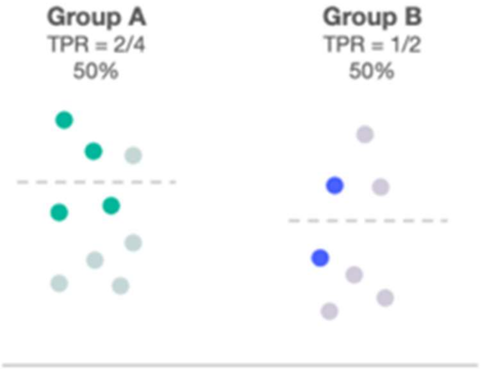


Figure 2. Equality of opportunity: an equal true positive rate (TPR) across groups. The dotted line represents the decision boundary. Data points above the dotted line have received a positive prediction. The color of the data points represents the actual labels. Bright colors indicate a positive, while muted colors indicate a negative. Figure adapted from (Cortez 2019).

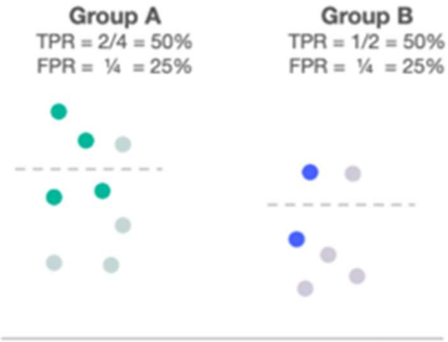


Figure 3. Equalized odds: an equal true positive rate (TPR) and an equal false positive rate (FPR) across groups. The dotted line represents the decision boundary. Data points above the dotted line have received a positive prediction. The color of the data points represents the actual labels. Bright colors indicate a positive, while muted colors indicate a negative. Figure adapted from (Cortez 2019).

Equalized odds is similar to equality of opportunity, but stricter. In addition to having an equal True Positive Rate, an equal False Positive Rate is required to satisfy equalized odds. For example, not only should a violator from a Western country be equally likely to be assigned a the high risk label as a violator from a non-Western country, but a non-violator from a Western country should be equally likely assigned a low-risk label as a non-violator from a non-Western country as well. In Figure 3 an example is given where equalized odds is satisfied. Note that in addition to having an equal TPR, as shown in Figure 2, we can also observe an equal FPR as 1/4 of all negatives are falsely predicted as being positive.

Calibration conditions on the labels as predicted by the model, as opposed to conditioning on the true labels, such as in equality of opportunity. Calibration is satisfied if, for instances predicted as being positive, the ratio of actual positives is equal between both sensitive subgroups. In addition, the same should hold for instances predicted as being negative, where, for instances predicted as being negative, the ratio of actual negatives is equal between both sensitive subgroups. This means that both Positive Predicted Values (PPVs) and Negative Predicted Values (NPV) should be equal between both subgroups. Figure 4 shows a scenario in which this is the case.

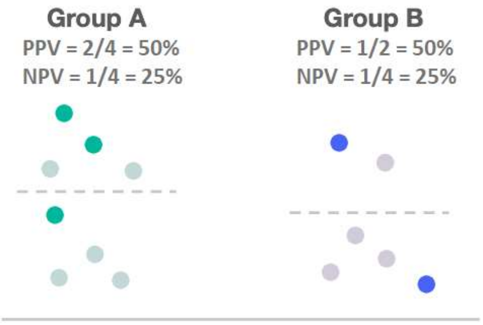


Figure 4. Calibration: an equal Positive Predictive Value (PPV) and an equal Negative Predicted Value (NPV) across groups. The dotted line represents the decision boundary. Data points above the dotted line have received a positive prediction. The color of the data points represents the actual labels. Bright colors indicate a positive, while muted colors indicate a negative.

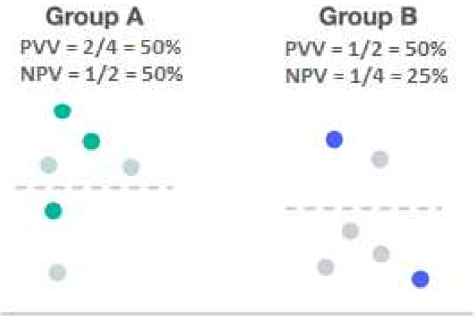


Figure 5. Predictive parity: an equal Positive Predictive Value (PPV). The dotted line represents the decision boundary. Data points above the dotted line have received a positive prediction. The color of the data points represents the actual labels. Bright colors indicate a positive, while muted colors indicate a negative.

Predictive parity is equal to calibration but weaker, as only the PPVs should be equal across both subgroups. Figure 5 shows a scenario in which predictive parity is satisfied.

3.2 Choosing and Interpreting a Metric

In most real world scenario's, not all fairness metrics can be satisfied simultaneously. For example, in case the "base-rate" (i.e. the ratio of actual positive instances out of all instances) differs between the classes in the sensitive attribute, it is impossible for any categorization model to satisfy measures that condition on the actual classes (e.g. equality of opportunity) and measures that condition on the predicted classes (e.g. calibration) at the same time. This naturally raises the question on how to choose an appropriate metric.
Here, we highlight some of the considerations that can be made as suggested in two sources that describe a decision model for choosing an appropriate fairness metric, namely the Fairness Compass proposed in (Ruf and Detyniecki 2021) and the Fairness Tree proposed in the Acquitas Tool[2] and described in (Muhammad 2022). We describe two of the principal considerations suggested in both sources, namely whether one is interested in *equality in outcome representation or equality in prediction errors* and

whether predictions are used to decide on something that is *desirable or undesirable* from the perspective of the individual.

The first consideration that needs to be made according to the sources is whether the developer preferably desires equality in outcome representation or equality in prediction errors between both groups. For example, if the oversight body is only interested in an equal proportion of containers being labeled as low risk between both Western and non-Western countries, a measure that captures equality in predictions, such as demographic parity, might be relevant. This may possibly be applicable in cases in which one wants to “correct” for selection bias that is present in the data. For example, in case companies from non-Western countries operating containerships have been more thoroughly and frequently inspected in the past, the available training data may reflect this skewness with respect to country origin. In this case, you may want to correct for such measurement errors and simply assume that the distributions between containerships from both countries should be equal. However, as explained in (Muhammad 2022), in case one assumes there are valid reasons for the found disparities between the subgroups that should be taken into account, equality of representation metrics may not be desirable. Namely, when using these metrics for algorithm optimization, the algorithm is forced to treat the different groups differently such as to achieve the same outcome ratios. In case there are valid reasons for the disparities, these reasons will then be ignored by the algorithm. This leads to discrimination effects, as it “causes otherwise similar people to be treated differently” (Muhammad 2022). In case the base-rate between the different groups differs due to causes that are considered valid (i.e. due to “relevant characteristics” from a fairness perspective), it may be better to opt for an error-based metric, since these take into account the actual outcome representations (i.e. base-rates) in the groups.

When then opting for an error-based metric, one of the principal considerations to be made is whether the model’s output is used to decide on something that is desirable or undesirable (or “rewarding” or “punitive”) from the perspective of the individuals in the subgroups. Namely, when the algorithm is used to decide on whether to give an individual something that is undesirable, such as a financial punishment, fairness metrics that condition on the algorithm’s predictions (e.g. calibration) are probably better suited for the application at hand. The reason being that it is likely that unjustified penalties or punishments have a higher impact on the perceived fairness than missed penalties or punishments. On the other hand, when the model is used to decide on whether to give an individual something that is desirable, such as whether an individual deserves a financial benefit, metrics that condition on the actual classes (e.g. equality of opportunity) are likely to be more relevant. The assumption being that not giving individuals a reward or providing them assistance when they actually deserve the reward is likely to have a higher impact on the perceived fairness than giving individuals a reward that did not deserve the reward. In our example of companies operating containerships, the oversight body uses the model to decide on whether a containership should be inspected or not. From the perspective of the company operating the containerships, an inspection is undesirable (e.g. due to a time investment). Following the reasoning in the two sources, a metric that conditions on the predictions is therefore more probable to align with our conception of what is fair, as it ensures that Western and non-Western companies operating containerships that are obeying the rules are not being inspected at unequal rates. For safety inspections on a ship, on the other hand, it is important to warn shipowners in case there is an important safety problem. Receiving a warning when there is actually no safety issue does not have a large impact and such false alarms are therefore less likely to evoke a feeling of unfairness.

It is important to note that the metrics of fairness should most often not be interpreted as an exact measurement of what is considered to be fair. Notions of fairness typically relate to the process that precedes and determines a decision. For example, discrimination is concerned with *unjustified* distinguishing between groups and prejudice relates to opinions based on *insufficient knowledge or thought*. Fairness may therefore not always be reflected in the decision-outcomes themselves. Given that the fairness metrics described in this section are solely based on the decision-outcomes across groups, they should better be considered as indicators of fairness, and not as mathematical "definitions" of fairness. Hence, one should be aware that it may not be sufficient to optimize the model on a specific fairness metric, as the model may be unfair in ways that are not captured by the metric.

4 Solutions to Bias: Fairness Algorithms

In order to mitigate the problem of bias, researchers have developed fairness algorithms. These algorithms can be divided into three categories: preprocessing, in-processing, and postprocessing algorithms. In this section, we describe these algorithm categories and for each category provide some examples of algorithms that exist.

4.1 Preprocessing algorithms

Preprocessing methods manipulate the data that is used as input to the model. Examples include: **Optimized Preprocessing** (Calmon, et al. 2017)**, Disparate Impact Remover** (Feldman, et al. 2015), **Reweighing** (Kamiran and Calders 2012), **Massaging** (Kamiran and Calders 2012) and **Learning Fair Representations** (Zemel, et al. 2013). The principal difference between the existing preprocessing techniques is whether the algorithm affects the *outcome label*, the *features* used to predict the outcome label or applies a sampling technique in the training dataset.

An example of an algorithm that affects the outcome labels is **Massaging**. Massaging works by changing the labels of some of the instances in the training dataset such that distribution of the desirable and undesirable classes in the different sensitive subgroups becomes more equal. The instances in which the output label is changed are not picked at random, but a ranking algorithms is used that calculates the probability of the instances of belonging to the other class. The outcome label of instances that are likely to belong to the other class is then altered. This can be regarded as a considerably radical bias mitigation procedure since the actual outcome labels are changed. It can be expected that it has a relatively high impact on the predictive performance of the model on new instances, since these instances will not have similarly modified outcome labels.

An example of an algorithm that alters the predictive features is **Disparate Impact Remover**. Disparate Impact Remover first defines the “disparate impact”, which is the proportion of individuals in the unprivileged group (the protected group) that received the positive outcome divided by the proportion of individuals in the privileged group that received the positive outcome. Hence, a disparate impact of 1 illustrates a dataset where the outcome is equally represented over both groups and a disparate impact below 1 suggest an unequal distribution. Given there is likely to exist a trade-off between the accuracy of the model and permutated data to reduce bias, a desired or “acceptable” disparate impact is selected, such as 0,8. Using an algorithm that is detailed in (Feldman, et al. 2015), the predictive features are then altered in such a way that this disparate impact is obtained. Further, the algorithm restricts on rank preservation, which means that the ranking of the individuals on the different features (i.e., from low values to high values), is maintained after the algorithm has been applied.

An example of a preprocessing bias mitigation algorithm that does not affect the outcome labels or the predictive features is **Reweighing**. Reweighing works by adopting a sampling procedure such that the under and overrepresentation of a binary outcome label in two sensitive subgroups is equalized. It works by computing weights that determine the likelihood of different instances to be sampled. Hence, it affects the representation of positive and negative instances in the sensitive subgroups in the sampled training dataset, but not any of the predictive feature or outcome label values corresponding to those instances.

4.2 In-processing algorithms

In-processing methods target the model while it is training to reduce bias in the model. On a technical level, the primary difference between the different algorithms that we encountered relate to two different approaches to optimize for fairness. The first approach consists of algorithms that combine the solutions proposed by a predictor model that optimizes for accuracy and a fairness model that optimizes for a fairness metric to create a solution that is optimized for both objectives (the combination of objectives may be regarded as a “meta” objective). These algorithms then differ in the way that the solutions of both algorithms are obtained and combined. The other general approach is to apply a regularization term in the objective function that constraints the model to find solutions that are not only accurate but also fair. On a practical level, the different algorithms may be appropriate for different types of models and fairness metrics. Below we provide some of the algorithms that we found.

Adversarial Debiaser is an algorithm proposed in (Zhang, Lemoine and Mitchell 2018) that builds upon the idea of adversarial models presented in (Goodfellow, et al. 2014) . Adverserial Debiaser simultaneously trains a predictor model with the objective to maximize prediction accuracy and an adversary model with the objective to minimize the ability to determine the protected attribute from the predictions of the predictor. In the ideal situation, the adversary would not be able to predict the protected attribute based on the predictions of the predictor model. The predictive power of the adversarial model is then used to alter the parameters in the predictor model, such that the predictive performance of the adversarial model is reduced. This forces the predictor model to find a new optimum that does not directly or indirectly utilize co-variance between the predictor features and the sensitive attribute.

Gerry Fair (Kearns, et al. 2018) learns a classifier that is optimized on fairness with respect to multiple protected attributes or a combination of protected attributes. The authors of the paper argue that approaches focusing on pre-defined groups such as race or gender are susceptible to “fairness gerrymandering[3]”. According to the chosen definitions of fairness, the classifier may seem fair for individual groups, but it may still be unfair with respect to certain subgroups, such as subgroups that have a specific combination of protected attribute values (sometimes referred to as “intersectional fairness”). The authors propose two algorithms to solve this problem, of which Gerry Fair is one. Gerry Fair is based on fictitious play learning rules derived from the field of game theory. Here, two "players" attempt to find an optimal solution for a different objective, namely one optimizing for accuracy (or loss reduction) and the other optimizing for a group fairness metric. The different subgroups that are taken into account by the Gerry Fair algorithm when optimizing for group fairness are simply all possible combinations of the protected attribute categories. The authors show that their algorithm can be used for demographical parity, false positive rate and false negative rate group fairness metrics. These are then implemented using a one-versus-the-others approach for each separate subgroup. The advantage of this algorithm is that it allows for taking fairness considerations with respect to multiple possible protected attributes into account.

Similar to Gerry Fair, **Exponentiated Gradient Reduction** (Agarwal, Beygelzimer, et al. 2018) adopts the paradigm of using two "players" that optimize a different objective. An exponentiated gradient function is used to find the Lagrange multipliers that result in the most accurate and fair solution. It supports at least regression, and several fairness notions (demographic parity, true positive rate and error rate).

Grid Search Reduction (Agarwal, Beygelzimer, et al. 2018, Agarwal, Dudík and Wu, Fair Regression: Quantitative Definitions and Reduction-based Algorithms 2019) is equal to the Exponentiated Gradient Reduction algorithm, except that it uses grid search to find the optimal solution as opposed to a gradient function. Hence, if the full grid is searched, it is guaranteed to find the optimal solution, but may likely take more time to run.

Meta Fair Classifier (Celis, et al. 2018) learns a regularization term that constraints the classifier to learn to generate predictions that do not only optimize accuracy (decrease loss), but fairness with respect to the specified fairness metric as well. The algorithm can be used to optimize classification models on a wide range of group fairness metrics. Further, like Gerry Fair, it can take group fairness with various subgroups based on multiple attributes into account.

Prejudice Remover (Kamishima, et al. 2012) aims to decrease "indirect prejudice", which is defined as a statistical relationship between a sensitive feature and the target label. Like Meta Fair Classifier, Prejudice Remover adds a regularization term that controls the trade-off between accuracy and indirect prejudice. The algorithm defines a "prejudice index", which contains the mutual information between the sensitive attribute and the target feature. The regularizer directly minimizes this index by penalizing high prejudice indices.

4.3 Postprocessing algorithms

Postprocessing algorithms attempt to de-bias the output of a model by adapting the model's output such that it satisfies the constraints implied by a specific fairness metric. Hence, these algorithms do not alter the data or the model, but instead modify the output directly.

Equalized Odds Postprocessing (Hardt, Price and Srebro 2016) changes output labels in order to optimize for equality of odds. Specifically, it makes sure that neither false positives nor false negatives disproportionately affect values of a sensitive attribute. The algorithm only uses true labels and predicted labels, so it can be used as post-processing after any classifier. Equalized odds can in theory be applied to both categorical and continuous outcomes.

Calibrated Equalized Odds Postprocessing (Pleiss, et al. 2017) can be understood as an extension of Equalized Odds, the extension being that the algorithm takes into account calibration. Calibration is about the actual output and the expected output given by a system. A model is well calibrated if for a set of people (or datapoints in general) that have received a predicted probability *p*, a *p* fraction of this set are positive instances of the classification problem. Perhaps even more importantly, “if we are concerned about fairness between two groups G1 and G2 (e.g. African-American defendants and white defendants) then we would like this calibration condition to hold simultaneously for the set of people within each of these groups as well”. According to the authors, calibration and equalized odds are mutually exclusive in most cases. In some cases it is possible with a relaxed calibration, and can be achieved with post-processing. The goal of the algorithm is to find the optimal solution for satisfying fairness constraints while preserving calibration.

Reject Option Classification (Kamiran, Karim and Zhang 2012) changes output labels close to the decision boundary since, according to the authors of the paper, “discriminatory decisions are often made close to the decision boundary because of decision maker’s bias”. The input for this algorithm is a representation (a dictionary) for the unprivileged and privileged group.

5 Fairness Toolkits

Several open-source toolkits have been developed that contain algorithms and pre-defined metrics that are supposed to be readily used to measure or improve the fairness of AI models. In this section, we describe some of the strengths and weaknesses of these toolkits based on two studies. First, a systematic review study by (Richardson and Gilbert 2021) that provides an overview of several prominent fairness tools that exist. Second, an exploratory assessment study by (Lee en Singh 2021) that conducted usability tests, interviews and surveys to derive the strengths and limitations of some of the available fairness tools, as well as general gaps of current toolkits with respect to requirements from the data science field.

(Lee en Singh 2021) identified six toolkits that are prominent in the sense that they contain relevant implementations of fairness-related methodologies, are relatively easily findable and open-source. These are scikit-fairness, IBM Fairness 360 (AIF360), Aequitas Tool, Google What-if tool, and Fairlearn. Examples of other toolkits as described in Richardson and Gilbert (2021) include Fairness Indicators, LinkedIn's Fairness Toolkit (LiFT), PyMetrics Audit-AI and ML Fairness Gym. These toolkits allow for:

- Evaluating a model against pre-defined fairness metrics.
- Observing how different decision thresholds affect outcomes on a fairness metric.
- Explainability, as they can be integrated with explainability toolkits.
- Visualizations, such as visualizations of performance on fairness and accuracy metrics.
- Applying bias-mitigation algorithms.

The usability study in (Lee en Singh 2021) indicates that the Fairlearn and Google What-if tools are the most usable for users with pre-existing technical and fairness expertise as they contain the most comprehensive possibilities for customized analyses. The Fairlearn toolkit contains a relatively extensive set of fairness metrics and bias mitigation algorithms. The Google What-if is praised for its customizable visualization functionalities, including counter-factual explanation visualizations that allow the user to observe how a change in the input space relates to a change in the output space of the model. Further, as opposed to the other toolkits, both tools provide the possibility to test regression models in addition to classification models, such as an estimated probability of fraud in a fraud-detection model. For beginner data scientists, however, (Lee en Singh 2021) find that scikit-fairness may be better suitable as it easily integrates with popular scikit-learn machine learning modules. In case the user has a limited technical background, the Aequitas Tool may be helpful for it contains clear guidance on how to use the program.

Richardson and Gilbert (2021) and Lee and Singh (2021) also describe several limitations of the toolkits and pitfalls when using them. First, users of the fairness toolkits often indicate that they are overwhelmed by the level of expertise that is required to use them. There exists a steep learning curve of using the toolkits and given the complexity and variety of fairness metrics and bias mitigation algorithms that exist, users need to choose for themselves but feel unqualified to do so. What makes this even more confusing is that the different toolkits provide different instructions to the user for choosing an appropriate fairness metric. Importantly, the toolkits seem to be developed for relatively straightforward use cases and are hard to use for more realistic and complex use cases. In such complex scenario's, a major concern is that the emphasis of the tools on technical solutions may lead to a "formalism trap", where complex socio-technical aspects during model development that can induce unfairness are ignored. This can induce a false feeling of confidence that that the model is fair. Best-practices during the entire life-cycle of model development should be integrated next to any possible use of fairness-related techniques as provided in the fairness toolkits.

Some additional information on some of the toolkits is provided below. For more in-depth information on these toolkits, we refer the reader to the documentation (provided in the links) as well as the systematic review in (Richardson and Gilbert 2021).

Fairlearn is a community-driven Python library that contains an interactive visualization dashboard and bias mitigation algorithms. The visualization dashboard can be used to measure and visualize which groups might be negatively impacted by the model. It helps in navigating the trade-off between performance (e.g. overall predictive accuracy) and fairness of the model as it allows to compare multiple models in terms of performance and fairness. The bias mitigation module can be installed and the algorithms can be used as wrappers around standard machine learning algorithms, such as those implemented in Scikit-learn. The mitigation algorithms consist of pre-processing algorithms that use "re-weighing" (refer to explanation in this document), as well as post-processing algorithms (refer to section) that adapt the model's output in order to optimize for a selected fairness metric. These algorithms work for both classification and regression models, where, in the case of a regression model, the output is converted to categories using one or multiple decision boundaries.

- https://fairlearn.org/v0.7.0/about/index.html

AI Fairness 360 is one of the most extensive open-source toolkits that exists. It was developed by IBM and contains several metrics to measure fairness and algorithms to improve fairness. It contains individual fairness metrics (e.g. group distortion) next to group-fairness metrics.

- https://aif360.mybluemix.net/

Fairness Indicators is developed and maintained by Google and enables computation of various of the group fairness metrics, both for binary classifiers and multiclass classifiers. The toolkit requires that the user provides one or two models, that can subsequently be studied and compared using the tool. The tool is embedded within the TensorFlow Python package, but models that are created using other packages, such as Scikit-learn, can be used as well. The tool allows to study model performance for different groups and compute confidence intervals to obtain an indication of the certainty of the evaluation. The tool allows for intersectional analyses. It is specifically designed to cope with large datasets, such as those generated by “billion user” systems used by Google itself. Hence, the fairness indicators can be computed on enormous amounts of data.

- https://www.tensorflow.org/tfx/guide/fairness_indicators

The **What-If Tool** is developed by Google as well. It contains visualization functions that help in deriving insights into the workings of the model, including potential biases. It specifically allows for providing counterfactual explanations of the workings of the model.

- https://pair-code.github.io/what-if-tool/

Aequitas was developed by (Saleiro, et al. 2018) and is a relatively simple toolkit with a single purpose, namely to audit the binary output of a model using one or more of the pre-defined fairness metrics. It provides multiple functions that can be used to compute and visualize the distribution of errors across different subgroups, which can highlight subgroups that may be disadvantaged by the model. When solely concerned with identifying group differences in the distribution of errors of a binary model, this toolkit may be a good choice as it is very simple and easy to use. In case a regression model is used, the user has to provide a decision boundary to convert the continuous output into two categories.

http://aequitas.dssg.io/

6 Fairness Guidelines and Checklists

As discussed in the previous section, the use of technical solutions, such as bias mitigation algorithms and optimizing on fairness metrics, is not straightforward and the complex socio-technical aspects during model development should be incorporated as well. To include best socio-technical practices during the entire life-cycle of AI model development, researchers, practitioners and institutions have developed handbooks, guidelines and checklists that translate high-level ethical principles into more actionable protocols, heuristics and questions. In this section, we describe some of our overarching observations concerning the (dis)similarities, usability and limitations of these documents. It is not intended as a comprehensive documentation of the different handbooks.

A portion of the documents that we base our observations on was found in the review studies by (Richardson and Gilbert 2021) as well as (Madaio, et al. 2020). Given that some of the handbooks have only relatively recently been published, however, we consulted our own network as well. The documents include: Ethics Guidelines for Trustworthy AI, capAI, Non-discrimination by Design Handbook, Fairness Handbook, Data Ethics Framework, Ethics & Algorithms Toolkit, Ethics in Tech Practice A Toolkit, and Deon. Some of these documents are briefly described at the end of this section.

6.1 Problem Space

One of characteristics that sometimes differentiates between the different handbooks lies in the problem(s) towards which they are directed. For example, whereas the Non-discrimination by Design Handbook is primarily devoted to incorporating best fairness practices into the data science workflow, the Data Ethics Framework is targeted towards three principles, namely transparency, accountability and fairness. Further, perpendicular to the problem area that the handbook is focused on is the perspective from which it is considered a problem. That is, is fairness, and possibly other valuable principles, considered to be desirable primarily from a juridical or ethical perspective? The Non-discrimination by design handbook, for example, has a highly legal perspective on the problem of discrimination and, consequently, attempts to describe best-practices such as to conform to the law. The Fairness Handbook, on the other hand, views the problem primarily from an ethical perspective. Interestingly, however, the best practice "solutions" to the problem are largely the same between the different handbooks, as we discuss in the next subsection.

6.2 Solution Space

Most of the handbooks and guidelines describe best-practice principles for fair AI using a process-based approach. That is, the process of AI model development is broken down by the various stages of AI model development (i.e. the AI “life-cycle”). For each of these stages, the handbooks provide heuristics, questions and practices that they deem relevant and practical for enhancing the fairness of the AI model. These practices are generally socio-technical, in the sense that they may have a social, organizational or technical nature. The life-cycle stages that are defined in the different handbooks are not always identical, but generally resemble the development stages in data science projects defined in the well-known CRISP-DM model (Chapman, et al. 2000) (Martínez-Plumed, et al. 2021), namely: business understanding, data understanding, data preparation, modeling, evaluation and deployment.

In the *business understanding* phase, the “real-world” problem needs to be defined and possibly translated into an AI problem. Some of the principles that are often stressed in the fairness handbooks during this stage are preventing of falling into the “solutionist trap” (i.e. an overreliance on technology as a solution to real-world problems), reflecting on how the AI model will impact citizens and society and defining potential sensitive subgroups that may be harmed by the model. Examples of guidelines that are generally recommended during the *data understanding* and *data preparation* phases, are observing whether any sensitive attributes are present in the data as well as features that statistically relate to such attributes (i.e. as “proxies”). Also, observing whether sensitive subgroups are equally represented in the data and considering sampling techniques to mitigate any representation disparities between such groups is usually recommended. Principles that are typically stressed in the *modeling* phase are choosing a human-interpretable model over a more complex “black-box” model whenever the possible decrease in predictive performance allows it, measuring feature importance to indicate which variables are important for the model in making its predictions, and analyzing how the features with high importance relate to sensitive subgroups. In the *evaluation* phase, handbooks usually focus on evaluating fairness using fairness metrics, such as demographic parity or calibration (see section Fairness Metrics), and reconsider whether the model may cause any harm to individuals or sensitive sub-groups. Examples of best-practice recommendations during the *deployment* phase are to inform end-users about the capabilities and potential shortcomings of the model and incorporating a monitoring plan such as to test for “concept drift” (loss of modeling validity over time due to a changing environment) or a “reinforcing feedback loop”, where historical biases are first learned and then reinforced by deploying the model.

We did not find major differences between the handbooks in terms of the nature and type of the solutions they provide. Some handbooks incorporate slightly more information on technical solutions, such as some of the bias mitigation algorithms that exist and how to choose a fairness metric. Other focus primarily on organizational and process solutions, such as thinking about how the use of an algorithm might affect individuals. The solutions in the handbooks seem to be primarily targeted towards the project leader of the data science project or the data scientist that is actually developing the model. The solutions that are directed towards these actors generally do not have an imperative nature. They are mostly provided as options or possibly suggestions, not as “must do’s”. The reason is probably that the optimal low-level actions to achieve a high-level ethical objective such as fairness are highly use-case specific. This is mentioned in (Richardson and

Gilbert 2021) as one of the usability limitations of most Toolkits and Handbooks, as it remains difficult to determine what to do in any specific use-case context.

6.3 Discussion & Conclusion

The different handbooks, guidelines and checklists sometimes have slightly different focus areas, since some of the documents are aimed towards additional high-level principles next to AI fairness. The solutions that are provided tend to have largely the same nature as they generally cover social and organizational practices as well as technical solutions that can be implemented at the various stages of the developmental life-cycle of AI models. The major differences between the various handbooks, however, lies in the extensivity and readability/usability of the documents. The Fairness Handbook, for example, comprises almost 70 pages with detailed information on the concept of fairness, various types of bias and how to mitigate bias issues. The Digital Decisions Tool, on the other hand, only contains a few questions that may be asked per developmental stage. Further, the readability also differs largely between the documents in our view. For example, where the Fairness Handbook is clearly structured and provides easy-to-follow details and information on fairness-related concepts, capAI has a relatively unclear scope and chapter-structure.

6.4 Brief Description of the Handbooks

The **Non-discrimination by Design Handbook** (2021) was commissioned by the Dutch Ministry of the Interior and Kingdom Relations and collaboratively developed by researchers from Tilburg University, Eindhoven University of Technology, Free University of Brussels, as well as the Netherlands Institute for Human Rights. The document is targeted towards project leaders working in AI projects and contains guidelines on the questions and principles that should be leading in the development and implementation of non-discriminatory AI systems. The questions and principles are composed from three perspectives that are, according to the authors, relevant when designing fair AI systems, namely a *legal*, a *technical* and an *organizational* perspective. For every of the developmental phases, the document provides questions and principles that relate to the different perspectives, as well as brief illustrations of how these questions and principles may be addressed in hypothetical examples. The legal perspective with its associated questions and principles is relatively extensive and informative compared to other guidelines that exist. This can be helpful when juridical context is needed in a project.

- handbook non-discrimination by design(ENG).pdf

The **Fairness Handbook** (2022) was developed by the Municipality of Amsterdam and provides a handbook that is solely devoted to implementing fairness principles when developing AI models. The document provides details on fairness-related concepts, such as bias and different types of harms that may be done to individuals. In contrast to the other handbooks, the relevant biases that may occur in every developmental stage are clearly laid-out. Further, it provides instructions and explanations of different fairness metrics that may be used. Another strong point is the explanation of different “traps” that can be encountered during the translation of the real-world problem to an AI problem (i.e. Business understanding phase in CRISP-DM).

- The Fairness Handbook (amsterdamintelligence.com)

Ethics Guidelines for Trustworthy AI (2019) was developed by the European Union High-level Expert Group. It has a broader scope than the previous two handbooks as it focusses on “thrustworthy AI”, which the authors define as incorporating three principles, namely that trustworthy AI is *legally compliant*, *ethical* and *technically robust*. The document first explains four high-level ethical principles that organizations developing AI models should adhere to, namely *respect for human autonomy, prevention of harm, fairness* and *explicability.* It then sets out seven key requirements for AI models. There are *human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental wellbeing; accountability.* It explains technical and non-technical methods to implement these requirements. These requirements should be addressed continuously during the AI system’s life cycle. Finally, the document operationalizes these requirements by providing an assessment list with a proposed set of questions that should help in achieving the trustworthy AI requirements. This set of questions, however, should be tailored to the specific use-case.

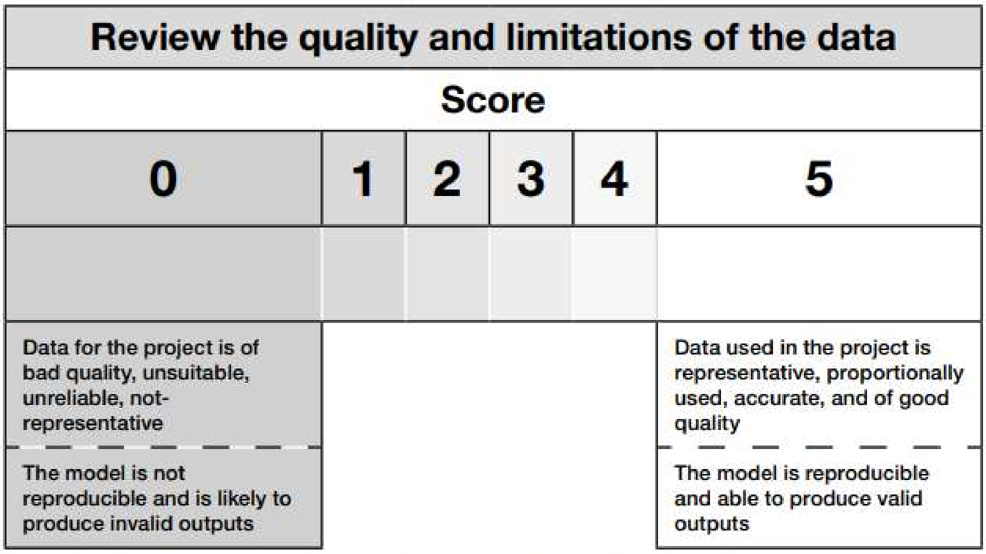
- Ethics guidelines for trustworthy AI | Shaping Europe’s digital future (europa.eu)

capAI (2022) was developed by researchers of Oxford University and is devoted to providing organizations within the European Union (EU) with practical guidance on how to ensure that their AI models comply with EU rules and regulations. It is specifically concerned with providing guidelines on how to conform to the proposal for the European law named *Artificial Intelligence ACT* (AIA)[4]. Like Ethics Guidelines for Thrustworthy AI, the scope of capAI is broader than just fairness, as it is concerned with AI being legally compliant, technically robust and ethical (the principles of AIA). Interestingly, although it is specifically designed the conform to AIA, its recommendations and best-practices on the topic of fairness do not significantly differ from the previous described documents on fair or ethical AI.

- https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064091

The Data Ethics Framework was commissioned by the government of the United Kingdom and designed to promote appropriate and responsible data use in public bodies. It is aimed at anyone working directly or indirectly with data in the public sector. Further, the document is not only concerned with actualizing fairness principles, but also transparency and accountability principles in services that rely on data. Hence, the document has a relatively broad scope. The framework is split into three overarching principles, namely *transparency*, *accountability* and *fairness*, as well as five specific actions that should be carried out in projects that rely on data. The actions that are defined are listed as follows:

1. Define public benefit and user need.
2. Involve diverse expertise.
3. Comply with the law.
4. Check the quality and limitations of the data and the model.
5. Evaluate and consider wider policy implications.



Review the quality and limitations of the data

Score

0	1	2	3	4	5
Data for the project is of bad quality, unsuitable, unreliable, not-representative					Data used in the project is representative, proportionally used, accurate, and of good quality
The model is not reproducible and is likely to produce invalid outputs					The model is reproducible and able to produce valid outputs

Figure 6. Self-assessment item for the "Check the quality and limitations of the data and the model" action in the Data Ethics Framework. Figure adapted from (Data Ethics Framework 2020)

The actions are ordered such that they tend to become more relevant as the project proceeds. For every action, a variety of questions, as well as general best-practice heuristics, are formulated that are categorized by the ethical principle that they relate to. An example question that is concerned with fairness is "*How has the data being used to train a model been assessed for potential bias?*", where they suggest you should consider both historical biases and selection biases. Using a self-assessment methodology, the user of the framework can give his or her project a score from one to five, which can consequently be used to determine ethical weak-spots in the project. An example of such a self-assessment item is provided in Figure 6.

- Data Ethics Framework (publishing.service.gov.uk)

The **Digital Decisions Tool** was designed by the Center for Democracy & Technology[5]. It is a process based interactive tool that attempts to translate principles for fair decision-making into a series of questions that can be addressed during the various stages of development of the AI tool or service. When consulting the tool for possible relevant questions during the evaluation stage of model development, for example, one of the questions posed by the tool is: "*Are errors evenly distributed across all demographics?*". The tool does not suggest any possible interventions or actions that may be considered to measure or improve the fairness of the model. Figure 7 provides an illustration of the tool.

- Digital Decisions Tool - Center for Democracy and Technology (cdt.org)

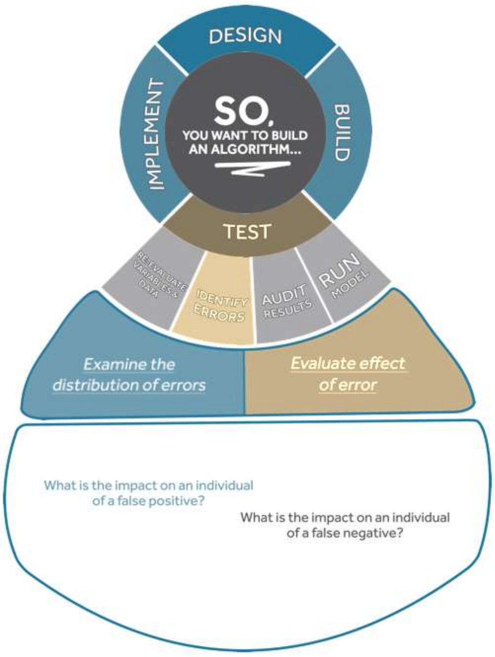


Figure 7. Illustration of the Digital Decisions Tool. Figure adapted from (Digital Decisions Tool 2017).

7 References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A Reductions Approach to Fair Classification. *arXiv:1803.02453*.

Agarwal, A., Dudík, M., & Wu, Z. S. (2019). Fair Regression: Quantitative Definitions and Reduction-based Algorithms. *arXiv:1905.12843*.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine Bias*. (ProPublica) Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Binns, R. (2020). On the Apparent Conflict Between Individual and Group Fairness. *Conference on Fairness, Accountability, and Transparency (FAT)* (pp. 514-524). New York: Association for Computing Machinery.

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, (pp. 3995-4004). Retrieved from http://papers.nips.cc/paper/6988-optimized-pre-processing-for-discrimination-prevention.pdf

Celis, L. E., Huang, L., Keswani, V., & Vishnoi, N. K. (2018). Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. *arXiv:1806.06055*.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide.

Cortez, V. (2019, September 24). *How to define fairness to detect and prevent discriminatory outcomes in Machine Learning*. Retrieved from https://towardsdatascience.com/how-to-define-fairness-to-detect-and-prevent-discriminatory-outcomes-in-machine-learning-ef23fd408ef2

Data Ethics Framework. (2020). Retrieved May 4, 2022, from UK Government: https://www.gov.uk/government/publications/data-ethics-framework

Digital Decisions Tool. (2017). (Center for Democracy & Technology) Retrieved May 4, 2022, from https://www.cdt.info/ddtool/

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Theoretical Computer Science Conference.* New York: Association for Computing Machinery.

(2019). *Ethics Guidelines for Trustworthy AI.* European Commission, High-Level Expert Group, Brussels.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and Removing Disparate Impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 259–268).

Retrieved from https://doi.org/10.1145/2783258.2783311

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im)possibility of fairness. *arXiv:1609.07236*.

Goodfellow, I., Ouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengo, Y. (2014). Generative adversarial models. *Communications of the ACM, 63*(11), 139-144.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems.*

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems, 33*, 1-33. doi:https://doi.org/10.1007/s10115-011-0463-8

Kamiran, F., Karim, A., & Zhang, X. (2012). Decision Theory for Discrimination-Aware Classification. *IEEE 12th International Conference on Data Mining.*

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer. *Machine Learning and Knowledge Discovery in Databases*, (pp. 35-50).

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. *arXiv:1711.05144*.

Lee, M. S., & Singh, J. (2021). The Landscape and Gaps in Open Source Fairness Toolkits. *Conference on Human Factors in Computing Systems*, (pp. 1-13). doi:10.1145/3411764.3445261

Madaio, M. A., Stark, L., Vaughan, J. W., & Wallach, H. (2020). Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. *CHI Conference on Human Factors in Computing Systems*, (pp. 1-14). doi:10.1145/3313831.3376445

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., . . . Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering, 33*(8), 3048-3061.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys, 54*(6), 1-35. Retrieved from https://doi.org/10.1145/3457607

Merriam-Webster. (n.d.). *Gerrymandering*. Retrieved from https://www.merriam-webster.com/dictionary/gerrymandering

Muhammad, S. (2022). *The Fairness Handbook.* Amsterdam: Gemeente Amsterdam.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On Fairness and Calibration. *Advances in Neural Information Processing Systems.*

Richardson, B., & Gilbert, J. E. (2021). A Framework for Fairness: A Systematic Review of Existing. doi:arXiv:2112.05700

Ruf, B., & Detyniecki, M. (2021). Towards the Right Kind of Fairness in AI. *arXiv:2102.08453*. Retrieved from https://arxiv.org/abs/2102.08453

S. Barocas, S. M.-T. (2021, December 7). Retrieved from The FAT-ML Workshop Series on Fairness, Accountability, and Transparency in Machine Learning: https://www.fatml.org/resources/principles-for-accountable-algorithms

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., . . . Ghani, R. (2018). Aequitas: A Bias and Fairness Audit Toolkit. doi:10.48550

Sandvig, C. (2014). Seeing the Sort: The Aesthetic and Industrial Defense of "The Algorithm". *Journal of the New Media Caucus, 10*.

Tversky, A., & Kahneman, D. (1981). The Framing of decisions and the psychology of choice. *Science, 211*(4481), 453-458.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning Fair Representations. *Proceedings of the 30 th International Conference on Machine Learning*, (pp. 325-333).

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, (pp. 335–340).

Artificial Intelligence

Anna van Buerenplein 1
2595 DA Den Haag
www.tno.nl