CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Assessing and enforcing fairness in the AI lifecycle
Authors	Calegari, Roberta;Castañé, Gabriel G.;Milano, Michela;O'Sullivan, Barry
Publication date	2023
Original Citation	Calegari, R., G. Castañé, G., Milano, M. and O'Sullivan, B. (2023) 'Assessing and enforcing fairness in the AI lifecycle', in Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, Macau, SAR China: International Joint Conferences on Artificial Intelligence Organization, pp. 6554–6562. https://doi.org/10.24963/ijcai.2023/735.
Type of publication	Article (peer-reviewed)
Link to publisher's version	https://doi.org/10.24963/ijcai.2023/735 - 10.24963/ijcai.2023/735
Rights	Accepted version © 2023 the authors - https://creativecommons.org/licenses/by/4.0/
Download date	2026-10-09 11:03:36
Item downloaded from	https://hdl.handle.net/10468/15259

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Assessing and Enforcing Fairness in the AI Lifecycle

Roberta Calegari[1*], **Gabriel G. Castañé**[2], **Michela Milano**[1] and **Barry O'Sullivan**[2]
[1]ALMA MATER STUDIORUM–Università di Bologna, Italy
[2]Insight SFI Research Centre for Data Analytics, University College Cork, Ireland
{roberta.calegari, michela.milano}@unibo.it, {gabriel.castane, barry.osullivan}@insight-centre.org

Abstract

A significant challenge in detecting and mitigating bias is creating a mindset amongst AI developers to address unfairness. The current literature on fairness is broad, and the learning curve to distinguish where to use existing metrics and techniques for bias detection or mitigation is difficult. This survey systematises the state-of-the-art about distinct notions of fairness and relative techniques for bias mitigation according to the AI lifecycle. Gaps and challenges identified during the development of this work are also discussed.

1 Introduction

Artificial intelligence (AI) systems exploiting machine learning (ML) algorithms trained on data from different domains – banking, education, legal, or human resources – often support humans in their decision-making processes. Despite the many benefits decision support systems may offer in economic, speed and accuracy of solutions, we run the risks of discriminating and biasing societal groups of individuals if training data is biased [Leavy *et al.*, 2021].

The scientific community has been extensively working to address fairness issues through solutions, methods, and metrics to avoid bias by increasing awareness of the impact of AI failures on developers and industries. However, the literature is wide, many existing approaches can be applied only to specific types of bias and, thus, it is not so obvious how to find the right approach to be applied to a given setting. Furthermore, it is key to ensure that solutions to preventing bias are applied at the right time in the development of AI systems. Incorporating them into the AI lifecycle enables AI practitioners to bring them to their daily practices. Therefore, despite the many attempts to compile information on a survey for fairness in ML [Pessach and Shmueli, 2022; Mehrabi *et al.*, 2021; Caton and Haas, 2020], it is important to couple the concept of fairness with the technique that can be exploited to achieve it, positioning them in the AI lifecycle.

The criteria to systematise the state-of-the-art chosen in this work use, as inputs, existing work on fairness in ML, and enlarge the discussion focusing on the two activities that must be run in the AI lifecycle in terms of fairness: *i) fairness awareness*, how to measure and assess fairness (or bias, Section 2), and *ii) fairness reparation/mitigation*, how to mitigate bias in models when necessary and in which step of the AI lifecycle (Section 3). Using the work in this paper, researchers and developers can find the appropriate technique to use in their application scenario, depending on the most efficient moment of the AI lifecycle. Current gaps and future opportunities in the field are as well identified.

Section 2 discusses the notion of fairness along with related metrics. Fairness enforcing techniques, step two of the process, are then reviewed and divided into categories that consider fairness metrics and the phase in the AI lifecycle (Section 3). Finally, the work ends by discussing challenges and potential barriers (Section 4).

2 Assessing Fairness

Two elements are required for fairness awareness: *i)* definition of the fairness notions and *ii)* a quantitative mechanism to measure them. Fairness notions are context-dependent and encompass how society perceives what is fair in the case at hand. They can be evaluated through a statistical formula – fairness metric – providing a quantitative way to measure fairness. As a result, developing quantitative formulations of fairness metrics becomes challenging due to the process of capturing all the nuances of fairness that can arise on technical, societal and legal aspects [Chierichetti *et al.*, 2019]. Of course, fairness metrics in AI applications must be balanced with other and often conflicting AI metrics—related to other trustworthy requirements, such as accuracy and performance.

Dozens of competing fairness metrics are proposed in the literature. Frequently, these are bespoke to some scenarios, each with specific advantages and drawbacks. Many different propositions were made to categorise these metrics, but none of these is complete [Howard *et al.*, 2016]. This work encompasses relevant fairness metrics categorisation, discussing aspects that should be considered when selecting a particular metric. Figure 1 sets out a general framework for systematising the different fairness notions - labelled from 1) to 6).

2.1 Procedural Fairness

Procedural fairness is a concept inherited from administrative law concerned with equality of treatment within the process that carries out a decision, i.e., fair treatment of people.

*Contact Author

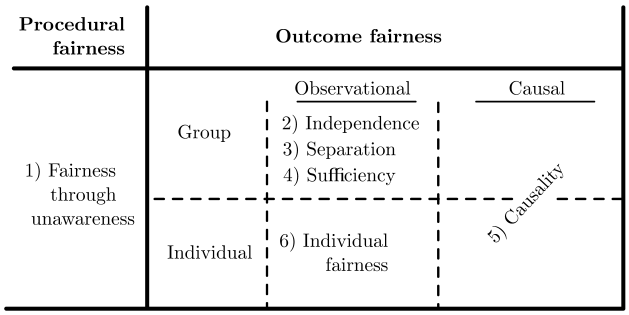

Figure 1: Organising framework of algorithmic fairness metrics

In the computational area, specifically for AI algorithms, the concept relates to the information that must be considered in decision-making. This has often led to being interpreted in the literature as not including sensitive attributes in the AI algorithm. Omission of the sensitive attributes or *fairness through unawareness* – 1) in the Figure 1 – is here the main approach. However, the model accuracy is reduced and the *discrimination effects* do not improve as a consequence of neglecting relationships with proxy variables, ignoring that prejudice may not be caused by a single variable but rather by a combination of several ones. Omissions potentially increase bias or concealment of discrimination [Bacelar, 2021].
Metrics: fairness through unawareness [Dwork *et al.*, 2012].

2.2 Outcome Fairness

Outcome fairness is the term used to define equality ('fair result') of the outcomes of the decision making processes. The literature is wider in this category and can be classified into two orthogonal groups of two dimensions each: individual vs. group notions of fairness, and observational vs. causal approaches. In the first, the *individual* notions of fairness compare single outcomes for individuals to establish their fairness while *group* notions of fairness work on outcomes aggregated by several individuals belonging to the same sensitive category. These two dimensions are not mutually exclusive. In the second case, notions of fairness can be classified as *observational*, described as joint distributions of observable aspects such as outcomes, decisions, features, and sensitive attributes; or *casual* in case the causal inference is required to acquire knowledge about variables and their (co)relations.

Observational Fairness. Some of the advantages of the observational definitions are the easiness of the state and a lightweight formalism. Moreover, assumptions are excluded from the inner workings of the classifier, the impact of the decisions, and possible correlations between features and outcomes. However, a major drawback is that they share limitations in the scope of the evaluation of the available data. They do not evaluate what is not observable [Kilbertus *et al.*, 2017].

Four categories compose the set of observational fairness, three for the *group* notion and one for the *individual* category. The *group* notion of fairness considers as the core of fairness definitions and metrics some of the fundamental aspects of a classifier. These are *a)* the sensitive variable S as number of groups to be measured, *b)* the target variable Y as the prediction classes, and *c)* R as the classification score. Based on these, three subcategories arise based on the "non-discrimination" statistical criteria: independence, separation and sufficiency – labelled in Figure 1 as 2), 3) and 4) respectively. Note that, there is a relation of mutual exclusion between the three subcategories that make them pairwise incompatible. Observational fairness can be ensured also at the individual level by applying the same criteria to single individuals instead of groups (6) in Figure).
Independence metrics: statistical parity, group fairness, demographic parity, conditional statistical parity.
Separation metrics: equal opportunity, equalised odds, balance for the negative class, balance for the positive class, predictive equality, equalised correlations.
Sufficiency metrics: groups calibration, predictive parity.
Individual fairness metrics: individual fairness.

Causal Fairness. *Causality*-based criteria, 5) in Figure 1, employs additional knowledge (e.g. external experts) to discover the causal structure of the case at hand. The structure can be analysed by comparing different outcomes of counterfactual scenarios and by modifying sensitive attributes that belong to this. For example, the what-if scenario: "what would have been the decision if that individual/group had a different race?" Then, counterfactual scenarios can be compared and evaluated, highlighting fairness discrepancies. Causality-based notions are richer than observational notions and permit the selection of which causal paths – from sensitive attribute to an outcome – can be legitimate or should be forbidden [Mhasawade and Chunara, 2021]. Causal notions are promising, but the fairness notion is governed by the construction of causal graphs, often difficult to be built due to lack of knowledge or computationally expensive.
Metrics: unresolved discrimination, counterfactual fairness.

2.3 Running Example on Recruiting Tool

To illustrate the difference between metrics, let us consider an example of an AI recruiting tool in which the system could potentially discriminate across sensitive attributes, like for instance gender. A naive approach of *fairness through unawareness* would consist of removing the dataset's sensitive attributes. The method is proven to be inefficient in some cases [Dwork *et al.*, 2012]. This means in our example that although gender features are omitted from the dataset, other unknown correlated features might remain (e.g., marital history) being "proxies" for revealing the gender, thus, the model remains biased.

Enforcing *independence* would ensure the equality of outcomes (selection) – the classification scores – as these should be independent of the sensitive attribute. For instance, via statistical parity, the outcome is the same for different sensitive groups (equal acceptance rate of males and females applying for a position); via demographic parity, each sensitive group – males and females – should receive the positive outcome – being hired – at equal rates; via conditional statistical parity, the outcome is the same for different sensitive groups, adding additional factors for consideration (e.g. divorced men and divorced women have the same rate of acceptance, and single men and single women have the same rate of acceptance).

A "stronger" criterion that could be utilised to enforce fairness is to enforce rules like "similar people should be treated

similarly", i.e., given a set of candidates that are qualified enough for the work, no bias should materialise when a subset of these candidates is chosen for the work. This could be achieved by enforcing *separation*, ensuring equality of errors. In terms of metrics, and related to the example, equal opportunity means having a similar rejection rate for males and females, for example, despite being qualified enough for selection. This means each group's likelihood of false positive and false negative predictions should be equal. With equalised odds, the chances of a false negative and a false positive should be the same for each sensitive group. In the example, the chance that you will be denied a job even though you are qualified should be the same whether you are a man or a woman, and the chance that you will be given a job even though you are unqualified should be the same whether you are a man or a woman. Ensuring balance for the negative class means the probability of getting a correct negative outcome is the same for each sensitive group, i.e., the probability that you will be denied a job when you are not qualified is the same whether you are a man or a woman (dual for the positive class). With predictive equality, the chances of predicting a false positive should be the same in each sensitive group, i.e., the chance that you will be given a job even though you are not qualified should be the same whether you are a man or a woman. One of the reasons that separation might be more desirable than independence is because there might be some correlation between the sensitive features and outcome.

The *sufficiency* notion (calibration) ensures that choices reflect the same accuracy per subgroup. In the example, it means the chances of males and females being qualified enough given the hiring decision should be the same. Exploiting sufficiency means looking for calibration, i.e., if we consider a set of people who receive a predicted probability of p, we would desire that the fraction of the members of this set that are positive instances of the classification problem is equal to p. Moreover, if we are concerned about fairness between two groups (e.g. male and female) then we would like this calibration condition to hold simultaneously for the set of people within each of these groups as well.

Finally, *causality* – exploiting the causal graph and the observed data – means to answer hypothetical questions of the form "What would the hiring decision have been in case I am a different gender?". Ensuring counterfactual fairness means that for any individual, the outcome does not change in the counterfactual scenario where the sensitive attributes change.

3 Enforcing Fairness

Figure 2 depicts the phases of the AI lifecycle. The figure shows three inner arrows encompassing the primary phases for the time of intervention in terms of fairness. This time is also commonly named fairness intervention time and defines the adequate phase in which fairness must be tackled in order to maximise the adequate result or to enable its applicability. The three phases are: pre-process, which comprises the data processing (from collection requirements to preparation); the in-process, encompassing the ML modelling, development and evaluation; finally, the post-processing, which is when the model is deployed, tuned and monitored.

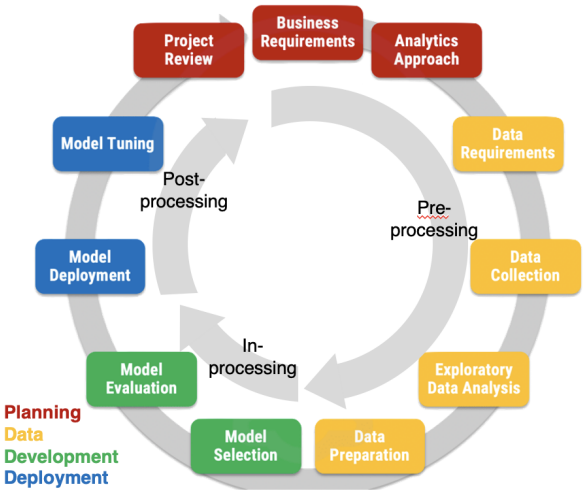

Figure 2: AI lifecycle & fairness intervention time

Pre-processing techniques approach the problem by removing the underlying discrimination from the data prior to modelling. This is argued in the literature to be the most flexible phase of repairing bias in the pipeline, as it makes no assumptions with respect to the choice of applied modelling technique. The methods, that modify the training data are at odds with policies like GDPR's right to an explanation, potentially introducing new biases. Sufficient knowledge of the data and veracity assumptions are required. *In-processing* techniques modify the traditional learning algorithms to account for fairness during the model training phase. They require a higher technological effort and integration with standard ML libraries to avoid porting challenges. *Post-processing* final classes of methods can be performed as post-training processing of the output scores of the classifier to make decisions fairer. The accuracy is suboptimal when compared to "equally fair" classifiers and could be the case that test-time access to protected attributes is needed, which may not be legally permissible.

Table 1 categorises existing fairness intervention approaches by their phase in the AI lifecycle and technique (rows), and which notion of fairness is sought (columns).

3.1 Pre-processing

Blinding. Blinding is a pre-processing technique that, according to Robertson [2016] "immunises" a predictor with respect to one or more sensitive variables. Authors define 'blinding predictors' as the process to protect attributes which are not direct inputs or features in the computation, thus there is no observable differentiation of results based on the attributes. For example, a predictor is gender blind if there is no observable differentiation of results based on gender.

Furthermore, a different process different to procedural fairness blinding via "immunity" is to blind the predictor via "omission". In this case, it allows fairness through unawareness, then explicitly prohibiting the input of any protected characteristics into the decision models. Unfortunately, omission has been shown to reduce model accuracy and not improve discrimination effects [Chen *et al.*, 2019].

Moreover, blinding in the scope of outcome fairness can lead to achieving independence. Feldman et al. [2015] present a method which removes dependencies between two variables but still allows these variables to be considered in the prediction process. However, the approach presents some implementation barriers and limitations such as the applicability to non-numeric data (i.e., categorical).

Both omission and immunity neglect relationships with proxy variables leading to potential increases in discrimination [Bacelar, 2021]. Blinding (or partial blinding) has also been used as a mechanism to conduct fairness audits, similar to causal model approaches.

Adversarial Learning. This technique is used in ML applications before a training process. There are adversaries that try to establish the fairness process. In case the model is not considered fair, the adversary's input is used for the model refinement. Adversaries are used as a pre-processing transformation process on the training data (e.g. [Feng *et al.*, 2019; Adel *et al.*, 2019]) often moving towards a notion of censoring the training data with similar objectives to blinding.

Independence is the notion of fairness exploited by existing adversarial learning approaches and specifically focused on disparate impact and statistical parity. Some approaches allow the classifier to be additionally evaluated on their individual fairness dimension [Feng *et al.*, 2019].

The main advantage of adversarial learning approaches in fairness is that they can take into account several constraints at the same time, often by treating the paradigm as a blackbox. However, it has been documented that the applicability of these methods often lacks stability, making them difficult to be trained consistently [Bacelar, 2021]. In particular, in transfer learning, it happens that the protected variable is established only for a limited number of samples.

Causal Approaches. The discovery of causal relationships between variables and data is the focus of causal methods recently applied to fairness [Kilbertus *et al.*, 2017; Mhasawade and Chunara, 2021; Gupta *et al.*, 2018; Chiappa, 2019; Kusner *et al.*, 2017], as it is shown in Table 1. Specifically, to detect proxies of sensitive variables causal approaches are well suited for purpose. However, computational resources are high as the technique used to model conditional assumptions between variables is by using directed acyclic graphs.

Fairness metrics involved in causal approaches for both *group* and *individual* spanning from unresolved discrimination and proxy discrimination to counterfactual fairness respectively. The metric compares a decision from two complementary perspectives: its fairness towards the individual (actual world) and its fairness in positioning the individual within a different demographic group (counterfactual world).

The information required to build the context knowledge of a causal model must be precise to adequately examine the scope, hence making its construction challenging or not always accessible. Therefore, they have been criticised for not well examining their applicability in practice [Mhasawade and Chunara, 2021].

Relabelling and Perturbation. Altering the distribution of the variables in the training set is an approach that was largely investigated and can be categorised into two techniques: relabelling and perturbation. Relabelling, some authors referred to it as data-massaging, is the process of modifying or flipping the labels of training data instances to ensure that the proportion of positive instances is equal for all protected groups [Calders and Verwer, 2010; Kamiran *et al.*, 2010; Luong *et al.*, 2011; Kamiran and Calders, 2012]. Perturbation changes directly the value of the dependent variable [Wang *et al.*, 2019]. The main drawback of the approach is in the legal scope. The modification of the data via relabelling and perturbation is not always legitimate, hence changes to the data should be minimised. Furthermore, in some cases classifiers are negatively affected by altering the training data in an attempt to mitigate them. For these reasons, continuous (re)assessing fairness metrics and decisions is of paramount importance but also costly.

Table 1 shows that frequently relabelling and perturbation-based approaches are used as pre-processing techniques to prepare for an in-processing approach and reach fairer algorithms. However, some approach exist for the post-processing phase where the probability of having a positive decision as an outcome when this is altered by the modification of the probabilities in the model [Calders and Verwer, 2010; Kamiran *et al.*, 2010]. Finally, the most exploited fairness notion, as shown in the Table is Independence – demographic parity metrics with some exceptions of conditional statistical parity metrics – and according to Luong [2011]. However, recent work shows an algorithm that targets samples with an individual bias for remediation in order to improve both individual and group fairness metrics [Lohia *et al.*, 2019].

(Re)sampling. The main objective of sampling methods is to create a set of representatives that trains robust algorithms on detecting groups of the data that can be (are) underprivileged. Decoupled classifiers and multitask learning emerged as promising techniques has it is shown in Table 1 [Awasthi *et al.*, 2021; Dwork *et al.*, 2018]. The main objective of the authors on these approaches is to find the most accurate models for given subgroups (decoupled classifiers) or consider the observation of different subgroups (multitask learning). In both cases, the training data is decoupled, splitting it into subgroups according to sensitive variables or learned as part of the pre-processing phase. However, there are challenges in the selection of groups related to the processes of ensuring balance, atomicity, and robustness. Some of the consequences of not performing an adequate selection of groups can lead to overfitting, issues with minimising fairness metric(s), and/or other theoretical violations.

Reweighting. Reweighting is a technique that assigns weights to instances of the training data while leaving the data itself unaltered—in contrast to relabelling, perturbation, and transformation. With appropriate sampling – in comparison to relabelling and blinding approaches – reweighting can achieve high(er) accuracy. However, an issue is related to the classifier's stability and robustness. Moreover, the process becomes more opaque and therefore less explainable. Usually reweighing is applied as a pre- and in-processing approach. For example, Kamiran [2012] assigns weights based on the probability of an instance belonging to a particular

class and sensitive value pairing (pre-processing). Whereas, Krasanakis [2018] is an approach that first exploits an unweighted classifier for learning sample weights and then retrains their classifier using these weights (mixing pre- and in-processing).

The aims of reweighting are *i)* to set a lower/higher weight (importance) to some sensitive training samples [Kamiran and Calders, 2012] *ii)* to specify the frequency count of a kind of instance [Calders and Verwer, 2010], and *iii)* to increase the stability of the classifier [Krasanakis *et al.*, 2018].

3.2 In-process

Adversarial Learning. Most approaches in this area exploit notions of fairness within the adversary to apply feedback for model tuning as a form of in-processing where the adversary penalises the model if a sensitive variable is predictable from the dependent variable [Edwards and Storkey, 2015; Beutel *et al.*, 2017; Feng *et al.*, 2019].

Constraint Optimisation & Regularisation. In-processing (constraint) optimisation approaches have similar goals to fairness regularisation methods; hence, we present them together. The notions of fairness are included in constraint optimisation using several mechanisms such as the notions of fairness in the classifier loss function operating on the confusion matrix during the training of the model, with the incorporation of additional constraints – precision or budget – to improve the accuracy fairness trade-off [Goh *et al.*, 2016] or with the reduction of the problem to a cost-sensitive classification one [Goh *et al.*, 2016]. The challenge is in balancing conflicting constraints that potentially lead to unstable training.

Table 1 classifies several works that are into in-process techniques and that deal with group fairness in particular with *independence*, [Zemel *et al.*, 2013; Agarwal *et al.*, 2018; Louizos *et al.*, 2015; Goh *et al.*, 2016; Zafar *et al.*, 2017a; Kamishima *et al.*, 2012; Liu and Vicente, 2021], *separation*, [Corbett-Davies *et al.*, 2017; Zafar *et al.*, 2017b; Woodworth *et al.*, 2017; Quadrianto and Sharmanska, 2017; Bechavod and Ligett, 2017; Pessach and Shmueli, 2021],and *sufficiency* notion. An interesting approach is the one by Ignatiev [2020] that reconsiders the criterion of fairness through unawareness but proposes a semantic definition. Ignatiev [2020] discuss a formal method for certifying fairness through unawareness, developed criteria for assessing fairness in ML models and bias in datasets and relating fairness with explanations and robustness.

In the case of regularisation methods applied to fairness, the technique applied is to add one or more penalty terms into the model that evaluates the classifier behaviour seeking for biased outcomes [Heintz *et al.*, 2021; Kamishima *et al.*, 2012; Liu and Vicente, 2021; Bechavod and Ligett, 2017; Pessach and Shmueli, 2021]. However, key challenges to its applicability are related to the fact that regularisation methods are either non-convex in nature or convexity is reached at a high computational cost, therefore not all fairness measures are equally affected by regularisation parameters. To balance the fairness and accuracy some works augment the (convex) loss function of the classifier including fairness constraints (e.g. [Kamishima *et al.*, 2012; Liu and Vicente, 2021]). Finally, the outcomes of applying regularisation on different terms and penalties are data-sets dependant, i.e. the choice affects the trade-off between both accuracy and fairness [McCarthy and Narayanan, 2023].

3.3 Post-process

Calibration. Calibration is defined as: "to ensure that positive predictions proportion equals positive examples proportion" [Pleiss *et al.*, 2017]. In the context of fairness, this should also hold for all subgroups existing in the data. This applicability was explored on multiple protected groups and/or using multiple fairness criteria at once and has been shown by authors to be impossible [Pleiss *et al.*, 2017]. However, there are approaches to handle the impasse of achieving calibration and other fairness measures. For example, in the post-processing phase and to achieve a balance between accuracy and fairness, the individuals were randomised. But this solution affects outcome, as randomised individuals can be negatively impacted and the overall accuracy of the model adversely affected as shown by Pleiss [2017].

Relabelling. Some approaches exist for the post-processing phase of relabelling. These show the probability of having a positive decision as an outcome when this is altered by the modification of the probabilities in the model[Calders and Verwer, 2010; Kamiran *et al.*, 2010]. Furthermore, recent work shows an algorithm that targets samples with an individual bias for remediation in order to improve both individual and group fairness metrics [Lohia *et al.*, 2019].

Thresholding. This is a post-processing approach that determines threshold values via measures such as equalised odds for different protected groups. This ensures a balance between true and false positive rates can be found, minimising the expected classifier loss [Woodworth *et al.*, 2017]. The fairness notions considered are independence and separation.

An evaluation of protected group thresholds exploiting logistic regression was done by Menon [2018]. The authors used fairness boundaries to illustrate the misalignment between threshold values. Some alternative works propose other methods, like shifting decision boundaries using post-processing regularisation [Kamiran and Calders, 2012; Dwork *et al.*, 2012]. Some other approaches learnt a threshold value after training an ensemble of decoupled ensembles such that the difference between protected and non-protected groups is below some user-specified threshold [Woodworth *et al.*, 2017; Hardt *et al.*, 2016; Menon and Williamson, 2018]. It is worth mentioning that thresholding methods claim fair notions of equity when the threshold is correctly selected.

4 Discussion

4.1 Which Mechanisms and When?

Determining the right notion of fairness to be used must take into account the proper legal, ethical, and social context. As discussed in Section 2, different fairness notions exhibit different behaviour and must be executed in different phases. For this reason, the mechanisms present specific advantages and disadvantages.

Pre-processing mechanisms can be used with any classification algorithm, and this is an advantage. However, this may hamper the explainability of the results. In addition, they are not tailored to a specific classification algorithm and thus, the accuracy obtained at the end of the process is uncertain. This hampers the evaluation of the tradeoff fairness-accuracy.

Post-processing mechanisms can be used with any classification algorithm (as in pre-process). However, applying them in late-stage typically produces poorer results. It is more applicable to fully remove some kind of bias (as disparate impact), but often the desired measure is not achieved and it can deliberately damage accuracy for some individuals in order to compensate others (also related to legal issues and economic controversies in affirmative action). Post-process approaches should require humans at the end of the loop (decision makers) managing information about the group to which individuals belong (usually unavailable due to legal/privacy reasons).

One of the main advantages of the in-process mechanisms is the required trade-off between accuracy and fairness. It can be clearly defined in the objective function. However, mechanisms are tightly coupled with the algorithm itself.

Hence, the selection of the best phase in which to act has dependencies with the data, the availability of the sensitive attributes at testing time, and the fairness notion selected (since some can only be applied in certain phases). Different algorithms usually differ on the input requirements. Foremost among these is the encoding of sensitive attributes, the support for multiple sensitive attributes, and the support for categorical attributes or the transformations required by the algorithm. These context setups can vary between applications, and related choices directly affect the accuracy and fairness of a fairness-aware classifier. Finally, it should be considered that often algorithms are fragile: they are sensitive to variations in the input.

4.2 Why Fairness in the AI Lifecycle?

The need of merging fairness in the AI lifecycle is to incorporate fairness needs into the software operations, making it more sustainable from social and technical perspectives. The more *complexity* is added to AI operations, the less sustainable and, in particular, fair they become. We aim to help developers in a practical manner. To provide an understanding of the current fairness needs associated with each phase in the AI lifecycle when operating ML software continuously. Incorporating fairness seamlessly after the software is operational is in many cases unrealistic given this complexity. That is why the mapping is a first step that can be used as a robust pillar for stakeholders to tackle bias in a structured manner.

In addition, AI software is traditionally bespoke or domain-driven, focusing on addressing specific problems to solve. Thus, incorporating new components into the flow can be perceived as a significant constraint. But, the need to give a fair decision is directly related to the software's sustainability and, therefore, it becomes necessary. This work shows the first steps towards a methodological approach between software operations and a social contract underlying those operations. This will ensure continuity and fairness in the software's decision-making process.

4.3 Conclusions. Gaps and Challenges.

There are several challenges associated with the incorporation of fairness into the AI lifecycle. First, there is an *educational* aspect of AI practitioners. Software professionals must be able to create, maintain and continuously adapt their software as part of their lifecycle. Developers can perceive fairness as a burden for the processes, in spite of the heavy impact that bias can have on the business outcomes, mainly in reputation.

Furthermore, a second aspect is the lack of a *methodological approach* to tackle fairness in the different stages of the AI lifecycle. Although there are schemes well-defined for the production of AI systems, AI is frequently focused on addressing specific problems, and there is no methodology for incorporating fairness into the AI lifecycle. This is a major challenge for developers to understand the actions to be taken on each phase – data, model, development, operations – or, more important, the responsible to execute these, which, frequently, is assigned to different persons or even companies.

From a technical perspective, *diversification* is needed beyond existing algorithms and datasets. Literature focuses on supervised learning with an emphasis on binary classification. Furthermore, realistic and representative datasets must improve and new techniques are needed on the full set of features to avoid stability issues. In addition, *interpretable and transparent* approaches are required. The ability of humans to read the outcomes and understand the fairness process is key to building the trust of users in AI and, in some domains, it is enforced by law. Causality-based approaches provide a better understanding of the unfairness roots, improving explainability and the selection of measures and mechanisms.

Fairness metrics need to be balanced between individual and group notions of fairness by the model optimisations. Existing works are mainly focused on group fairness with respect to independence metrics due to: low development effort, low computational costs, and/or easing users' fairness understanding. More effort is required to strengthen existing metrics but also to research the development of new ones if necessary.

Experimentation environments are required to provide an easy playground to test different notions and techniques, comparing the results and having a better awareness of bias implications.

This work aims to put the first steps towards facilitating developers' seamless understanding of how to incorporate fairness into operations, however, incorporating additional processes can be challenging. To conclude, it is needed the development of fairer algorithms but most importantly the design of procedures to reduce biases in the data, for instance integrating humans and algorithms in the decision workflow. However, thus far, it seems that biased algorithms are easier to fix than biased humans or procedures.

		Procedural	Outcome					
			Group Fairness				*Individual Fairness*	
		Fairness through unawareness	Independence	Separation	Sufficiency	Causality	Causality	Individual fairness
Pre-process	Blinding	[Chen *et al.*, 2019]	[Feldman *et al.*, 2015]					
	Adversarial Learning		[Feng *et al.*, 2019] [Adel *et al.*, 2019]					[Feng *et al.*, 2019]
	Causal					[Kilbertus *et al.*, 2017] [Mhasawade and Chunara, 2021] [Gupta *et al.*, 2018]	[Kusner *et al.*, 2017] [Chiappa, 2019]	
	Relabelling		[Calders and Verwer, 2010] [Kamiran *et al.*, 2010] [Luong *et al.*, 2011] [Kamiran and Calders, 2012] [Wang *et al.*, 2019]					
	Resampling			[Awasthi *et al.*, 2021] [Dwork *et al.*, 2018]				
	Reweighing		[Kamiran and Calders, 2012] [Calders and Verwer, 2010] [Calders and Verwer, 2010]					
In-process	Adversarial Learning		[Edwards and Storkey, 2015] [Beutel *et al.*, 2017] [Madras *et al.*, 2018] [Feng *et al.*, 2019]					
	Constraint Optimization	[Ignatiev *et al.*, 2020]	[Zemel *et al.*, 2013] [Aïvodji *et al.*, 2021] [Louizos *et al.*, 2015] [Goh *et al.*, 2016] [Zafar *et al.*, 2017a] [Detassis *et al.*, 2020] [Agarwal *et al.*, 2018]	[Corbett-Davies *et al.*, 2017] [Zafar *et al.*, 2017b] [Woodworth *et al.*, 2017] [Quadrianto and Sharmanska, 2017] [Detassis *et al.*, 2020] [Aïvodji *et al.*, 2021]	[Corbett-Davies *et al.*, 2017]			[Dwork *et al.*, 2012]
	Regularization		[Kamishima *et al.*, 2012] [Liu and Vicente, 2021]	[Bechavod and Ligett, 2017] [Pessach and Shmueli, 2021]				
	Reweighing		[Kamiran and Calders, 2012] [Calders and Verwer, 2010] [Krasanakis *et al.*, 2018]					
Post-process	Calibration				[Pleiss *et al.*, 2017]			
	Relabelling		[Kamiran *et al.*, 2010] [Calders and Verwer, 2010] [Lohia *et al.*, 2019]					[Lohia *et al.*, 2019]
	Thresholding		[Kamiran and Calders, 2012] [Hardt *et al.*, 2016] [Dwork *et al.*, 2012]	[Woodworth *et al.*, 2017] [Woodworth *et al.*, 2017] [Menon and Williamson, 2018]				

Table 1: Fairness awareness via fairness notions (columns) and related intervention techniques in the AI lifecycle (rows).

Acknowledgements

The work has been supported by the AEQUITAS project funded by the European Union's Horizon Europe Programme (Grant Agreement No. 101070363). This publication has also emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 12/RC/2289-P2 which is co-funded under the European Regional Development Fund.

References

[Adel *et al.*, 2019] Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. One-network adversarial fairness. In *AAAI*, volume 33, pages 2412–2420, 2019.

[Agarwal *et al.*, 2018] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on ML*, pages 60–69. PMLR, 2018.

[Aïvodji *et al.*, 2021] Ulrich Aïvodji, Julien Ferry, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. Faircorels, an open-source library for learning fair rule lists. In *International Conference on Information & Knowledge Management*, pages 4665–4669, 2021.

[Awasthi *et al.*, 2021] Pranjal Awasthi, Alex Beutel, Matthäus Kleindessner, Jamie Morgenstern, and Xuezhi Wang. Evaluating fairness of machine learning models under uncertain and incomplete information. In *2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 206–214, 2021.

[Bacelar, 2021] Marley Bacelar. Monitoring bias and fairness in machine learning models: A review. *ScienceOpen Preprints*, 2021.

[Bechavod and Ligett, 2017] Yahav Bechavod and Katrina Ligett. Penalizing unfairness in binary classification. *arXiv:1707.00044*, 2017.

[Beutel *et al.*, 2017] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv:1707.00075*, 2017.

[Calders and Verwer, 2010] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, 21(2):277–292, 2010.

[Caton and Haas, 2020] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *arXiv:2010.04053*, 2020.

[Chen *et al.*, 2019] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Conference on fairness, accountability, and transparency*, pages 339–348, 2019.

[Chiappa, 2019] Silvia Chiappa. Path-specific counterfactual fairness. In *AAAI*, volume 33, pages 7801–7808, 2019.

[Chierichetti *et al.*, 2019] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvtiskii. Matroids, matchings, and fairness. In *22nd Conference on Artificial Intelligence and Statistics*, pages 2212–2220. PMLR, 2019.

[Corbett-Davies *et al.*, 2017] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *23rd acm sigkdd International Conference on knowledge discovery and data mining*, pages 797–806, 2017.

[Detassis *et al.*, 2020] Fabrizio Detassis, Michele Lombardi, and Michela Milano. Teaching the old dog new tricks: supervised learning with constraints. In *NeHuAI@ ECAI*, pages 44–51, 2020.

[Dwork *et al.*, 2012] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *3rd innovations in theoretical computer science Conference*, pages 214–226, 2012.

[Dwork *et al.*, 2018] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on fairness, accountability and transparency*, pages 119–133. PMLR, 2018.

[Edwards and Storkey, 2015] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv:1511.05897*, 2015.

[Feldman *et al.*, 2015] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *21th ACM International Conference on knowledge discovery and data mining*, pages 259–268, 2015.

[Feng *et al.*, 2019] Rui Feng, Yang Yang, Yuehan Lyu, Chenhao Tan, Yizhou Sun, and Chunping Wang. Learning fair representations via an adversarial framework. *arXiv:1904.13341*, 2019.

[Goh *et al.*, 2016] Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. *Advances in Neural Information Processing Systems*, 29, 2016.

[Gupta *et al.*, 2018] Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. Proxy fairness. *arXiv:1806.11212*, 2018.

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

[Heintz *et al.*, 2021] Fredrik Heintz, Michela Milano, and Barry O'Sullivan. *Trustworthy AI-Integrating Learning, Optimization and Reasoning*. Springer, 2021.

[Howard *et al.*, 2016] Rebecca Howard, Anne Tallontire, Lindsay Stringer, and Rob Marchant. Which "fairness", for whom, and why? an empirical analysis of plural notions of fairness in fairtrade carbon projects, using q methodology. *Env science & policy*, 56:100–109, 2016.

[Ignatiev *et al.*, 2020] Alexey Ignatiev, Martin C Cooper, Mohamed Siala, Emmanuel Hebrard, and Joao Marques-

Silva. Towards formal fairness in machine learning. In *CP*, pages 846–867. Springer, 2020.

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

[Kamiran *et al.*, 2010] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*, pages 869–874. IEEE, 2010.

[Kamishima *et al.*, 2012] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on machine learning and knowledge discovery in databases*, pages 35–50. Springer, 2012.

[Kilbertus *et al.*, 2017] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30, 2017.

[Krasanakis *et al.*, 2018] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *2018 WWW Conference*, pages 853–862, 2018.

[Kusner *et al.*, 2017] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Adv. in neural information processing systems*, 30, 2017.

[Leavy *et al.*, 2021] Susan Leavy, Eugenia Siapera, and Barry O'Sullivan. Ethical data curation for ai: An approach based on feminist epistemology and critical theories of race. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 695–703, 2021.

[Liu and Vicente, 2021] Suyun Liu and Luis Nunes Vicente. The sharpe predictor for fairness in machine learning. *arXiv:2108.06415*, 2021.

[Lohia *et al.*, 2019] Pranay K Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R Varshney, and Ruchir Puri. Bias mitigation post-processing for individual and group fairness. In *International Conference on acoustics, speech and signal processing*, pages 2847–2851. IEEE, 2019.

[Louizos *et al.*, 2015] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv:1511.00830*, 2015.

[Luong *et al.*, 2011] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. k-nn as an implementation of situation testing for discrimination discovery and prevention. In *17th International Conference on Knowledge discovery and data mining*, pages 502–510, 2011.

[Madras *et al.*, 2018] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on ML*, pages 3384–3393. PMLR, 2018.

[McCarthy and Narayanan, 2023] Michael B McCarthy and Sundaraparipurnan Narayanan. Fairness–accuracy tradeoff: activation function choice in a neural network. *AI and Ethics*, pages 1–10, 2023.

[Mehrabi *et al.*, 2021] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), jul 2021.

[Menon and Williamson, 2018] Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pages 107–118. PMLR, 2018.

[Mhasawade and Chunara, 2021] Vishwali Mhasawade and Rumi Chunara. Causal multi-level fairness. In *AAAI/ACM Conf. on AI, Ethics, and Society*, pages 784–794, 2021.

[Pessach and Shmueli, 2021] Dana Pessach and Erez Shmueli. Improving fairness of artificial intelligence algorithms in privileged-group selection bias data settings. *Expert Systems with Applications*, 185:115667, 2021.

[Pessach and Shmueli, 2022] Dana Pessach and Erez Shmueli. A review on fairness in ml. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[Pleiss *et al.*, 2017] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

[Quadrianto and Sharmanska, 2017] Novi Quadrianto and Viktoriia Sharmanska. Recycling privileged learning and distribution matching for fairness. *Advances in Neural Information Processing Systems*, 30, 2017.

[Robertson and Kesselheim, 2016] Christopher T Robertson and Aaron S Kesselheim. *Blinding as a solution to bias: Strengthening biomedical science, forensic science, and law*. Academic Press, 2016.

[Wang *et al.*, 2019] Hao Wang, Berk Ustun, and Flavio Calmon. Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In *International Conference on ML*, pages 6618–6627. PMLR, 2019.

[Woodworth *et al.*, 2017] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953. PMLR, 2017.

[Zafar *et al.*, 2017a] Muhammad B. Zafar, Isabel Valera, Manuel G. Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *AI and Statistics*, pages 962–970. PMLR, 2017.

[Zafar *et al.*, 2017b] Muhammad Bilal Zafar, Isabel Valera, Manuel G. Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *26th International Conference on WWW*, pages 1171–1180, 2017.

[Zemel *et al.*, 2013] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on ML*, pages 325–333. PMLR, 2013.