What Is Fairness? Implications For FairML

Ludwig Bothmann
Department of Statistics
LMU Munich, Germany
ludwig.bothmann@stat.uni-muenchen.de

Kristina Peters
Faculty of Law
LMU Munich, Germany
kristina.peters@jura.uni-muenchen.de

Bernd Bischl
Department of Statistics
LMU Munich, Germany
bernd.bischl@stat.uni-muenchen.de

Abstract

A growing body of literature in fairness-aware ML (fairML) aspires to mitigate machine learning (ML)-related unfairness in automated decision making (ADM) by defining metrics that measure fairness of an ML model and by proposing methods that ensure that trained ML models achieve low values in those measures. However, the underlying concept of fairness, i.e., the question of what fairness is, is rarely discussed, leaving a considerable gap between centuries of philosophical discussion and recent adoption of the concept in the ML community. In this work, we try to bridge this gap by formalizing a consistent concept of fairness and by translating the philosophical considerations into a formal framework for the evaluation of ML models in ADM systems. We derive that fairness problems can already arise without the presence of protected attributes, pointing out that fairness and predictive performance are not irreconcilable counterparts, but rather that the latter is necessary to achieve the former. Moreover, we argue why and how causal considerations are necessary when assessing fairness in the presence of protected attributes. Eventually, we achieve greater linguistic clarity for the discussion of fairML by clearly assigning responsibilities to stakeholders inside and outside ML.

1 Introduction

The machine learning (ML) community has produced numerous contributions on the topic of fairness-aware ML (fairML) in recent years. However, a fundamental question remains: **What is fairness**? This question is not so easy to answer and often skipped; instead of asking “what is fairness”, the questions of “how to measure fairness of ML models” and “how to make ML models fair” are pursued. This paper does not intend to criticize individual approaches that address those latter questions and often propose important solutions.

Rather, the aim is to make explicit the premises that underlie the various understandings of fairness and the approaches to solving fairness problems. In doing so, a largely concordant understanding can be elaborated that is based on a rich foundation in the history of philosophy. Subsequently, we show that the conception of fairness depends on multilayered normative evaluations; any discussion of fairML is reliant on adopting those normative stipulations. The basis for fair decisions is always the question of the equality of the people treated with respect to the subject matter concerned. With this decision basis, a decision rule is to be established, which in turn can be adapted to the concrete needs as a result of normative stipulations. Based on this basic concept of fairness, we turn to the question of to what extent ML models can induce unfair treatments in automated decision making (ADM).

Preprint. Under review.

1.1 Our Contributions

- To the best of our knowledge, this is the first contribution that formalizes a **consistent concept of fairness** derived from philosophical considerations and translates it into a formal framework for the evaluation of ML models in ADM systems, hence bridging the gap between centuries of philosophical discussion and the recent adoption in the ML community.
- By doing so, we precisely delineate fairML's contribution to questions of fairness in ADM from the responsibilities of other scientific fields as well as from the socio-political discussion and from fundamental legislative decisions, with the result of **greater linguistic clarity**.
- We argue that an ML model cannot be unfair **per se** due to the lack of an action aspect, and derive that fairness problems can arise **using ML models in ADM systems** if the model is not **individually well-calibrated** – even if no protected attributes are present – hence pointing out that predictive performance is paramount for fairness rather than a tradeoff to it.
- We argue that in the presence of protected attributes, a **fictitious, normatively desirable world** is conceived with respect to which ML models are to be trained and evaluated, conclude that fairness criteria must make causal considerations, and elaborate on how counterfactuals should be used for fairness evaluation.
- A practical advantage of our approach (compared to, e.g., [16], who acknowledge that "[...] access to the distance metric [...] is one of the most challenging aspects of our framework") is that **we only require non-ML users to answer three normative questions** that they are already accustomed to answering – also without using ML.
- In addition, we offer a concept for **fair but unequal treatment of unequal individuals**.

1.2 Related Work

What exactly is understood by fairness is not disclosed by central laws [14; 17] or statements from politics [18]. Because of this, there exists a broad and now almost unmanageable body of literature on the topic of fairness in general and, in particular, on the fairness of ADM – especially in the social sciences, law, and more recently in ML [for overviews see, e.g., 1; 5; 8; 12; 13; 37]. Most of their arguments, however, are presuppositional and start a step further than we do here. Explanations of what fairness really means, if they occur, usually turn out to be vague [7; 15; 21; 26; 31].

There is a growing awareness in the ML literature that basic assumptions that are not made explicit do indeed matter [19; 36; 38], but the reappraisal of these basic assumptions is still in its infancy. Adding to the many concepts of **group fairness** [see, e.g., 46, for an overview], most recently, the notion of **individual fairness**, highlighted in particular by [16], has received considerable resonance [6; 13; 19; 34]. This requires that similar individuals should be treated similarly and reflects a demand already formulated comparably by [22] more than 50 years ago (see also [25; 45] regarding the debate on test (un)fairness in mid 20th century). Compared to [16], [43] assume for a ranking task that "[u]nfairness occurs when an agent with higher merit obtains a worse outcome than an agent with lower merit." At first glance, [34] go in a different direction with their concept of **individual counterfactual fairness**, according to which a fair decision exists if it turns out to be the same in the real world and in a fictitious world in which the individual in question belongs to a different group. As described later, these definitions – which take up causal concepts [see also 27; 29; 33] – seem to produce useful results and are also supported by our considerations, but nevertheless do not represent essential aspects of the fairness concept. We note that our concept can be used to resolve confusions that have recently arisen in view of supposedly different definitions of fairness [19; 33; 34; 38].

In order to reveal the basic formal structure of the concept of fairness, it is not enough to go back to the great works of the 20th century, since these also take a step further and usually focus on which material criteria should be taken into account in the context of a just or fair distribution [10]. These considerations already build on the concept that we will present below [41; 42]. For this reason, our considerations lend themselves equally well as a complement to the manageable ML literature that ties in with corresponding theories of distributive justice [9; 20; 27; 32].

2 The basic structure of fairness

The general understanding of fairness is regularly characterized as (i) typically concerning the treatment of people by people [3; 11; 15; 31] and (ii) not being described as a concept, but merely by referring to normatively charged synonyms – such as justice, equality, or absence of discrimination. At first glance, one might think that a more detailed definition of the term is superfluous and that fairness may be difficult to define, yet intuitively graspable. Suppose, for example, that we have a cake from which two people are to receive a portion. It seems (initially) "fair" if each person gets one half of the cake. Fairness, then, is equated with "equality" or "equal treatment". But what if one person is starving and the other is well-fed and satiated? What if the cake is supposed to be a reward for a service previously rendered and one person has done twice as much as the other?

2.1 Basis for decision: equality or inequality

In these considerations, sometimes referred to as the difference between "equality" and "equity", lies the basic problem of arguments about fairness. These always depend on the reference point of the evaluation: Is it solely the distribution of the asset at hand? Or should the point of reference also be the person concerned? What is fair is then determined by who is affected.

These rather trivial considerations can be translated into a theoretical framework, i.e., a **formal basic structure**. The fundamental aspects of this concept were already developed by Aristotle in his analysis of the nature of justice [3] (which is not being considered by [36; 47] in their approaches to invite Aristotle into the discussion about fairness) and still provide a viable foundation for contemplating fairness today. Here, justice can be understood as mere adherence to the standards agreed upon in society (e.g., in laws), but also refers to the idea of equality. This idea of equality is the common root of what is meant by "justice" or "fairness" when used as a critical concept. Equality demands that equals are to be treated equally and unequals unequally. In other words, if unequals are treated equally, this is unfair. If equals are treated unequally, this is also unfair.

Consequently, the decision-making basis for treatment is the question of whether or not people are equal. However, equality is a **strongly normatively loaded term**, because people possess infinitely many qualities and are therefore never equal. In relation to certain situations, however, there is a normative stipulation that this difference between people should be irrelevant. Aristotle stated that this is particularly the case in the private relations: if two people conclude a contract and it is a question of whether performance and consideration are balanced, it is irrelevant who these people are. Similarly, it is irrelevant whether a rich person kills a poor person or vice versa. Because the distributional decision here can be made by means of a simple calculation, Aristotle speaks in this respect of **arithmetic or continuous proportionality**.

In other situations, however, equality is said to depend on the characteristics of the people concerned; the relevance of the characteristics for the assessment of equality is decided normatively. Aristotle calls this the "worthiness" of people. For example, take the tax rate: In many societies, those who earn more also pay a higher tax rate, and vice versa. Because this kind of distribution decision must consider the balance of a more complex ratio, Aristotle speaks here of the **geometric or discrete proportionality**. The distribution ratio results in dependence on the worthiness negotiated in the political dispute: the ratio of the worthiness of person i to the assets distributed to them must correspond to the ratio of the worthiness of person j to the assets distributed to them.

As evidenced, **equality is always the result of a normative stipulation**. This is accompanied by an evaluation of what is to be brought into a relationship of equality – only the things or assets that are distributed (arithmetic proportionality), or also the people (geometric proportionality). For Aristotle, this depends on the subject area concerned: private dealings are decided by arithmetic proportionality, and government distribution is decided by geometric proportionality. Today, it is part of the political dispute whether private matters may be left in this sphere or whether state intervention according to the principles of the state distributive system is deemed necessary. Due to the normativity of the concept of equality, the assessment of whether actions constitute equal or unequal treatment – and are "fair" or "unfair" –, can vary widely depending on the system of norms involved. However, certain moments of consensus are now emerging, at least in certain regions of the world – for example, with regard to the unequal treatment of women or ethnic groups.

Nowadays, the so-called **Protected Attributes** (PAs) play a special role in the decision-making process, e.g., the characteristics listed in the US Civil Rights Act of 1964, in the Charter of Fundamental Rights of the European Union, or in Article 3 of the German Constitution. In some cases, the comparison of two people must not be based on these PAs (**PA-neutrality**). Because the attributes, if they do not exist, cannot act as causal factors either, this decision is accompanied by the consideration to ignore the consequences of these attributes as well, e.g.: if ethnicity has an influence on an offender's probability of recidivism, it seems natural to choose not to take this into account. Here, an underlying consideration may be that ethnicity is not the direct reason for the higher probability of recidivism, but rather that ethnicity has complex consequences for socialization processes, which then in turn have an effect on the higher recidivism probability, e.g., an average lower level of education or a certain place of residence. A society may take responsibility for these consequences, wanting to keep them out of the decision-making process. However, because no monocausal processes occur in the life of an individual – i.e., in the example, ethnicity is not the only causal factor for the level of education – this is again a social negotiation process, at the end of which a normative decision is made as to who is to be attributed responsibility for which processes. These considerations make it evident that eventually a fictitious world massively corrected by normative evaluations becomes the basis for deciding on the treatments of individuals.

Conversely, there are constellations in which the PAs are specifically targeted in order to justify the inequality of people and thus their unequal treatment in the form of a preference for the feature bearers (**PA-focus**). The perspective depends on a normative decision as to whether the protection of the feature bearers in the respective subject area is to be ensured "only" by means of an exclusion of these features, or whether reality is to be actively reshaped according to certain objectives. In the words of Aristotle, in the first case, a diminished worthiness must not be based on the protected attributes, while in the second case, the same attribute is invoked to establish a higher worthiness. Thus, even the complex reality of today still fits into Aristotle's concept – while at the same time it is evident that multi-layered normative questions are involved here. Even in the area of PAs, there is thus no fixed "equal treatment" or "fairness".

2.2 Decision rule: equal treatment

Once the basis for decision has been established, a decision rule must be drawn up in a second step, which determines the extent to which equality or inequality is to be taken into account. In general, this is given by the idea of equality or proportionality. In the case of geometric proportionality, it must hold that the ratio of treatment t to worthiness w is the same for any comparison of two individuals i and j, i.e., $\frac{t^{(i)}}{w^{(i)}} = \frac{t^{(j)}}{w^{(j)}} = k \quad \forall i, j \in \{1, \dots, n\}$. In other words, we can define a treatment function $s : W \to T,\ w \mapsto t$, where typically $W \subseteq \mathbb{R}$ and $T \subseteq \mathbb{R}$:

Definition 1 (Geometrically fair treatment). A treatment $t^{(i)}$ of an individual i is called **geometrically fair** iff it is a linear function of the individual's worthiness $w^{(i)}$, i.e., $t^{(i)} = s(w^{(i)}) = k \cdot w^{(i)}$, where the value of $k \in \mathbb{R}$ is a normative choice.

In the simpler case of arithmetic proportionality, a fair treatment does not depend on the worthiness:

Definition 2 (Arithmetically fair treatment). A treatment $t^{(i)}$ of an individual i is called **arithmetically fair** iff it is the same as for any other individual, i.e., $t^{(i)} = s(w^{(i)}) = k\ \forall i$ does not depend on the individual's worthiness $w^{(i)}$, where the value of $k \in \mathbb{R}$ is a normative choice.

However, it may be decided to modify the treatment function $s(\cdot)$ in a more flexible way, e.g., if a tax rate is raised with higher income, but this is done step-wise rather than continuously – and above a certain income, not at all. The function $s(\cdot)$ can thus be normatively corrected:

Definition 3 (Fair treatment). A treatment $t^{(i)}$ of an individual i is called **fair** iff it is determined by a normative function of the individual's worthiness $w^{(i)}$, i.e., $t^{(i)} = s(w^{(i)})$, where $s(\cdot)$ is a monotonic function.

This guarantees (i) equal treatment of equals (as measured by w) and (ii) unequal treatment of unequals in a normatively defined way that is then considered to be fair. Even if it would be mathematically straight-forward to allow for $s(\cdot)$ being more flexible, a non-monotonic $s(\cdot)$ would not be in line with the above philosophical concept. Note that one will demand strict monotonicity if unequals are always to be treated unequally, and that geometrically and arithmetically fair treatments result by special choices of $s(\cdot)$. Thus, again, valuation decisions arise that add another normative dimension to the decision process. The duty of ML in the ADM process will be to provide an estimate of the individual's worthiness $w^{(i)}$, which serves as basis for the decision of their treatment.

3 Role of ML in the ADM process

Now that we have defined the general concept of fairness, we can turn to answering the question of to what extent an ML model used in the context of an ADM process can be unfair. First, in Section 3.1, we will review the steps of an ADM process with respect to the applicability of the notion of fairness and illustrate the application of the concept using the COMPAS example.

3.1 Fairness in the ADM process

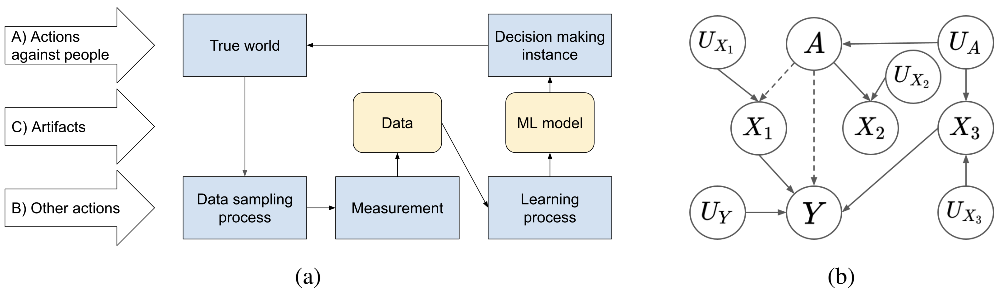

Figure 1: (a) Simplified ADM process, following [44]. (b) Following the notation of [34], A is a PA, X_* are sets of observable features, U_* are latent background variables, Y is the output. Arrows denote causal effects; in the fictitious world, dashed lines are removed. Note that PA may still have causal effects on a set X_2 of observable features, as long as this does not imply effects on Y.

The ADM process is explained in detail in [44], we consider a reduced version (see Figure 1a). We divide the process into three categories: (A) actions against people, (B) other actions, and (C) artifacts. As shown in Section 2, the notion of fairness is applicable only to category (A) and, hence, not directly applicable to the ML model. However, the results of categories (B) and (C) establish the basis of decisions for actions against people: the basis of decisions – called "worthiness" in the prior section – is the measure of equality of individuals, with respect to the decision to be made. In a classification scenario, this metric is the individual probability of success $\mathbb{P}(y = 1|i) = \pi^{(i)}$; in a regression scenario, it is the individual expected value $\mathbb{E}(y|i) = \mu^{(i)}$. Thus, indirectly, the elements of categories (B) and (C) can induce unfair actions. (We will hold to the classification scenario in the following for notational clarity, but switching to the regression scenario is straightforward, remedying a critique towards fairML's focus on classification formulated by [25].) Note that the normative decision can be made to consider not only a $\pi^{(i)}$ – which is yet to be estimated – as the basis for decision-making, but to consider additional factors; the treatment is then a function $s(\pi^{(i)}, \mathbf{v}^{(i)})$ of $\pi^{(i)}$ and other variables $\mathbf{v}^{(i)}$ that are not to be used as input to the ML model. Before we shed more light on the role of ML in this process, we highlight the crucial difference between the ML model and actions based on it and analyze the effect of the occurrence of PAs.

In the much-cited COMPAS example [2], the individual probability $\pi^{(i)}$ of recidivism within two years ($y \in \{0, 1\}$) [35] is the basis for deciding how to treat defendant i. Since $\pi^{(i)}$ is unknown, an ML model $\hat{\pi}(\cdot)$ is used to estimate it, based on a feature vector $\mathbf{x}^{(i)}$. The goal of the ML model [see, e.g., 40] is to assign different – and as accurately as possible – recidivism probabilities to different individuals, based on the observed features:

Definition 4 (Statistically discriminative model w.r.t. feature X**).** An ML model $\hat{\pi}(\cdot)$ is called **statistically discriminative w.r.t. feature** X if there is at least one pair of individuals i and j who differ only with respect to feature X and are assigned different predicted probabilities $\hat{\pi}^{(i)} \neq \hat{\pi}^{(j)}$.

This is either the case or not the case for each feature; however, a fairness evaluation can only be applied to the action based on it. From the definitions above, we derive a similar notion as [22]:

Definition 5 (Descriptively unfair treatment). Assume a pair of individuals i and j who differ only with respect to feature X. Assume that feature X is not a causal reason for a difference in the true probabilities, i.e., $\pi^{(i)} = \pi^{(j)}$. A treatment is called **descriptively unfair w.r.t. feature** X if these individuals are treated differently, i.e., $t^{(i)}(= s(\hat{\pi}^{(i)})) \neq t^{(j)}(= s(\hat{\pi}^{(j)}))$, in a process due to differing estimated individual probabilities $\hat{\pi}^{(i)} \neq \hat{\pi}^{(j)}$.

Example 1: Assuming that the recidivism probability π does not causally depend on the ethnicity X, yet two persons who differ only with respect to ethnicity (i.e., have the same true recidivism probability) would be assumed by the ML model to have different recidivism probabilities, and therefore there would be different judicial decisions, then this unequal treatment would be descriptively unfair – regardless of whether ethnicity is a PA or not, because equals would be treated unequally.

If, on the other hand, X is causal for a difference in π, then a differing decision basis due to X, i.e., $\pi^{(i)} \neq \pi^{(j)}$ and a resulting difference in treatment, i.e., $t^{(i)} \neq t^{(j)}$ cannot be said to be descriptively unfair – unequals are treated unequally. This evaluation may change with the introduction of PAs, as derived in Section 2 and similarly noted by [36], as the assessment of equality changes:

Definition 6 (Normatively unfair treatment). Assume a pair of individuals i and j who differ only with respect to feature A. Assume that feature A is a causal reason for a difference in the true probabilities, i.e., $\pi^{(i)} \neq \pi^{(j)}$. Assume that feature A is a PA. A treatment is called **normatively unfair w.r.t. feature** A if these individuals are treated differently, i.e., $t^{(i)}(= s(\hat{\pi}^{(i)})) \neq t^{(j)}(= s(\hat{\pi}^{(j)}))$, in a process due to differing estimated individual probabilities $\hat{\pi}^{(i)} \neq \hat{\pi}^{(j)}$, as feature A must not be invoked for the determination of equality, i.e., the decision basis for the treatment.

Example 2: Suppose that the probability of recidivism π depends causally on the ethnicity X, but the judicial decision should not depend on the ethnicity for normative reasons, e.g., because society decides not to let the individual bear the responsibility for the grievance that ethnicity is causal for the probability of recidivism, but to take it from them and bear it as a whole society.

So, the measure of equality here is no longer the true recidivism probability π in the real world, but rather the true recidivism probability $\tilde{\pi}$ in a fictitious, normatively desired world – a world where the PAs have no causal effects on the probability, neither directly nor indirectly, i.e., $\tilde{\pi}^{(i)} = \tilde{\pi}^{(j)}$ for the pair considered in Definition 6 (see also Figure 1b). Once we have moved to this world after defining the PAs, we can use $\tilde{\pi}^{(i)}$ (instead of $\pi^{(i)}$) as a measure of equality; hence, a fair treatment would result by $t^{(i)} = s(\tilde{\pi}^{(i)})$. The role of ML is to estimate $\pi^{(i)}$ or $\tilde{\pi}^{(i)}$ as accurately as possible.

3.2 Contribution of ML – without PA

We first consider the situation where no PA is present and show that, even then, the use of an ML model can induce fairness problems. In Section 3.3, we consider the situation where PAs are present.

3.2.1 Can ML induce unfairness, and if so, how?

In Section 2, we saw that the treatment $t^{(i)}$ of an individual i based on a decision basis $w^{(i)}$ is fair iff it follows the normative decision rule $s(\cdot)$, i.e., iff $t^{(i)} = s(w^{(i)})$. In Section 3.1, we saw that in the context of an ADM process, the decision basis $w^{(i)}$ often corresponds to the individual probability $\pi^{(i)} = \mathbb{P}(y = 1|i)$ for an event $y \in \{0, 1\}$. Since the true individual probability $\pi^{(i)}$ is unknown in practice, we use data to estimate $\pi^{(i)}$. In doing so, two key steps introduce imprecision that can induce unfair treatment in the outcome and are described below (see also Table 1).

Table 1: Coarse information due to finite feature space $\mathcal{X}$ and estimation via ML introduce errors. Point of reference changes with introduction of PAs (last row, see Section 3.3).

	Fair treatment	Treatment w/ coarse information	Treatment w/ ML model
True world	$s(\pi^{(i)})$	$s(\pi(\mathbf{x}^{(i)}))$	$s(\hat{\pi}(\mathbf{x}^{(i)}))$
Fictitious world	$s(\tilde{\pi}^{(i)})$	$s(\tilde{\pi}(\tilde{\mathbf{x}}^{(i)}))$	$s(\hat{\tilde{\pi}}(\tilde{\mathbf{x}}^{(i)}))$

Coarsening of information. The first step is to coarsen the information by basing the treatment not on the true individual probability $\pi^{(i)}$ but on a group probability $\pi(\mathbf{x}^{(i)})$ that assigns the same value to all individuals $I_{\mathbf{x}} = \{i : \mathbf{x}^{(i)} = \mathbf{x}\}$ with the same combination of observed features $\mathbf{x}$. Naturally, the function $\pi : \mathcal{X} \rightarrow [0, 1]$ is as true and unknown as the individual probabilities $\pi^{(i)}$. For any feature combination $\mathbf{x}$, this function is the best possible approximation of the different $\pi^{(i)}$ of the different individuals $I_{\mathbf{x}}$ sharing that feature combination, based on the available p features.

This coarsening of information introduces fairness problems, since individuals $I_{\mathbf{x}}$ are treated the same even though they are – except for the features collected – not the same. Except for degenerate special cases where all individuals in a group $I_{\mathbf{x}}$ are identical or for a few individuals where $\pi^{(i)} = \pi(\mathbf{x}^{(i)})$ happens to hold, this results in $\pi^{(i)} \neq \pi(\mathbf{x}^{(i)}) \Rightarrow s(\pi^{(i)}) \neq s(\pi(\mathbf{x}^{(i)})) = t^{(i)} \; \forall i \in \{1, \ldots, n\}$ (if $s(\cdot)$ is injective), so (almost) all individuals are treated unfairly – even when $\pi(\cdot)$ is known.

This step is only appreciated in a few works, such as [19; 31; 38; 43]. However, it is very important for fairness considerations of ADM systems, since this means that already reducing the information regarding an individual to finitely many features is a gateway to unfairness – even if we knew the true $\pi(\cdot)$, and even before we considered how to estimate $\pi(\cdot)$ and also before introducing PAs.

Estimation by ML. Estimating the unknown function $\pi(\cdot)$ by $\hat{\pi}(\cdot)$ introduces two types of error: The **estimation error** comes from the fact that the learner has only a finite amount of training data $\mathcal{D} = \left(\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(n)}, y^{(n)}\right)\right)$ available. This error converges to 0 for $n \to \infty$. Should the true $\pi(\cdot)$ not be part of the hypothesis space $\mathcal{H}$, a non-reducible **approximation error** remains.

3.2.2 Evaluation regarding fairness

Evaluation of the ML model $\hat{\pi}(\cdot)$ can be done at two levels, namely (i) with respect to $\pi(\cdot)$ or (ii) with respect to $\pi^{(i)}$. Considering level (i), for cases in which every conceivable feature combination $\mathbf{x} \in \mathcal{X}$ is represented sufficiently often (only possible in the case of few categorical features), the mean observed and predicted probabilities can be compared directly – for example, via the L2 norm of the differences of the group means. This value should be as small as possible on a test set. (Note, however, that using the L2 norm is another normative choice.) Since this (a) is only conceivable for special data situations and (b) ignores the imprecision introduced by coarsening $\pi^{(i)}$ via $\pi(\mathbf{x}^{(i)})$, an evaluation directly with respect to $\pi^{(i)}$ seems more purposeful: We recall that a treatment $t^{(i)}$ is fair iff $t^{(i)} = s(\pi^{(i)})$, so here iff $s(\hat{\pi}(\mathbf{x}^{(i)})) = s(\pi^{(i)})$. Thus, a sufficient condition for a fair treatment is $\hat{\pi}(\mathbf{x}^{(i)}) = \pi^{(i)}$, which can be seen as an individual version of well-calibration [31; 46]:

Definition 7 (Individually well-calibrated model): An ML model $\hat{\pi}(\cdot)$ is called **individually well-calibrated for individual** i if $\hat{\pi}^{(i)} = \pi^{(i)}$; it is called **individually well-calibrated** if $\hat{\pi}^{(i)} = \pi^{(i)} \; \forall i$.

(Note that a direct generalization to regression is possible by replacing true and estimated probability by true and estimated expectation.) For a strictly monotone function $s(\cdot)$, this is also a necessary condition for fairness, i.e., a treatment $s(\hat{\pi}(\mathbf{x}^{(i)}))$ is to be called unfair if the model is not individually well-calibrated for individual i. Although the notion of fairness refers to actions against individuals, empirical evaluation cannot be performed individually; evaluation is only possible by considering appropriate groups. This poses a massive problem, since any group definition runs the risk of assigning individuals to a group that is inappropriate with respect to their true probability $\pi^{(i)}$. In particular, it falls short to define a group based only on a single feature. Rather, it is necessary to consider all computationally identifiable subgroups, as is done, e.g., in [19; 24; 28; 30], and check well-calibration on all these subgroups.

3.3 Contribution of ML – with PA

Even without the presence of PAs, the use of empirical methods can induce fairness problems. We now analyze what changes through the introduction of PAs. Again, we consider first the process of finding a model and then the question of evaluating the model with respect to fairness.

3.3.1 Can ML induce unfairness, and if so, how?

In Section 3.1, it was shown that in the presence of PAs, in order to be able to achieve a fair treatment, the basis for decisions is PA-neutral, i.e., the normative decision can be made to use the PA-neutral true probability $\tilde{\pi}^{(i)}$ instead of $\pi^{(i)}$ as a measure of equality. This $\tilde{\pi}^{(i)}$ describes for an individual i the probability $\mathbb{P}(\tilde{y} = 1|i)$ for an event $\tilde{y} \in \{0, 1\}$ in a fictitious, normatively desired world in which the PA has no causal influence on this event, neither directly nor indirectly. As above, treatment $t^{(i)}$ of an individual i is then said to be fair iff it follows the normatively specified decision rule $s(\tilde{\pi}^{(i)})$, with changed decision basis $\tilde{\pi}^{(i)}$. Thus, in the COMPAS example, we consider the individual recidivism probability $\tilde{\pi}^{(i)}$, **in a fictitious world where the PA has no causal influence on recidivism**.

Since the PA is not supposed to have a causal influence on $\tilde{y}$ in this fictitious world, the normative decision can be made to exclude certain indirect influences of the PA on $\tilde{y}$ via other features. Therefore, **the feature vector of individual i is also to be considered PA-neutral**, i.e., $\tilde{\mathbf{x}}^{(i)}$. This does not mean that the PA is removed (hence, goes beyond "fairness through unawareness" [20]), nor that all influences of the PA are fictitiously removed. Rather, only those that (possibly via detours) have an effect on $\tilde{y}$ are removed (see dashed lines in Figure 1b). Note that, while related, this is a different idea than that formulated in Lemma 1 of [34], where the prediction is a function of the non-descendants of A; rather, we allow the prediction to be a function of the descendants of A (set X_1 in Figure 1b) as well but rely on the fictitious world counterpart $\tilde{X}_1$ of these features. In the COMPAS example, individuals might not only have a different recidivism probability when the influence of the PA is removed, but also, e.g., a different income or residence, assuming that in the real world the PA also has a causal influence on these features.

There is also **no reason to assume that the relation of event and features is identical in the real and fictitious world after potentially causal influences are removed**, so the function $\tilde{\pi}(\cdot)$ we are looking for is also potentially different from $\pi(\cdot)$. In the COMPAS example, the influence of features such as income and residence on the recidivism probability might be different in the two worlds.

Thus, we are confronted with a different situation than in Section 3.2, since we want to learn contexts in a fictitious world, considering a different data generating process. We can now again consider the two steps of introducing imprecision in empirical use cases (see also Table 1, bottom row).

Coarsening of information. Again, the first step is a coarsening of information, since instead of individual probabilities $\tilde{\pi}^{(i)}$, group probabilities $\tilde{\pi}(\tilde{\mathbf{x}}^{(i)})$ are considered. The function $\tilde{\pi}(\cdot)$ is again true (in the fictitious world) and unknown to us, and $\tilde{\pi}(\tilde{\mathbf{x}}^{(i)})$ is the best approximation of $\tilde{\pi}^{(i)}$ possible via $\tilde{\mathbf{x}}^{(i)}$. Thus, by the same reasoning as above, even knowing $\tilde{\pi}(\cdot)$, treatment based on $\tilde{\pi}(\tilde{\mathbf{x}}^{(i)})$ can be unfair, since unequals (with respect to $\tilde{\pi}^{(i)}$) are treated equally (since they have identical $\tilde{\mathbf{x}}^{(i)}$) – and this is before we have even considered estimating $\tilde{\pi}(\cdot)$.

Estimation by ML. In the second step, the function $\tilde{\pi}(\cdot)$ is estimated by $\hat{\tilde{\pi}}(\cdot)$, introducing an error. This estimation should be performed on a random sample $\tilde{\mathcal{D}} = \left(\left(\tilde{\mathbf{x}}^{(1)}, \tilde{y}^{(1)}\right), \ldots, \left(\tilde{\mathbf{x}}^{(n)}, \tilde{y}^{(n)}\right)\right)$ of the fictitious world. The practical problem is that if we do not have these data from the fictitious world, we must rely on data $\mathcal{D}$ from the real world. However, with these data, in turn, we do not estimate $\tilde{\pi}(\cdot)$, but rather $\pi(\cdot)$. Even if we had a one-time access to the fictitious world, providing us training data $\tilde{\mathcal{D}}$ with which we could determine $\hat{\tilde{\pi}}(\cdot)$, we would again have only $\mathbf{x}^{(i)}$ available at prediction time, so we could not use the learned model for predicting $\hat{\tilde{\pi}}(\tilde{\mathbf{x}}^{(i)})$. To effectively solve this problem, we would need to develop a way to create mappings from the real to the fictitious world, i.e., $m_y : \mathcal{Y} \rightarrow \mathcal{Y},\ y \mapsto \tilde{y}$ and $m_{\mathbf{x}} : \mathcal{X} \rightarrow \tilde{\mathcal{X}},\ \mathbf{x} \mapsto \tilde{\mathbf{x}}$. Naturally, we would still have the usual estimation and approximation error. We now reflect on the evaluation of such a model $\hat{\tilde{\pi}}(\cdot)$.

3.3.2 Evaluation regarding fairness

Let us first assume that, in the absence of $\tilde{\mathcal{D}}$, we have learned a model based on $\mathcal{D}$. We obtain an estimate $\hat{\tilde{\pi}}(\mathbf{x}^{(i)})$, i.e., an approximation of $\tilde{\pi}^{(i)}$ by observable data $\mathbf{x}^{(i)}$. (One possibility is to simply consider $\hat{\pi}(\mathbf{x}^{(i)})$ – although it remains highly unclear how well this model approximates $\tilde{\pi}^{(i)}$.) As above, we can perform the evaluation of the model $\hat{\tilde{\pi}}(\mathbf{x}^{(i)})$ at two levels, namely (i) with respect to $\tilde{\pi}(\cdot)$ or (ii) with respect to $\tilde{\pi}^{(i)}$, where again only the second variant is constructive. The treatment $t^{(i)} = s(\hat{\tilde{\pi}}(\mathbf{x}^{(i)}))$ is fair iff $s(\hat{\tilde{\pi}}(\mathbf{x}^{(i)})) = s(\tilde{\pi}^{(i)})$. Thus, no sufficient condition for fairness is $\hat{\tilde{\pi}}(\mathbf{x}^{(i)}) = \pi^{(i)}\ \forall i$. A sufficient condition for fairness, however, is:

Lemma 1. A treatment $s(\hat{\tilde{\pi}}(\mathbf{x}^{(i)}))$ of individual i is fair if $\hat{\tilde{\pi}}(\mathbf{x}^{(i)}) = \tilde{\pi}^{(i)}$.

For a strictly monotonic $s(\cdot)$, this condition is also necessary for a fair treatment. However, we can only test this condition with access to $\tilde{y}^{(i)}\ \forall i$. Alternatively, we can formulate a necessary condition:

Lemma 2. Assume that $s(\cdot)$ is strictly monotonic; a treatment $s(\hat{\tilde{\pi}}(\mathbf{x}^{(i)}))$ of individual i is only fair if $\hat{\tilde{\pi}}(\mathbf{x}^{(i)}) = \hat{\tilde{\pi}}(\tilde{\mathbf{x}}^{(i)})$.

Then, all who are PA-neutrally equal are treated equally; however, we can only test this condition if we have access to $\tilde{\mathbf{x}}^{(i)}\ \forall i$. This condition is only necessary because, e.g., $\hat{\tilde{\pi}}(\cdot) = 0.5$ satisfies the condition, but treatment based on this condition is not fair unless indeed $\tilde{\pi}^{(i)} = 0.5\ \forall i$. (Moreover, such an ML model is useless because it does not contain any information about individual i. Regardless, the important point is that unequals will also be treated equally under this model, which is contrary to fairness.) Consequently, another condition must be defined to ensure that such trivial models are excluded. Ideally, the chosen model should also perform well – which, again, can only be verified with knowledge of $\tilde{y}^{(i)}\ \forall i$. Alternatively, validity of counterfactuals $\tilde{\mathbf{x}}^{(i)}$ can be required (see below).

Note on the validity of fairness-aware training: Following the idea of post-processing [12; 23], we could train the model on $\mathcal{D}$ and evaluate it with respect to test performance, iteratively satisfying the condition $\hat{\tilde{\pi}}(\mathbf{x}^{(i)}) = \hat{\tilde{\pi}}(\tilde{\mathbf{x}}^{(i)})\ \forall i$ necessary for fairness as well as possible. For this approach we would need a valid method to generate counterfactuals $\tilde{\mathbf{x}}^{(i)}$. However, this methodological challenge is compounded by a more fundamental problem: without knowledge of the $\tilde{y}^{(i)}$, it is unclear to what extent the resulting model (which continues to be trained only on real-world data) can adequately learn the relationships in the fictitious world. Conversely, knowing the $\tilde{y}^{(i)}$ and $\tilde{\mathbf{x}}^{(i)}$, one could train the model directly with $\tilde{\mathcal{D}}$, which would render the post-processing approach obsolete.

We conclude that (i) proof of fairness is possible via checking the sufficient condition of Lemma 1 – if $\tilde{y}^{(i)}$ can be generated, and that (ii) proof of unfairness is possible via the negation of the necessary condition of Lemma 2 – if counterfactuals $\tilde{\mathbf{x}}^{(i)}$ can be generated, whereby an additional quality criterion must be added to exclude trivial (and thus, in practice, useless) counterfactuals.

We have seen that for a valid training of a model in the fictitious world as well as for its evaluation regarding fairness, methods should be found to create mappings from the real to the fictitious world. Some minimal requirements for those mappings are formulated in Appendix A, leaving a thorough treatment of this important topic for further research.

4 Conclusion and Discussion

Despite a rapidly growing body of literature addressing the topic of fairML, there was still a lack of a unified theoretical foundation. By drawing on basic philosophical ideas, the fairness debate could be placed on such a foundation. The basic axioms that underlie the fairness debate even now – without being made explicit – were identified, and their relationship to one another was worked out. It was necessary to separate the basic structure of fairness from material aspects. Based on this, the normative stipulations could be precisely identified, which must take place outside ML. For the case of PA-neutrality, three normative questions must be addressed for the task at hand: (i) Which attributes are defined as PAs (if there are any)? (ii) How is the basis for decisions, i.e., the measure of equality of individuals ($w^{(i)}$), defined? (iii) How is the treatment function ($s(w^{(i)})$) defined?

ML models provide estimates of this basis for decisions and are not unfair per se, but can induce unfairness. Regardless of the presence of PAs, fairness problems can arise with ML, e.g., if the model is not individually well-calibrated. Thus, we do not see a tradeoff between fairness and predictive performance, but predictive performance is essential for fairness. Since individual well-calibration cannot be checked empirically, approaches that consider all computationally identifiable subgroups should be pursued. In general, once PAs are defined, fairness cannot be evaluated easily because the ground truth of the fictitious world (where the causal effects of PAs have been removed) is missing; an approximation to this fictitious world are counterfactuals. The concept of fairness is hence intrinsically related to causal considerations. However, causal questions can, in general, not be answered with purely observational data [29; 34; 39]; for fairML, this means that causal approaches must come into focus even more.

Negative societal impact. As for other technical solutions, overtrusting techniques to overcome fairness problems can have negative societal impacts. However, our hope is that by explicitly tasking non-ML users with answering three broadly understandable normative questions and by linking the ML model evaluation to the respective answers, we contribute a step in the direction of bringing societal needs and technical solutions closer together. Nevertheless, if the answers establish undesirable, e.g., discriminatory norms, then there is no technical possibility to remedy this. Therefore, it is of paramount importance that the normative stipulations – especially in critical applications – are transparent and result in a broad societal discussion.

Limitations and future research. This paper considers fairness questions of ADM systems, where an ML model is used to predict some characteristic of interest in the subject matter concerned. Other applications of ML models – such as diagnostic usage as advocated in, e.g., [4] – are out of scope. Considering PAs, we focus on how to reflect **PA-neutrality** in the ADM process – it would be interesting to tackle incorporating **PA-focus** as well. Furthermore, we derive both that counterfactuals should be used to evaluate fairness as well as how these counterfactuals should be used but do not provide concrete algorithms on how to create these counterfactuals, leaving this for future research.

References

[1] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach. A Reductions Approach to Fair Classification. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR, July 2018. URL `https://proceedings.mlr.press/v80/agarwal18a.html`.

[2] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016. URL `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

[3] Aristotle. *The Nicomachean ethics (book V)*. Oxford World's Classics. Oxford University Press, 2009. ISBN 978-0-19-921361-0. doi: 10.1093/actrade/9780199213610.book.1.

[4] C. Barabas, M. Virza, K. Dinakar, J. Ito, and J. Zittrain. Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 62–76. PMLR, Jan. 2018. URL `https://proceedings.mlr.press/v81/barabas18a.html`. ISSN: 2640-3498.

[5] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. URL `http://www.fairmlbook.org`.

[6] Y. Bechavod, C. Jung, and S. Z. Wu. Metric-Free Individual Fairness in Online Learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11214–11225. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/80b618ebcac7aa97a6dac2ba65cb7e36-Paper.pdf`.

[7] S. Beck, A. Grunwald, K. Jacob, and T. Matzner. *Künstliche Intelligenz und Diskriminierung*. Lernende Systeme, 2019.

[8] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, 50(1):3–44, Feb. 2021. ISSN 0049-1241. doi: 10.1177/0049124118782533. URL `https://doi.org/10.1177/0049124118782533`. Publisher: SAGE Publications Inc.

[9] R. Binns. Fairness in Machine Learning: Lessons from Political Philosophy. In S. A. Friedler and C. Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 149–159. PMLR, Feb. 2018. URL `https://proceedings.mlr.press/v81/binns18a.html`.

[10] C. Calsamiglia. *Decentralizing equality of opportunity and issues concerning the equality of educational opportunity*. Yale University, 2005.

[11] Cambridge Dictionary. Fairness, 2022-05-11. URL `https://dictionary.cambridge.org/de/worterbuch/englisch/fairness`.

[12] S. Caton and C. Haas. Fairness in Machine Learning: A Survey. *arXiv:2010.04053 [cs, stat]*, Oct. 2020. URL `http://arxiv.org/abs/2010.04053`. arXiv: 2010.04053.

[13] A. Chouldechova and A. Roth. The Frontiers of Fairness in Machine Learning. *arXiv:1810.08810 [cs, stat]*, Oct. 2018. URL `http://arxiv.org/abs/1810.08810`. arXiv: 1810.08810.

[14] Y. D. Clarke. Text - H.R.2231 - 116th Congress (2019-2020): Algorithmic Accountability Act of 2019, Nov. 2019. URL https://www.congress.gov/bill/116th-congress/house-bill/2231/text. Archive Location: 2019/2020.

[15] J. Dator. Chapter 3. What Is Fairness? In J. Dator, R. C. Pratt, and Y. Seo, editors, *Fairness, Globalization, and Public Institutions*, pages 19–34. University of Hawaii Press, Honolulu, Dec. 2017. ISBN 978-0-8248-4196-6. doi: 10.1515/9780824841966-004. URL https://www.degruyter.com/document/doi/10.1515/9780824841966-004/html.

[16] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[17] EU. EUR-Lex - 32016R0679 - EN - EUR-Lex, 2016. URL https://eur-lex.europa.eu/eli/reg/2016/679/oj. Doc ID: 32016R0679 Doc Sector: 3 Doc Title: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

[18] Executive Office of the President, C. Muñoz, M. Smith, and D. Patil. Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights, May 2016. URL https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

[19] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im)possibility of fairness. *arXiv:1609.07236 [cs, stat]*, Sept. 2016. URL http://arxiv.org/abs/1609.07236. arXiv: 1609.07236.

[20] P. Gajane and M. Pechenizkiy. On formalizing fairness in prediction with machine learning. *CoRR*, abs/1710.03184, 2017. URL http://arxiv.org/abs/1710.03184. arXiv: 1710.03184.

[21] B. Green. Fair Risk Assessments: A Precarious Approach for Criminal Justice Reform. *5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2018.

[22] R. M. Guion. Employment tests and discriminatory hiring. *Industrial Relations: A Journal of Economy and Society*, 5(2):20–37, 1966. Publisher: Wiley Online Library.

[23] M. Hardt, E. Price, and N. Srebro. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://papers.nips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html.

[24] U. Hebert-Johnson, M. Kim, O. Reingold, and G. Rothblum. Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1939–1948. PMLR, July 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

[25] B. Hutchinson and M. Mitchell. 50 Years of Test (Un)Fairness: Lessons for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 49–58, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287600. URL https://doi.org/10.1145/3287560.3287600. event-place: Atlanta, GA, USA.

[26] M.-T. Hütt and C. Schubert. Fairness von KI-Algorithmen. In K. Mainzer, editor, *Philosophisches Handbuch Künstliche Intelligenz*, pages 1–22. Springer Fachmedien Wiesbaden, Wiesbaden, 2020. ISBN 978-3-658-23715-8. doi: 10.1007/978-3-658-23715-8_39-1. URL http://link.springer.com/10.1007/978-3-658-23715-8_39-1. Series Title: Springer Reference Geisteswissenschaften.

[27] M. Kasy and R. Abebe. Fairness, Equality, and Power in Algorithmic Decision-Making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 576–586, Virtual Event Canada, Mar. 2021. ACM. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445919. URL https://dl.acm.org/doi/10.1145/3442188.3445919.

[28] M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR, July 2018. URL https://proceedings.mlr.press/v80/kearns18a.html.

[29] N. Kilbertus, M. Rojas Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding Discrimination through Causal Reasoning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f5f8590cd58a54e94377e6ae2eded4d9-Abstract.html.

[30] M. P. Kim, A. Ghorbani, and J. Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

[31] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In C. H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 43:1–43:23, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-95977-029-3. doi: 10.4230/LIPIcs.ITCS.2017.43. URL http://drops.dagstuhl.de/opus/volltexte/2017/8156. ISSN: 1868-8969.

[32] M. Kuppler, C. Kern, R. L. Bach, and F. Kreuter. Distributive Justice and Fairness Metrics in Automated Decision-making: How Much Overlap Is There? *arXiv:2105.01441 [cs, stat]*, May 2021. URL http://arxiv.org/abs/2105.01441. arXiv: 2105.01441.

[33] M. J. Kusner and J. R. Loftus. The long road to fairer algorithms. *Nature*, 578(7793):34–36, Feb. 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/d41586-020-00274-3. URL http://www.nature.com/articles/d41586-020-00274-3.

[34] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual Fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

[35] J. Larson, S. Mattu, L. Kirchner, and L. Angwin. How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*, 2016. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm#:~:text=For%20most%20of%20our%20analysis,of%20the%20COMPAS%20administration%20date.%E2%80%9D.

[36] M. Lee, L. Floridi, and J. Singh. From Fairness Metrics to Key Ethics Indicators (KEIs): A Context-Aware Approach to Algorithmic Ethics in an Unequal Society. *SSRN Electronic Journal*, Jan. 2020. doi: 10.2139/ssrn.3679975.

[37] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A Survey on Bias and Fairness in Machine Learning. *arXiv:1908.09635 [cs]*, Sept. 2019. URL http://arxiv.org/abs/1908.09635. arXiv: 1908.09635.

[38] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum. Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, 8(1):141–163, Mar. 2021. ISSN 2326-8298, 2326-831X. doi: 10.1146/annurev-statistics-042720-125902. URL http://arxiv.org/abs/1811.07867. arXiv: 1811.07867.

[39] J. Pearl, M. Glymour, and N. P. Jewell. *Causal Inference in Statistics*. John Wiley & Sons Ltd, 2016.

[40] D. Pedreshi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 560–568, New York, NY, USA, Aug. 2008. Association for Computing Machinery. ISBN 978-1-60558-193-4. doi: 10.1145/1401890.1401959. URL `https://doi.org/10.1145/1401890.1401959`.

[41] J. A. Rawls. *A Theory of Justice*. Belknap Press of Harvard Univ. Press, 2003.

[42] J. E. Roemer. *Theories of Distributive Justice*. Harvard University Press, 2018.

[43] A. Singh, D. Kempe, and T. Joachims. Fairness in ranking under uncertainty. *Advances in Neural Information Processing Systems*, 34, 2021.

[44] H. Suresh and J. V. Guttag. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9, Oct. 2021. doi: 10.1145/3465416.3483305. URL `http://arxiv.org/abs/1901.10002`. arXiv: 1901.10002 version: 5.

[45] R. L. Thorndike. Concepts of culture-fairness. *Journal of Educational Measurement*, 8(2): 63–70, 1971. Publisher: Wiley Online Library.

[46] S. Verma and J. Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, pages 1–7, Gothenburg Sweden, May 2018. ACM. ISBN 978-1-4503-5746-3. doi: 10.1145/3194770.3194776. URL `https://dl.acm.org/doi/10.1145/3194770.3194776`.

[47] R. Zwick and N. J. Dorans. Philosophical perspectives on fairness in educational assessment. In *Fairness in educational assessment and measurement*, pages 283–298. Routledge, 2016.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See paragraph "Limitations and future research" in Section 4
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See paragraph "Negative societal impacts" in Section 4
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] All results and conclusions are thoroughly explained in the main body of the paper. Since our main conclusions have a more philosophical than a mathematical character, we feel that textual explanations are most constructive here.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A] We did not run any experiments since this a purely theoretical paper
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A] We did not run any experiments since this a purely theoretical paper
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] We did not run any experiments since this a purely theoretical paper

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] We did not run any experiments since this a purely theoretical paper

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A] We are neither using existing assets nor curating/releasing new assets
   (b) Did you mention the license of the assets? [N/A] We are neither using existing assets nor curating/releasing new assets
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A] We are neither using existing assets nor curating/releasing new assets
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We are neither using existing assets nor curating/releasing new assets
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] We are neither using existing assets nor curating/releasing new assets
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did neither use crowdsourcing nor conducted research with human subjects
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We did neither use crowdsourcing nor conducted research with human subjects
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We did neither use crowdsourcing nor conducted research with human subjects

A Appendix

Requirements for $m_{\mathbf{x}} : \mathcal{X} \to \tilde{\mathcal{X}},\ \mathbf{x} \mapsto \tilde{\mathbf{x}}$: All normatively equal individuals must be assigned the same counterfactual $\tilde{\mathbf{x}}^{(i)}$. To avoid trivial mappings to, e.g., 0, validity of the counterfactuals $\tilde{\mathbf{x}}^{(i)}$ must also be required: The PA-independent contexts within the feature space $\mathcal{X}$ of the real world must be found in the fictitious world, only the contexts evaluated as PA-dependent are fictitiously removed. In other words, feature groups X_2 and X_3 in Figure 1b remain untouched.

Requirements for $m_y : \mathcal{Y} \to \mathcal{Y},\ y \mapsto \tilde{y}$: All normatively equal individuals must be assigned the same counterfactual $\tilde{y}^{(i)}$. In addition, to avoid trivial mappings to, e.g., 1, validity of counterfactuals $\tilde{y}^{(i)}$ must be required: The PA-independent relations within $(\mathcal{X}, \mathcal{Y})$ of the real world must be found in the fictitious world, only the PA-dependent relations are fictitiously removed – here in comparison to above thus including the target variables.