What is Fairness? On Protected Attributes and Fictitious Worlds

LUDWIG BOTHMANN, LMU Munich & Munich Center for Machine Learning (MCML), Germany
KRISTINA PETERS, LMU Munich, Germany
BERND BISCHL, LMU Munich & Munich Center for Machine Learning (MCML), Germany

A growing body of literature in fairness-aware machine learning (fairML) aims 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 to ensure that trained ML models achieve low scores on these metrics. However, the underlying concept of fairness, i.e., the question of what fairness is, is rarely discussed, leaving a significant gap between centuries of philosophical discussion and the 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 training and evaluation of ML models in ADM systems. We argue that fairness problems can arise even without the presence of protected attributes (PAs), and point out that fairness and predictive performance are not irreconcilable opposites, but that the latter is necessary to achieve the former. Furthermore, we argue why and how causal considerations are necessary when assessing fairness in the presence of PAs by proposing a fictitious, normatively desired (FiND) world in which PAs have no causal effects. In practice, this FiND world must be approximated by a warped world in which the causal effects of the PAs are removed from the real-world data. Finally, we achieve greater linguistic clarity in the discussion of fairML. We outline algorithms for practical applications and present illustrative experiments on COMPAS data.

Keywords: fairness-aware machine learning, causality, philosophical foundations, pre-processing

Reference Format:
Ludwig Bothmann, Kristina Peters, and Bernd Bischl. 2025. What is Fairness? On Protected Attributes and Fictitious Worlds. In *Proceedings of Fourth European Workshop on Algorithmic Fairness (EWAF'25).* Proceedings of Machine Learning Research, 28 pages.

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 easily answered and is often circumvented; 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, in doing so, often propose important solutions for specific problems. 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 broadly consistent understanding can be based on a

Authors' Contact Information: Ludwig Bothmann, LMU Munich & Munich Center for Machine Learning (MCML), Germany, ludwig.bothmann@lmu.de; Kristina Peters, LMU Munich, Germany, kristina.peters@jura.uni-muenchen.de; Bernd Bischl, LMU Munich & Munich Center for Machine Learning (MCML), Germany, bernd.bischl@stat.uni-muenchen.de.

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

rich foundation in the history of philosophy. Subsequently, we show that the conception of fairness depends on multilayered normative evaluations; any discussion of fairML relies 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. On this 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 essential concept of fairness, we turn to the questions of to what extent ML models can induce unfair treatments in automated decision-making (ADM), and of how to implement these normative stipulations both in training an ML model and in using its predictions in ADM.

1.1 Our Contributions

In this paper, we formalize a *consistent concept of fairness* derived from philosophical considerations and translate it into a formal framework for the design of ML models in ADM systems, hence bridging the gap between centuries of philosophical discussion and the recent adoption in the ML community (Section 2). 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 legislative decisions; this distinction offers *greater linguistic clarity* for the discussion of algorithmic fairness (Sections 2 and 3.1). We argue that an ML model cannot be unfair per se because a predictive model alone does not make decisions or execute material actions. Rather, *fairness problems can arise* using ML models in ADM systems if the model is not individually well-calibrated – *even if no protected attributes (PAs) are concerned*. Hence, we point out that predictive performance is paramount for fairness rather than a tradeoff to it (Section 3.2). *In the presence of PAs*, we tackle the goal of PA-neutrality (leaving PA-focus to future research, see Section 2.1 for the distinction of these two concepts): a fictitious, normatively desired (FiND) world is conceived, where the PAs have no causal effects. This FiND world must be approximated by a warped world, and ML models are to be trained and evaluated in the warped world. We emphasize that *fairness criteria must make causal considerations* (Section 3.3). In addition, our concept offers an explanation for the possibility for fair but unequal treatment of unequal individuals – also called vertical equity [11]. In the appendix, we outline first *algorithms for practical applications* and show illustrative *empirical results* on COMPAS data (Appendix C). However, our focus is on the theoretical problems and fundamental challenges surrounding fairML. While we investigate and diagnose these thoroughly, we also indicate how these problems might be overcome. We do not present final solutions to all open questions but put the necessary emphasis on clarifying the structural problems in fairML to enable future research to investigate proper solutions that could help mitigate ethical issues of ADM.

1.2 Motivating Examples

Our concept aligns well with the notions of *substantive equality* and *bias transforming fairness metrics* as described, e.g., by Wachter et al. [74] (Appendix B compares our framework with these notions). Before presenting our framework in detail, we illustrate our understanding with two examples:

Recidivism Prediction. Consider a Person of Color who is charged with a crime and whose pretrial release must be decided by a judge. As a basis for this decision, a two-year recidivism rate is predicted by an ML model. However, if the ML model is trained on historical data, it tends to perpetuate historical discrimination such as racial profiling or, in some cases, racially discriminatory practices by prosecutors or judges. This means that a high recidivism rate may reflect historical discrimination rather than the individual's actual likelihood of committing a crime. Such an approach is called *bias preserving* by Wachter et al. [74], which includes approaches that optimize for "classical fairML metrics" such as predictive parity or equalized odds: While they try to equalize error rates in different subgroups as defined by PAs, they still aim to predict well in the – potentially biased – real world. In contrast, we take a *bias transforming* approach by imagining a FiND world in which certain causal effects on recidivism that concern PAs are eliminated, using the prediction derived in that FiND world as the basis for our decision. This would underestimate the true likelihood of being convicted of a crime in the real world – assuming that the real world is still characterized by racial discrimination – but would more fairly reflect the individual's likelihood of being convicted of a crime in an unbiased world. We will use this example throughout the paper.

College Admission. Consider a Black student who, because of structural inequalities, has attended underfunded schools with fewer educational opportunities, lived in a less safe neighborhood, and so on. Suppose that the college admissions decision is based on a prediction of her probability of educational success based on, among other things, her SAT score. A bias-preserving approach would be based on training an ML model on historical data, including biases that may have led to lower success rates for individuals from certain PA groups. This leads to several problems. First, the model might "correctly" (for the biased real world) predict further structural inequalities in college education, thus perpetuating these inequalities to the disadvantage of the individual. Furthermore, it might not be able to adequately predict the individual success rate, as it is likely that the Black student has greater potential than reflected by the SAT score because she achieved her score *despite all the obstacles* placed in her way. In our terms, this would mean that she actually has a higher task-specific merit (see definition below). And lastly, even if the person in question actually has a lower chance of success at college due to their poorer educational opportunities, the college could decide that the person should not have to take responsibility for this and still give the person the chance to make up for these deficits at university in order to break the circle of inequality. Technically, we aim to overcome these problems by removing the direct and indirect effects of racialisation on her SAT scores. Within such an approach, it is possible to calculate her actual task-specific merit by removing only some path-specific effects while preserving others. As a result, she would have a higher SAT score in the FiND world, potentially leading to a higher predicted probability of success, justifying her admission to college.

1.3 Related Work

What exactly is understood by fairness is not disclosed by central laws [20, 24] or statements from politics [25]. 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, 6, 9, 15, 19, 55]. However, most of their arguments are presuppositional and start a step further than we do here. When attempted, explanations of what fairness really means usually settle on vague definitions [8, 22, 31, 37, 45] – notable exceptions being Loi and Heitz [53] and Kong [46]. A fundamental critique of statistical criteria as fairness measures such as "false positive rate equality does not track anything about fairness, and thus sets an incoherent standard for evaluating the fairness of algorithms" by Long [54] is also formulated by [26, 35, 53], pointing out a special role of calibration. Reflections on fairness as not (only) a technical challenge can be found in, e.g., [32, 52, 68, 75].

There is a growing awareness in ML literature that basic assumptions that are not made explicit do indeed matter [28, 51, 57, 67], but the reappraisal of these basic assumptions is still in its infancy. Adding to the many concepts of *group fairness* [see, e.g., 72, for an overview], the notion of *individual fairness*, highlighted in particular by Dwork et al. [23], has resonated considerably [7, 19, 28, 48]. This concept requires that similar individuals should be treated similarly and reflects a demand already comparably formulated by Guion [33] more than 50 years ago [see also 36, 71, regarding the debate on test (un)fairness in the mid-20th century]. For ranking tasks, Singh et al. [69] assume that "[u]nfairness occurs when an agent with higher merit obtains a worse outcome than an agent with lower merit". *Individual counterfactual fairness* [48] goes in a different direction, stating that 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 protected group. This is arguably the work that is most related to our concept, and we differentiate our proposal from it in Appendix B. As described later, these definitions – which take up causal concepts [see also 16–18, 30, 40, 43, 49] – 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 [28, 48, 49, 57].

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 proceed a step further and usually focus on which material criteria should be taken into account in the context of a just or fair distribution [13]. These considerations already build on the concept that we will present below [65, 66]. 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 [10, 29, 40, 47].

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, 14, 22, 45] 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, e.g., 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: Task-specific (In-)Equality

In these considerations – sometimes referred to as the difference between "equality" and "equity" – lies the basic problem of arguments about fairness. These arguments always depend on the reference point of the evaluation: (1) Is it solely the distribution of the asset at hand? Or (2) should the point of reference also be the person concerned? In the second case, what is fair is 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 considered by [51, 52, 76] 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 are to be treated 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 exactly 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 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 (e.g., the price for a bottle of water in a specific supermarket is x – no matter who buys it). Similarly, for the assessment of a penalty, it is irrelevant whether a rich person kills a poor person or vice versa. Because the decision on the treatment here can be made by means of a simple calculation (e.g., the price for two bottles of water is $2x$ – no matter who buys it), Aristotle speaks in this respect of equality as *arithmetic or continuous proportionality*.

In other situations, equality is said to depend on some 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. We will also use the – more contemporary – term *task-specific merit* to refer to this concept of *task-specific equality*. For example, take the tax rate: usually, those who have a higher income 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 task-specific merit negotiated in the political dispute: the ratio of the task-specific merit of person i (e.g., the income) to the assets distributed to them (e.g., the tax rate) must correspond to the ratio of the task-specific merit 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 features of the people involved (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, e.g., when private companies discriminate racially. 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 and treatment of two people must not be based on these PAs (*PA-neutrality*). Since the attributes cannot act as causal factors – if they are not to be considered for comparisons – this decision is accompanied by the consideration of ignoring the consequences of these attributes as well, see e.g., the example of COMPAS [2] which will be used in later sections of this work: 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. Another consideration might be that detection rates and conviction rates of criminal offenses might vary across ethnicities (e.g., due to racial profiling or biased law enforcement agencies), leading to higher records of recidivism rates. Society may take responsibility for these consequences, wanting to keep them out of the decision-making process. However, because processes occurring in the life of an individual are usually not monocausal – 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. In Section 3 onwards, we will focus on the case of PA-neutrality, elaborate on how this fictitious world can be formalized and which consequences this has for the design of ML models that shall be used as components of ADM systems.

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 or a focus on their specific characteristics [also called *PA-focus* or *affirmative action*, see, e.g., 21, 27, 38, 42]. Examples could be (1) hiring policies that aim at achieving gender diversity or educational programs that actively target historically disadvantaged populations or (2) decisions on medical treatments that take into account, e.g., the physiological differences between people of different gender, also referred to as "gender medicine" [4]. The perspective depends on a normative decision as to whether the protection of the feature bearers in the respective task is to be ensured "only" by means of exclusion of these features and their effects (PA-neutrality), or whether reality is to be actively reshaped according to certain objectives (PA-focus). In Aristotle's words, in the first case, the protected attribute must not be used to decide on task-specific merit, while in the second case, the same attribute is used to determine higher merit. Thus, even the complex socio-political realities of the modern world fit into Aristotle's concept, while at the same time, it is clear that today complex normative decision-making processes are involved. Even in situations where we conceive PAs, there is thus no generally fixed "equal treatment" or "fairness", but every concrete task demands answers to the above normative questions. As mentioned at the beginning, the specific, technical framework proposed in the remainder of this paper focuses on PA-neutrality, while a corresponding framework for PA-focus – even if fitting into the same basic philosophical concept – is left to future research.

2.2 Decision Rule: Equal Treatment

Once the normative basis for a decision has been established, the second step is to develop a decision rule that determines the extent to which the equality or inequality thus conceived is to be taken into account. In the simpler case of arithmetic proportionality, a fair treatment does not depend on the task-specific merit:

DEFINITION 2.1 (ARITHMETICALLY FAIR TREATMENT). *A treatment $t^{(i)}$ of an individual i is called arithmetically fair if and only if it is the same as for any other individual, i.e., does not depend on an individual's task-specific merit $m^{(i)}$*

$$t^{(i)} = k \quad \forall i \in \{1, \dots, n\}, k \in \mathbb{R},$$

where $n \in \mathbb{N}$ is the number of individuals in the entire population.

In the case of geometric proportionality, it must hold that the ratio of treatment t to task-specific merit m is the same for any comparison of two individuals i and j, i.e.,

$$\frac{t^{(i)}}{m^{(i)}} = \frac{t^{(j)}}{m^{(j)}} = k \quad \forall i, j \in \{1, \ldots, n\}, k \in \mathbb{R}.$$

In other words, we can define a treatment function $s : M \to T,\ m \mapsto t$, where typically $M \subseteq \mathbb{R}$ and $T \subseteq \mathbb{R}$:

Definition 2.2 (Geometrically fair treatment). *A treatment $t^{(i)}$ of an individual i is called geometrically fair if and only if it is a linear function of the individual's task-specific merit $m^{(i)}$, i.e.,*

$$t^{(i)} = s(m^{(i)}) = k \cdot m^{(i)} \quad \forall i \in \{1, \ldots, n\}, k \in \mathbb{R}.$$

The task-specific merit $m^{(i)}$ may be directly identified with some observable feature $v^{(i)}$ (such as the income in the tax rate example), but it may also be defined as a latent and non-observable feature $z^{(i)}$. Finally, it may be a combination of observable and non-observable features, i.e., $m^{(i)} = f(v^{(i)}, z^{(i)})$, where $f(\cdot)$ is another normative function. If PAs are present, the effect of these features and hence the task-specific merit of an individual may be modified by the PAs, referring to their counterparts in a fictitious, normatively desired world, see Section 3 for details. In the example of pretrial decisions – such as in the COMPAS example introduced shortly in this paper – $v^{(i)}$ and $z^{(i)}$ may be the type of crime and probability of recidivism, respectively. The duty of ML in the ADM process will be to provide an estimate of the non-observable feature(s) $z^{(i)}$, where their meaning is possibly modified by the PAs. These estimates serve as the basis for assessing the task-specific equality of individuals, i.e., for deriving their task-specific merit $m^{(i)}$, which in turn is the basis for the decision of their treatment $t^{(i)}$ via $s(m^{(i)})$, see below.

It may be decided to modify the treatment function $s(\cdot)$ in a more flexible way, e.g., if a tax rate $s(\cdot)$ is raised with higher income $m^{(i)}$, 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 2.3 (Modified fair treatment). *A treatment $t^{(i)}$ of an individual i is called modified fair if and only if it is determined by a monotonic[1] function of the individual's task-specific merit $m^{(i)}$, i.e.,*

$$t^{(i)} = s(m^{(i)}) \quad \forall i \in \{1, \ldots, n\}.$$

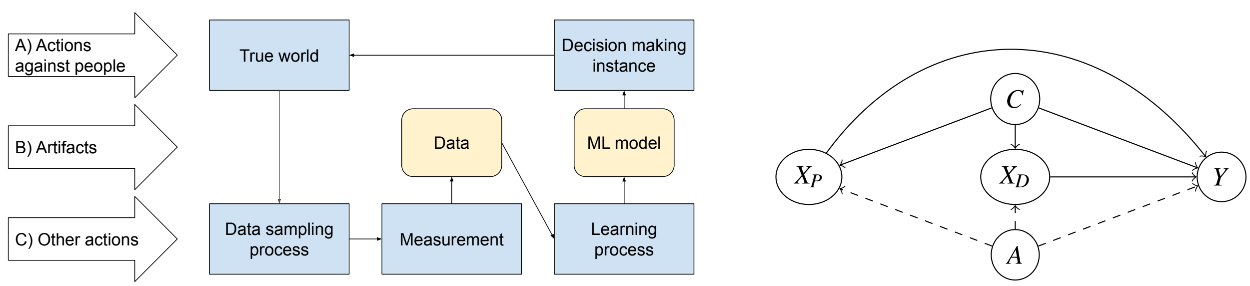
This approach still allows for (i) equal treatment of equals and (ii) unequal treatment of unequals in a normatively defined way that is then considered to be fair. To enhance readability, the term "fair treatment" refers to this notion in the remainder of the paper, if not stated otherwise. Even if it would be mathematically straightforward to allow for $s(\cdot)$ to be more flexible, a non-monotonic $s(\cdot)$ would not be in line with the above philosophical concept. One will demand strict monotonicity if unequals are always to be treated unequally. Geometrically and arithmetically fair treatments result from special choices of $s(\cdot)$. Thus, again, valuation decisions arise that add a normative dimension, making it three normative questions in this scenario of Def. 2.3: (1) How is the task-specific merit $m^{(i)}$, i.e., the measure of task-specific equality of individuals, defined? (2) Which attributes are defined as PAs (if there are any)? (3) Shall the treatment function $s(m^{(i)})$ be modified, and if so, how?

[1]Note that we do not require the function $s(\cdot)$ to be *strictly* monotonic in general; the function is allowed to have plateaus, such as in the tax example. The concrete form of the function and the concrete value of k are normative choices.

3 Role of ML in the ADM Process

Now that we have defined the general concept of fairness, we can turn to answer the questions of to what extent an ML model used in the context of an ADM process can contribute to unfair treatment and of what this entails for the design of ML models. First, in Section 3.1, we review the steps of an ADM process with respect to the applicability of the notion of fairness and illustrate it using the COMPAS example – focusing on the goal of PA-neutrality rather than PA-focus (see again Section 2.1). In Sections 3.2 and 3.3 we discuss the contributions of the ML model and propose concrete solutions for the situations without and with PAs, respectively. Appendix B compares our framework with two seemingly related concepts. We illustrate its application in Appendix C.

3.1 Fairness in the ADM Process



(b)

(a)

Fig. 1. (a) Simplified ADM process, following Suresh and Guttag [70]. (b) Illustrative DAGs for the COMPAS example. In the FiND world, only solid arrows exist, and in the real world, all arrows exist. Features are *ethnicity* (protected attribute A with classes *Non-White* and *White*), *age, gender* (confounder C), *number of priors* (X_P), and *charge degree* (X_D). The target is *recidivism in two years* (Y).

The ADM process is explained in detail in Suresh and Guttag [70], while we consider a reduced version (see Figure 1a). We divide the process into three categories: (A) actions against people, (B) artifacts, and (C) other actions. As shown in Section 2, the notion of fairness is applicable only to category (A) and not directly applicable to the ML model. However, the results of categories (B) and (C) affect the estimated "task-specific merit", i.e., the measure of task-specific equality of individuals that is the basis of decisions for actions against people. As mentioned above, the task-specific merit $m^{(i)}$ may consist of observable features $v^{(i)}$ (like the type of crime) and non-observable features $z^{(i)}$ (like the recidivism probability), i.e., $m^{(i)} = f(v^{(i)}, z^{(i)})$. For ease of presentation, we will ignore any observable features in the following and set $m^{(i)} = z^{(i)}$. For a classification task, $z^{(i)}$ is the individual probability of success $\mathbb{P}(y = 1|i) = \pi^{(i)}$ of an event $y \in \{0, 1\}$; for a regression task, 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 – or survival

– scenario is straightforward, thereby remedying a critique towards fairML's focus on classification formulated by [36]. Before we shed more light on the role of ML in this process (Sections 3.2 and 3.3), we highlight crucial differences between ML model and actions based on it as well as 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 ($y \in \{0,1\}$) [50] 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 this probability, based on a feature vector $\mathbf{x}^{(i)}$. The goal of the ML model [see, e.g., 62] is to assign different recidivism probabilities as accurately as possible to different individuals, based on the observed features:

DEFINITION 3.1 (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 $\hat{\pi}^{(i)} \neq \hat{\pi}^{(j)}$.*[2]

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 this evaluation. From this, we derive similar to [33]:

DEFINITION 3.2 (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 3.3. *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 (equal task-specific merit $\pi^{(i)} = \pi^{(j)}$) would be treated unequally (unequal treatment $t^{(i)} \neq t^{(j)}$).*

If, on the other hand, X is causal for a difference in π, i.e., $\pi^{(i)} \neq \pi^{(j)}$, then a differing decision basis due to X, i.e., $\hat{\pi}^{(i)} \neq \hat{\pi}^{(j)}$, 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 Lee et al. [51], as the assessment of task-specific equality changes:

DEFINITION 3.4 (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.,

$$t^{(i)}(= s(\hat{\pi}^{(i)})) \neq t^{(j)}(= s(\hat{\pi}^{(j)}))$$

[2]Note that from here on, individuals i and j are not limited to be members of a finite training set but are assumed to be members of a potentially infinite population, i.e., their feature vectors can be any point in the feature space.

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 task-specific equality.*

EXAMPLE 3.5. *Suppose that the probability of recidivism* π *depends causally on ethnicity A, but the judicial decision should not – for normative reasons – depend on differences in* π *due to A, 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 (e.g., via racial profiling and higher re-arrest rates), but to take it from them and bear it as a whole society.*

So, the measure of task-specific equality is no longer the true recidivism probability π in the real world, but the true recidivism probability ψ in a fictitious, normatively desired (FiND) world:

DEFINITION 3.6 (FIND WORLD). *A FiND world is a fictitious, normatively desired world, where the PAs have no causal effects on the target variable, neither directly nor indirectly, i.e.,* $\psi^{(i)} = \psi^{(j)}$ *for the pair of Definition 3.4.*

This mirrors the demand of several laws to not differentiate between individuals based on their PA (e.g., gender, ethnicity), which means that individuals are to be considered as being equal if they only differ in their PA. This also means that subsequent effects of their PA on other features must be eliminated, therefore removing all direct and indirect effects of the PA on the target to reach the FiND world (i.e., remove dashed arrows in the Directed Acyclic Graph (DAG) shown in Figure 1b).

In debates such as the discussion surrounding COMPAS, a fairly broad consensus has been reached that combating discrimination means not only tackling direct discrimination but also indirect discrimination through proxy variables. However, several papers in fairML [e.g., 16, 58–60] argue for removing only some path-specific effects where the path itself is deemed unfair. While in some cases, indirect discrimination might be allowed, Wachter [73] argues that "[i]ndirect discrimination is only lawful if a legitimate aim is pursued and the measure is necessary and proportionate" and describes examples of European case law where this issue has been debated, concluding that the hurdles for justifying indirect discrimination are rather high. As there may therefore be cases where some paths starting from the PA are not considered unfair, we extend our methodology to such admissible path-specific effects: Let X_{ps} be a feature where the path from the PAs through X_{ps} to the target is normatively defined as not being unfair. Then the effects of the PAs on this X_{ps} are not removed in the definition of the FiND world (and hence the corresponding arrows in Figure 1b remain solid):

DEFINITION 3.7 (FIND WORLD – ADMISSIBLE PATH-SPECIFIC EFFECTS). *A FiND world with admissible path-specific effects is a fictitious, normatively desired world, where PAs have no causal effects on the target variable, neither directly nor indirectly – with the exception of path-specific effects via some features* X_{ps}.

The question of which paths should be removed, and thus the concrete definition of the FiND world, is a purely normative and potentially highly controversial issue. Our framework covers a wide spectrum of potential answers to these normative questions.

It can be argued that there are several FiND worlds that fit either of the two definitions above. The concrete definition of the FiND world depends on the specific actions that are carried out to remove the dashed arrows in the DAG. This relates to the definition of what exactly the baseline "without PA effects" is. As an example, let us assume that *ethnicity* is a binary PA with values *Non-White* and *White*. For mapping the number of priors (X_P) into a world without effects of *ethnicity*, should we remove the effect of belonging to the *White* subgroup (as compared to the *Non-White* subgroup) on X_P? This would make the *Non-White* subgroup the baseline. Or should we remove

Table 1. Coarse information due to finite feature space $\mathscr{X}$ and estimation via ML introduce errors. Point of reference changes with presence of PAs (columns, see also Section 3.3).

	Real world	Warped world	FiND world
True target probability	$\pi^{(i)}$	$\varphi^{(i)}$	$\psi^{(i)}$
Fair treatment	$s(\pi^{(i)})$	$s(\varphi^{(i)})$	$s(\psi^{(i)})$
Treatment w/ coarse inf.	$s(\pi(\mathbf{x}^{(i)}))$	$s(\varphi(\tilde{\mathbf{x}}^{(i)}))$	$s(\psi(\mathbf{x}_F^{(i)}))$
Treatment w/ ML model	$s(\hat{\pi}(\mathbf{x}^{(i)}))$	$s(\hat{\varphi}(\tilde{\mathbf{x}}^{(i)}))$	$s(\hat{\psi}(\mathbf{x}_F^{(i)}))$

the effect of belonging to the *Non-White* subgroup (as compared to the *White* subgroup) on X_P? This would make the *White* subgroup the baseline. Alternatively, the baseline could be something in between, like the marginal distribution of X_P. The choice of baseline will change the concrete values predicted by an ML model for that world. However, all such baselines are equally valid when aiming at fair, i.e., PA-neutral, treatment. In the remainder, we will assume that a decision on the baseline has been made and work with the such defined FiND world.

Once we have moved to this world after defining the PAs, we use $\psi^{(i)}$ (instead of $\pi^{(i)}$) as a measure of task-specific equality, i.e., introducing PAs has changed the definition of task-specific merit in a specific use-case normatively; hence, a fair treatment would result by $t^{(i)} = s(\psi^{(i)})$. The role of ML is to estimate $\pi^{(i)}$ or $\psi^{(i)}$ accurately, respectively.

3.2 Contribution of ML – without PAs

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 turn to the situation where PAs are present.

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

In Section 2, we observed that the treatment $t^{(i)}$ of an individual i based on task-specific merit $m^{(i)}$ is fair if and only if it follows the normative decision rule $s(\cdot)$, i.e., if and only if $t^{(i)} = s(m^{(i)})$. In Section 3.1, we observed that in the context of an ADM process, the task-specific merit $m^{(i)}$ often corresponds to the individual probability $\pi^{(i)} = \mathbb{P}(y = 1|i)$. 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 (see also Table 1):

Coarsening of Information. The first step is to coarsen the information by basing the treatment not on the 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 $\pi : \mathscr{X} \to [0,1]$ 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. With the exception of 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

$$\pi^{(i)} \neq \pi(\mathbf{x}^{(i)}) \Rightarrow s(\pi^{(i)}) \neq s(\pi(\mathbf{x}^{(i)})) = t^{(i)} \; \forall i \in \{1, \dots, n\}$$

(if $s(\cdot)$ is injective), so (almost) all individuals are treated unfairly – even if $\pi(\cdot)$ is known.

This step is only appreciated in a few works, such as [28, 45, 57, 69]. 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 $\pi(\cdot)$, even before introducing PAs, and even before estimating an ML model.

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 $\mathscr{D} = \left(\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(n)}, y^{(n)}\right)\right)$ available. This error converges to 0 for $n \to \infty$. Should the true $\pi(\cdot)$ not be part of the hypothesis space $\mathscr{H}$, 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 $\mathbf{x} \in \mathscr{X}$ 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 with respect directly to $\pi^{(i)}$ seems more purposeful: We recall that a treatment $t^{(i)}$ is fair if and only if $t^{(i)} = s(\pi^{(i)})$, so here, if and only if $s(\hat{\pi}(\mathbf{x}^{(i)})) = s(\pi^{(i)})$. Thus, a sufficient condition for fair treatment is $\hat{\pi}(\mathbf{x}^{(i)}) = \pi^{(i)}$, which can be seen as an individual version of well-calibration [45, 72]:

DEFINITION 3.8 (INDIVIDUALLY WELL-CALIBRATED MODEL). *An ML model $\hat{\pi}(\cdot)$ is called individually well-calibrated for individual i if $\hat{\pi}(\mathbf{x}^{(i)}) = \pi^{(i)}$; the model is called individually well-calibrated if $\hat{\pi}(\mathbf{x}^{(i)}) = \pi^{(i)} \quad \forall i \in \{1, \ldots, n\}$.*

Note that a direct generalization to regression, survival tasks, etc., is possible by replacing the probabilities with expected values. For a strictly monotonic 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 in, e.g., [28, 34, 41, 44], and check well-calibration in all these subgroups. Since exact equality $\hat{\pi}(\mathbf{x}^{(i)}) = \pi^{(i)}$ is not achievable in practice, a pragmatic tolerance range should be conceived. The above equalities in the definitions of individual well-calibration are then relaxed to $\hat{\pi}(\mathbf{x}^{(i)}) \in (\pi^{(i)} - \varepsilon, \pi^{(i)} + \varepsilon)$ with a tolerated deviation of $\varepsilon \in \mathbb{R}^+$.

Here, also the group fairness metrics such as *equalized odds*, *predictive parity*, etc. [see, e.g., 72] have their value: We can think of them as *fairness-related performance metrics* that allow us to obtain more nuanced information on the model performance in subgroups as this goes beyond well-calibration. Also here, the subgroups are not defined by PAs – since there are no PAs in this scenario – but all computationally identifiable subgroups should be considered for a holistic view of the model's performance.

3.3 Contribution of ML – with PAs

We have observed that 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. First, we elaborate on the FiND world and its approximation via warping, and then we turn to the question of evaluating the model regarding fairness.

3.3.1 FiND world and approximation via warping

In Section 3.1, it was shown that in the presence of a PA, the basis for decisions must be PA-neutral to be able to achieve fair treatment, i.e., the normative decision can be made to use the PA-neutral true probability $\psi^{(i)}$ instead of $\pi^{(i)}$ as a measure of task-specific equality. This $\psi^{(i)}$ describes for an individual i the probability $\mathbb{P}(y_F = 1|i)$ for an event $y_F \in \{0,1\}$ in a FiND world in which the PA has no causal effect on this event, neither directly nor indirectly.[3] To differentiate the target in the FiND world from the originally observed event y in the real world, we denote it as y_F. As above, treatment $t^{(i)}$ of an individual i is then said to be fair if and only if it follows the normatively specified decision rule $s(\psi^{(i)})$, with changed decision basis $\psi^{(i)}$. Thus, in the example, we consider the individual recidivism probability $\psi^{(i)}$, *in a fictitious world where the PA has no causal effect on recidivism*.

Since the PA is not supposed to have a causal effect on y_F, the normative decision can be made to exclude certain indirect effects of the PA on y_F via other features. Therefore, *the feature vector of individual i is also to be considered PA-neutral*, i.e., $\mathbf{x}_F^{(i)}$. This does not mean that the PA is removed [hence, goes beyond "fairness through unawareness", see 29]. Rather, only those influences that (possibly via detours) affect y_F are removed (see dashed arrows in Figure 1b). In the COMPAS example, individuals might have not only a different recidivism probability when the effect 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 effect on these features.

It is also not certain that the relation of event and features is identical in the real and the FiND world after potentially causal effects are removed, so the function $\psi(\cdot)$ 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, with a different data-generating process (see also Table 1, last column).

Since we never have access to FiND-world data in an empirical use case, we must approximate this data. We dub this approximation *warping*, which results in data from the *warped world*. Targets and features in this world are denoted $\tilde{y}^{(i)}$ and $\tilde{\mathbf{x}}^{(i)}$, respectively, and the true target probability in the warped world is $\varphi^{(i)} = \mathbb{P}(\tilde{y} = 1|i)$. In Appendix C, we discuss and experimentally compare two concrete warping methods; for the remainder of this section, we simply assume that such a warping exists.

3.3.2 Evaluation regarding fairness

We assume that – based on warped-world data $\tilde{\mathscr{D}} = \left(\left(\tilde{\mathbf{x}}^{(1)}, \tilde{y}^{(1)}\right), \ldots, \left(\tilde{\mathbf{x}}^{(n)}, \tilde{y}^{(n)}\right)\right)$ – we have trained a model $\hat{\varphi}(\cdot)$. For an individual i, we obtain an estimate $\hat{\varphi}(\tilde{\mathbf{x}}^{(i)})$, i.e., an approximation of its warped-world probability $\varphi^{(i)}$ by warped features $\tilde{\mathbf{x}}^{(i)}$. Naturally, we introduce imprecision analogously as described above by coarsening information and estimation by ML, but this time in the warped world (see also Table 1, second column). Per definition, the

[3] While in the following we assume that all PA effects are removed for the definition of the FiND world, the normative decision can be made not to remove some admissible path-specific effects, see Definition 3.7. The following methodology also applies in this case, we just need to define some FiND world normatively.

treatment $t^{(i)} = s(\hat{\varphi}(\tilde{\mathbf{x}}^{(i)}))$ is fair if and only if $s(\hat{\varphi}(\tilde{\mathbf{x}}^{(i)})) = s(\psi^{(i)})$. Thus, a sufficient condition for fair treatment includes the FiND world probability $\psi^{(i)}$ (proved in Appendix A) :

PROPOSITION 3.9. *A treatment $t^{(i)} = s(\hat{\varphi}(\tilde{\mathbf{x}}^{(i)}))$ of individual i is fair if $\hat{\varphi}(\tilde{\mathbf{x}}^{(i)}) = \psi^{(i)}$. For a strictly monotonic $s(\cdot)$, this condition is also necessary for fair treatment.*

We can only test this condition with access to $\psi^{(i)}$, which we do not have in practical use cases. However, we can divide the problem into two independent tasks:

(1) – Warping: We have to find a warping such that the warped world approximates the FiND world as well as possible, i.e., that the structural causal models [SCM, see 61] in both worlds are similar and that the corresponding joint probability distribution $\tilde{\mathbb{P}}$ in the warped world approximates the joint probability distribution $\mathbb{P}_F$ in the FiND world. This means that warped-world data $\tilde{\mathscr{D}}$ can approximately be seen as a sample from the FiND world data-generating process $\mathbb{P}_F$ and that the individual observations approximate their FiND world counterparts, i.e., $\tilde{\mathbf{x}}^{(i)} \approx \mathbf{x}_F^{(i)}$ and $\tilde{y}^{(i)} \approx y_F^{(i)}$. Consequently, the warped world probabilities should also approximate their FiND world counterparts, i.e., $\varphi^{(i)} \approx \psi^{(i)}$.

(2) – ML model: With data from that warped world, we can train a model that optimizes predictive performance – as usually done in ML and without the need to account for PAs in any special sense (since this was already accounted for in the warping) and also without the need for fairness-specific evaluation metrics. Here we aim at a version of individual well-calibration, analogously as described in Section 3.2.2, but now concerning the warped world probabilities $\varphi^{(i)}$. We could think of this as being in the situation without PAs, and now we only have to make sure that the ML model predicts as accurately as possible. Hence, the evaluation of task (2) is straightforward and should be done as described in Section 3.2.2, i.e., considering all computationally identifiable subgroups and possibly making use of fairness-related performance metrics.

The better the warping (1) and the better the performance of the model (2), the closer we get to fair treatments:

$$t^{(i)} = s(\hat{\varphi}(\tilde{\mathbf{x}}^{(i)})) \overset{(1)}{\approx} s(\hat{\varphi}(\mathbf{x}_F^{(i)})) \overset{(2)}{\approx} s(\varphi^{(i)}) \overset{(1)}{\approx} s(\psi^{(i)})$$

This means the evaluation of the warping (1) is disentangled from the evaluation of the prediction model (2) in the warped world. However, concerning task (1), we still face the complication that we do not know the true FiND world distribution $\mathbb{P}_F$ nor have access to FiND world data to compare warped world data against. Evaluations should hence focus on the validity of the warping method itself, which includes (at least) two parts: (i) the causal effects on the paths descending from the PAs have to be estimated correctly, (ii) the warping method has to ensure that it only corrects for these causal effects of the PAs and that other influences on individual observations remain untouched. Another important point to keep in mind is that for now we assumed that the perceived DAG in the real world is correct. However, in practice we will face the challenge of deciding for a DAG based on expert knowledge and analytical tools such as causal discovery. This step brings along a certain amount of uncertainty that should be accounted for. We leave it to future work to develop concrete measures for the validity of a warping method and for incorporating the uncertainty regarding the true real-world DAG.

In the Appendix, Section B compares our framework to related concepts, Section C shows illustrative experiments, and Section D describes a workflow for an applied ADM use case.

4 Conclusion and Discussion

Despite a rapidly growing body of literature in fairML, a unified theoretical foundation was still lacking. By drawing on basic philosophical ideas, such a foundation to support the fairness debate was derived. This paper has identified the fundamental axioms that underpin the fairness debate to date, albeit without being made explicit, and has worked out their relationship to each other. It was necessary to separate the basic structure of fairness from normative specifications for specific scenarios – i.e., from material aspects. Based on this, the normative stipulations could be precisely identified, which must take place outside ML. In the case of PA-neutrality, three normative questions arise for the task at hand:

(1) How is the task-specific merit $m^{(i)}$, i.e., the measure of task-specific equality of individuals, defined (e.g., by identifying it with an (un-)observable feature or with a function of several such features)?
(2) Which attributes are defined as PAs (if there are any)?
(3) Shall the treatment function $s(m^{(i)})$ be modified, and if so, how?

ML models provide estimates of the task-specific merit 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; rather, predictive performance is essential for fairness. Since individual well-calibration cannot be checked empirically, approaches that consider all computationally identifiable subgroups should be pursued. Once PAs are defined we move to the FiND world – where causal effects of the PAs have been removed – and the task of ML models is to predict the such modified counterparts of the real-world quantities in order to estimate the task-specific merit of individuals.

Limitations and Future Research. While we present first algorithms and a workflow for applying our framework in practice in the Appendix, working out detailed warping methods and concrete techniques to measure the validity of the warped data is left to future research. This includes handling the inherent errors occurring by estimating the FiND world, such as the uncertainty regarding the true real-world DAG.

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., Barabas et al. [5] – are out of scope. Considering PAs, we focus on how to reflect *PA-neutrality* in the ADM process. For future studies, we posit that a worthwhile direction is to tackle the incorporation of *PA-focus* as well (which, again, fits into the same basic philosophical concept of normative, task-specific equality but demands different technical solutions).

As mentioned above, the focus of this work is on the theoretical problem of fairML, pointing out weaknesses in current proposals of fairML and showing a direction to overcome these. We did not want to hide these fundamental challenges by presenting a fully applied use case, hence pretending that we would have the perfect solution for all challenges. Therefore, we limited ourselves to some illustrative examples that we considered helpful to understand our framework, and only present these in the Appendix to further emphasize our focus on the theoretical discussion. The investigation of proper solutions for these challenges is important for future research in this field, and we hope that this work inspires other researchers to pursue this path, eventually leading to improved methods to deal with fairness issues in ML and ADM.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *Proceedings of the 35th International Conference on Machine Learning*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 60–69. https://proceedings.mlr.press/v80/agarwal18a.html

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals, and It's Biased Against Blacks. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] Aristotle. 2009. *The Nicomachean ethics (book V)*. Oxford University Press. https://doi.org/10.1093/actrade/9780199213610.book.1

[4] Giovannella Baggio, Alberto Corsini, Annarosa Floreani, Sandro Giannini, and Vittorina Zagonel. 2013. Gender Medicine: A Task for the Third Millennium. *Clinical Chemistry and Laboratory Medicine (CCLM)* 51, 4 (2013), 713–727. https://doi.org/10.1515/cclm-2012-0849

[5] Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. 2018. Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 62–76. https://proceedings.mlr.press/v81/barabas18a.html

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. MIT Press. http://www.fairmlbook.org

[7] Yahav Bechavod, Christopher Jung, and Steven Z. Wu. 2020. Metric-Free Individual Fairness in Online Learning. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 11214–11225. https://proceedings.neurips.cc/paper/2020/file/80b618ebcac7aa97a6dac2ba65cb7e36-Paper.pdf

[8] Susanne Beck, Armin Grunwald, Kai Jacob, and Tobias Matzner. 2019. *Künstliche Intelligenz und Diskriminierung*. Lernende Systeme, München. https://www.plattform-lernende-systeme.de/publikationen-details/kuenstliche-intelligenz-und-diskriminierung-herausforderungen-und-loesungsansaetze.html

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

[10] Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 149–159. https://proceedings.mlr.press/v81/binns18a.html

[11] Emily Black, Hadi Elzayn, Alexandra Chouldechova, Jacob Goldin, and Daniel Ho. 2022. Algorithmic Fairness and Vertical Equity: Income Fairness with IRS Tax Audit Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1479–1503. https://doi.org/10.1145/3531146.3533204

[12] Ludwig Bothmann, Susanne Dandl, and Michael Schomaker. 2023. Causal Fair Machine Learning via Rank-Preserving Interventional Distributions. In *Proceedings of the 1st Workshop on Fairness and Bias in AI co-located with 26th European Conference on Artificial Intelligence (ECAI 2023)*. CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3523/

[13] Caterina Calsamiglia. 2005. *Decentralizing Equality of Opportunity and Issues Concerning the Equality of Educational Opportunity*. Yale University. https://economics.mit.edu/sites/default/files/inline-files/Calsamiglia.pdf

[14] Cambridge Dictionary. 2022. *Fairness*. Cambridge University Press & Assessment. https://dictionary.cambridge.org/de/worterbuch/englisch/fairness

[15] Simon Caton and Christian Haas. 2024. Fairness in Machine Learning: A Survey. *ACM Computing Surveys*, 3616865. , 3616865 pages. https://doi.org/10.1145/3616865

[16] Silvia Chiappa. 2019. Path-Specific Counterfactual Fairness. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (2019), 7801–7808. https://doi.org/10.1609/aaai.v33i01.33017801

[17] Silvia Chiappa and William S. Isaac. 2019. A Causal Bayesian Networks Viewpoint on Fairness. In *Privacy and Identity Management. Fairness, Accountability, and Transparency in the Age of Big Data*. Vol. 547. Springer International Publishing, Cham, 3–20. https://doi.org/10.1007/978-3-030-16744-8_1 Series Title: IFIP Advances in Information and Communication Technology.

[18] Yoichi Chikahara, Shinsaku Sakaue, Akinori Fujino, and Hisashi Kashima. 2021. Learning Individually Fair Classifier with Path-Specific Causal-Effect Constraint. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. PMLR, 145–153. https://proceedings.mlr.press/v130/chikahara21a.html

[19] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* 63, 5 (2020), 82–89. https://doi.org/10.1145/3376898

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

[21] Stephen Coate and Glenn Loury. 1993. Will Affirmative-Action Policies Eliminate Negative Stereotypes? *American Economic Review* 83, 5 (1993), 1220–40. https://EconPapers.repec.org/RePEc:aea:aecrev:v:83:y:1993:i:5:p:1220-40

[22] Jim Dator. 2017. Chapter 3. What Is Fairness? In *Fairness, Globalization, and Public Institutions*, Jim Dator, Richard C. Pratt, and Yongseok Seo (Eds.). University of Hawaii Press, Honolulu, 19–34. https://www.degruyter.com/document/doi/10.1515/9780824841966-004/html

[23] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[24] EU. 2016. EUR-Lex - 32016R0679 - EN - EUR-Lex. 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).

[25] Executive Office of the President, Cecilia Muñoz, Megan Smith, and DJ Patil. 2016. Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf

[26] Will Fleisher. 2023. Algorithmic Fairness Criteria as Evidence. *SSRN*. https://doi.org/10.2139/ssrn.3974963

[27] Dean P. Foster and Rakesh V. Vohra. 1992. An Economic Argument for Affirmative Action. *Rationality and Society* 4, 2 (1992), 176–188. https://EconPapers.repec.org/RePEc:sae:ratsoc:v:4:y:1992:i:2:p:176-188

[28] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (Im)possibility of fairness: different value systems require different mechanisms for fair decision making. *Commun. ACM* 64, 4 (March 2021), 136–143. https://doi.org/10.1145/3433949

[29] Pratik Gajane and Mykola Pechenizkiy. 2017. On Formalizing Fairness in Prediction With Machine Learning. *arXiv:1710.03184*. http://arxiv.org/abs/1710.03184

[30] Przemyslaw A. Grabowicz, Nicholas Perello, and Aarshee Mishra. 2022. Marrying Fairness and Explainability in Supervised Learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1905–1916. https://doi.org/10.1145/3531146.3533236

[31] Ben Green. 2018. Fair Risk Assessments: A Precarious Approach for Criminal Justice Reform. *5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*. https://www.fatml.org/media/documents/fair_risk_assessments_criminal_justice.pdf

[32] Ben Green. 2022. Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness. *Philosophy & Technology* 35, 4 (2022), 90. https://doi.org/10.1007/s13347-022-00584-6

[33] Robert M Guion. 1966. Employment Tests and Discriminatory Hiring. *Industrial Relations: A Journal of Economy and Society* 5, 2 (1966), 20–37. https://doi.org/10.1111/j.1468-232X.1966.tb00449.x Publisher: Wiley Online Library.

[34] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In *Proceedings of the 35th International Conference on Machine Learning*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 1939–1948. https://proceedings.mlr.press/v80/hebert-johnson18a.html

[35] Brian Hedden. 2021. On Statistical Criteria of Algorithmic Fairness. *Philosophy & Public Affairs* 49, 2 (2021), 209–231. https://doi.org/10.1111/papa.12189

[36] Ben Hutchinson and Margaret Mitchell. 2019. 50 Years of Test (Un)Fairness: Lessons for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 49–58. https://doi.org/10.1145/3287560.3287600

[37] Marc-Thorsten Hütt and Claudia Schubert. 2020. Fairness von KI-Algorithmen. In *Philosophisches Handbuch Künstliche Intelligenz*, Klaus Mainzer (Ed.). Springer Fachmedien Wiesbaden, Wiesbaden, 1–22. http://link.springer.com/10.1007/978-3-658-23715-8_39-1

[38] Alexandra Kalev, Frank Dobbin, and Erin Kelly. 2006. Best Practices or Best Guesses? Assessing the Efficacy of Corporate Affirmative Action and Diversity Policies. *American Sociological Review* 71, 4 (2006), 589–617. https://doi.org/10.1177/000312240607100404

[39] Atoosa Kasirzadeh and Andrew Smart. 2021. The Use and Misuse of Counterfactuals in Ethical Machine Learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 228–236. https://doi.org/10.1145/3442188.3445886

[40] Maximilian Kasy and Rediet Abebe. 2021. Fairness, Equality, and Power in Algorithmic Decision-Making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 576–586. https://doi.org/10.1145/3442188.3445919

[41] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 2564–2572. https://proceedings.mlr.press/v80/kearns18a.html

[42] Stephen N. Keith, Robert M. Bell, August G. Swanson, and Albert P. Williams. 1985. Effects of Affirmative Action in Medical Schools. *New England Journal of Medicine* 313, 24 (1985), 1519–1525. https://doi.org/10.1056/NEJM198512123132406

[43] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding Discrimination through Causal Reasoning. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/f5f8590cd58a54e94377e6ae2eded4d9-Abstract.html

[44] Michael P Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-box Post-processing for Fairness in Classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, New York, NY, USA, 247–254. https://doi.org/10.1145/3306618.3314287

[45] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 67)*, Christos H. Papadimitriou (Ed.). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 43:1–43:23. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

[46] Youjin Kong. 2022. Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 485–494. https://doi.org/10.1145/3531146.3533114

[47] Matthias Kuppler, Christoph Kern, Ruben L. Bach, and Frauke Kreuter. 2021. Distributive Justice and Fairness Metrics in Automated Decision-making: How Much Overlap Is There? *arXiv:2105.01441*. http://arxiv.org/abs/2105.01441

[48] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

[49] Matt J. Kusner and Joshua R. Loftus. 2020. The Long Road to Fairer Algorithms. *Nature* 578, 7793 (2020), 34–36. https://doi.org/10.1038/d41586-020-00274-3

[50] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[51] Michelle Lee, Luciano Floridi, and Jat Singh. 2020. From Fairness Metrics to Key Ethics Indicators (KEIs): A Context-Aware Approach to Algorithmic Ethics in an Unequal Society. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3679975

[52] Michelle Seng Ah Lee, Luciano Floridi, and Jatinder Singh. 2021. Formalising Trade-offs Beyond Algorithmic Fairness: Lessons from Ethical Philosophy and Welfare Economics. *AI and Ethics* 1, 4 (2021), 529–544. https://doi.org/10.1007/s43681-021-00067-y

[53] Michele Loi and Christoph Heitz. 2022. Is Calibration a Fairness Requirement? An Argument from the Point of View of Moral Philosophy and Decision Theory. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2026–2034. https://doi.org/10.1145/3531146.3533245

[54] Robert Long. 2021. Fairness in Machine Learning: Against False Positive Rate Equality as a Measure of Fairness. *Journal of Moral Philosophy* 19, 1 (2021), 49–78. https://doi.org/10.1163/17455243-20213439 Brill.

[55] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6 (July 2022), 1–35. https://doi.org/10.1145/3457607

[56] Nicolai Meinshausen. 2006. Quantile Regression Forests. *Journal of Machine Learning Research* 7, 35 (2006), 983–999. http://jmlr.org/papers/v7/meinshausen06a.html

[57] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8, 1 (2021), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

[58] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. 2019. Learning Optimal Fair Policies. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 4674–4682. https://proceedings.mlr.press/v97/nabi19a.html ISSN: 2640-3498.

[59] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'18/IAAI'18/EAAI'18)*. AAAI Press, New Orleans, Louisiana, USA, 1931–1940. https://doi.org/10.5555/3504035.3504270

[60] Weishen Pan, Sen Cui, Jiang Bian, Changshui Zhang, and Fei Wang. 2021. Explaining Algorithmic Fairness Through Fairness-Aware Causal Path Decomposition. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. ACM, Virtual Event Singapore, 1287–1297. https://doi.org/10.1145/3447548.3467258

[61] Judea Pearl. 2009. *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press. https://yzhu.io/courses/core/reading/04.causality.pdf

[62] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware Data Mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08)*. Association for Computing Machinery, New York, NY, USA, 560–568. https://doi.org/10.1145/1401890.1401959

[63] Drago Plečko, Nicolas Bennett, and Nicolai Meinshausen. 2024. fairadapt: Causal Reasoning for Fair Data Preprocessing. *Journal of Statistical Software* 110, 4 (2024), 1–35. https://doi.org/10.18637/jss.v110.i04

[64] Drago Plečko and Nicolai Meinshausen. 2020. Fair Data Adaptation with Quantile Preservation. *Journal of Machine Learning Research* 21 (2020), 1–44. http://jmlr.org/papers/v21/19-966.html

[65] John A Rawls. 2003. *A Theory of Justice*. Belknap Press of Harvard Univ. Press. https://www.hup.harvard.edu/catalog.php?isbn=9780674000780

[66] John E Roemer. 2018. *Theories of Distributive Justice*. Harvard University Press. https://www.hup.harvard.edu/catalog.php?isbn=9780674879201

[67] Pola Schwöbel and Peter Remmers. 2022. The Long Arc of Fairness: Formalisations and Ethical Discourse. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 2179–2188. https://doi.org/10.1145/3531146.3534635

[68] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, Atlanta GA USA, 59–68. https://doi.org/10.1145/3287560.3287598

[69] Ashudeep Singh, David Kempe, and Thorsten Joachims. 2021. Fairness in Ranking under Uncertainty. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 11896–11908. https://proceedings.neurips.cc/paper/2021/file/63c3ddcc7b23daa1e42dc41f9a44a873-Paper.pdf

[70] Harini Suresh and John V. Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9. , 9 pages. https://doi.org/10.1145/3465416.3483305

[71] Robert L Thorndike. 1971. Concepts of Culture-Fairness. *Journal of Educational Measurement* 8, 2 (1971), 63–70. http://www.jstor.org/stable/1433959

[72] Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. In *Proceedings of the International Workshop on Software Fairness*. ACM, Gothenburg Sweden. https://doi.org/10.1145/3194770.3194776

[73] Sandra Wachter. 2019. Affinity Profiling and Discrimination by Association in Online Behavioural Advertising. *Berkeley Technology Law Journal* 35, 2 (2019). https://doi.org/10.2139/ssrn.3388639

[74] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law. *West Virginia Law Review* 123, 3 (Jan. 2021), 735–790. https://doi.org/10.2139/ssrn.3792772

[75] Pak-Hang Wong. 2020. Democratizing Algorithmic Fairness. *Philosophy & Technology* 33, 2 (2020), 225–244. https://doi.org/10.1007/s13347-019-00355-w

[76] Rebecca Zwick and Neil J Dorans. 2016. Philosophical Perspectives on Fairness in Educational Assessment. In *Fairness in educational assessment and measurement*. Routledge, 283–298. https://doi.org/10.4324/9781315774527