Yet Another Impossibility Theorem in Algorithmic Fairness

Fabian Beigang

Abstract

In recent years, there has been a surge in research addressing the question which properties predictive algorithms ought to satisfy in order to be considered fair. Three of the most widely discussed criteria of fairness are the criteria called equalized odds, predictive parity, and counterfactual fairness. In this paper, we will present a new impossibility result involving these three criteria of algorithmic fairness. In particular, we will argue that there are realistic circumstances under which any predictive algorithm that satisfies counterfactual fairness will violate both other fairness criteria, that is, equalized odds and predictive parity. As will be shown, this impossibility result forces us to give up one of four intuitively plausible assumptions about algorithmic fairness. We will explain and motivate each of the four assumptions and discuss which of them can plausibly be given up in order to circumvent the impossibility.

Keywords: algorithmic fairness, causal modeling, machine learning, impossibility theorem, AI ethics

1 Introduction

Machine learning techniques are increasingly applied to inform decisions in the public and economic sphere. Prominent examples are criminal sentencing, policing, hiring, and credit lending decisions. Since decisions in these domains potentially have a large positive or negative impact on individuals, they are subject to equality of opportunity and non-discrimination norms. When algorithms are applied to decision-making in the public and economic sphere, it is, therefore, critical to ensure that they too do not produce discriminatory outcomes or disproportionately harm specific social groups. Much recent research has gone into developing and discussing fairness constraints for machine learning models with the aim of providing tools for doing this.

In this paper, we consider the relation between three of the most popular fairness constraints: *counterfactual fairness*, *equalized odds*, and *predictive parity*. Counterfactual fairness formalizes the idea that in a given prediction, the protected characteristic (e.g. gender, ethnicity, or religion) should not make a (causal) difference to the prediction (Kusner et al., 2017). Equalized odds, in contrast, formalizes the idea that in a given population, the false positive and false negative error rates of a predictive model should be independent of the protected characteristic (Hardt et al., 2016). And lastly, predictive parity is concerned with the predictive value, that is, the probability that the predicted property is indeed present (or absent), given that an individual received a positive (or negative) prediction. Predictive parity formalizes the idea that in a given population, the predictive value of a model should be independent of the protected characteristic (Chouldechova, 2017).

The central contribution of this paper is an impossibility theorem with regard to the relation between the three criteria. It establishes that whenever the protected characteristic has some causal relevance to the variable that is to be predicted, a counterfactually fair predictive model will with logical necessity violate both, equalized odds and predictive parity. The result forces us to give up one of four individually plausible assumptions about algorithmic fairness. These assumptions are (1) that fairness requires that either equalized odds or predictive parity is satisfied, (2) that predictions should be counterfactually fair, (3) that protected characteristics (like age, gender, etc.) can, in some cases, influence the variable of interest for the prediction, and lastly, (4) that we can always find a fair way of making a prediction. A way to interpret this impossibility result is that we either have to accept that counterfactual fairness is not a requirement of fairness for predictive models, or that neither equalized odds nor predictive parity are requirements of fairness. If none of these two interpretations seem plausible, we either have to accept that there are situations for which no fair predictive models exist, or deny that the type of situation in which the impossibility arises ever occurs.

Other works have explored impossibilities and trade-offs between other fairness criteria. Most famously, Chouldechova (2017) and Kleinberg et al. (2016) have shown that under realistic conditions, equalized odds and predictive parity are mutually inconsistent. Attempts to reconcile weaker or approximate versions of the criteria are discussed in (Pleiss et al., 2017), who show that predictive parity and a weakened form of equalized odds can be reconciled when using a randomized prediction scheme, (Celis et al., 2019), who present an algorithmic approach that allows to satisfy approximate versions of multiple fairness criteria at the same time, and (Beigang, 2023), who shows that causally reinterpreted versions of equalized odds and predictive parity are mutually consistent.

Some earlier works have discussed the limitations of counterfactual fairness. A number of articles propose alternative causal fairness criteria which relax counterfactual fairness and would potentially avoid the results discussed here. Chiappa (2017) and Loftus et al. (2018) provide frameworks for analyzing whether individual causal paths in a model satisfy counterfactual fairness, allowing for the possibility of some of those paths to not be subject to fairness constraints. Kilbertus et al. (2017) present an alternative causal fairness constraint in which causal effects of the protected characteristic on the prediction that are not mediated by proxy variables are considered fair. Practical limitations of counterfactual fairness have been addressed by Kilbertus et al. (2020), Wu et al. (2019) and Russell et al. (2017). To our knowledge, no previous work discusses the incompatibilities presented in this article in depth.

The remainder of the paper is organized as follows. In section 2, we provide an introduction to the mathematical framework used in defining the fairness criteria and the proof of the impossibility theorem. In section 3, we introduce the three fairness criteria counterfactual fairness, equalized odds, and predictive parity. In section 4, we state and prove the impossibility theorem before then discussing ways to circumvent it in section 5. We close with a brief summary in section 6.

2 Mathematical framework

We begin by specifying the mathematical framework in which the fairness criteria are defined. This will moreover serve as a basis for the proof of the impossibility theorem presented in this paper.

The notation we use is as follows. We denote random variables by capital letters, e.g. X, their values by lower-case letters, e.g. x, and the domain of a variable X by D_X. A set of variables $X_1, X_2, ..., X_n$ is denoted in boldface by $\mathbf{X}$ with value $\mathbf{x}$. The domain $D_{\mathbf{X}}$ of a set of n variables is defined as the Cartesian product of the domains of the individual variables in the set, i.e. $D_{X_1} \times D_{X_2} \times ... \times D_{X_n}$. The probability that a variable X takes value x, $P(X = x)$, will be abbreviated by $P(x)$ when this is unambiguous. Two variables (and analogously sets of variables) X and Y are said to be conditionally independent given variable Z (in probability distribution $P(\cdot)$) if and only if $P(x \mid y, z) = P(x \mid z)$ for all $x \in D_X$, $y \in D_Y$ and $z \in D_Z$. We denote conditional independence by $(X \perp\!\!\!\perp Y \mid Z)$.

2.1 Causal models and counterfactuals

We next introduce a number of central concepts from the mathematical framework of causal modeling as developed by Pearl (2009). A *causal model* is defined as a triple $(\mathbf{U}, \mathbf{V}, F)$ such that (i) $\mathbf{U}$ is a set of variables whose values are determined by factors outside the present model, (ii) $\mathbf{V}$ is a set $\{V_1, V_2, ..., V_n\}$ of variables whose values are determined by other variables in the model, that is, by a subset of the variables in $\mathbf{U}$ and $\mathbf{V}$, and (iii) F is a set of structural equations $\{f_1, f_2, ..., f_n\}$ such that each structural equation f_i is a mapping from $D_{\mathbf{PA}_i}$ to D_{V_i}, with $\mathbf{PA}_i \subseteq \mathbf{U} \cup (\mathbf{V} \setminus V_i)$ (Pearl, 2009, p. 203). This means, a causal model encodes for each variable V_i how it functionally depends on the other variables in $\mathbf{V}$ and $\mathbf{U}$, or, in other words, what its direct causes are.

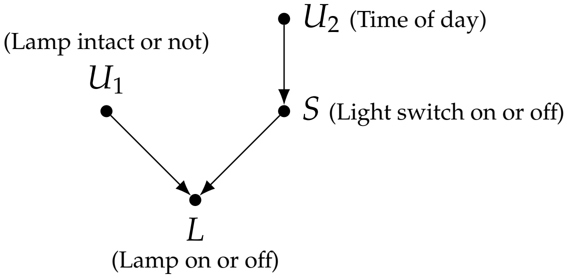

Figure 1: An example of a causal model (represented by its causal structure).

To illustrate this definition, take for example the causal model $(\mathbf{U}, \mathbf{V}, F)$ depicted in Figure 1, which represents the causal mechanism of a lamp. In this example, $\mathbf{U}$ contains the binary variables U_1 and U_2, which represent whether the lamp is intact or whether it is broken, and which time it is, respectively. $\mathbf{V}$ contains the variables L and S, which represent whether the light switch is in the ″On″ or in the ″Off″ position, and whether the lamp is on or off. F contains two structural equations:

$$l = f_1(s, u_1) = min(s, u_1)$$
$$s = f_2(u_2) = \begin{cases} 1 & \text{if } u_2 > 17 \\ 0 & \text{if } u_2 \leq 17 \end{cases}$$

First, it contains f_1, which specifies that the lamp is on whenever it is intact and the light switch is in the ″On″ position. This is formalized as $min(s, u_1)$: whenever the lamp is intact, $u_1 = 1$, and whenever the light switch is on, $s = 1$, and consequently, $L = min(s, u_1) = min(1, 1) = 1$ – the lamp is on. Whenever one of u_1 or s takes the value 0, representing that either the light switch is off or that the lamp is not intact, $l = min(s, u_1) = 0$ – the lamp is off. Secondly, f_2 specifies that whether the light switch is on depends on the time of day, in particular, whether it is after 17:00. There are no structural equations for the variables U_1 and U_2, as their respective values are determined by factors that are not represented in our model.

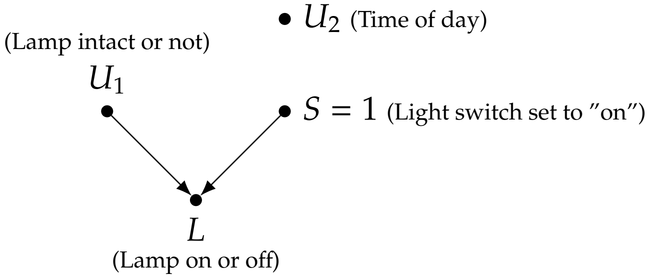

Figure 2: Submodel of the original causal model. By setting $S = 1$, the initial link from U_2 to S is deleted.

On the basis of the above definition of a causal model, we can introduce the notion of a *submodel*. A submodel of a causal model M is itself a causal model $M_{\mathbf{X=x}} = (\mathbf{U}, \mathbf{V}, F_{\mathbf{X=x}})$ where $F_{\mathbf{X=x}} = \{f_i : V_i \notin \mathbf{X}\} \cup \{\mathbf{X} = \mathbf{x}\}$ for a particular realization $\mathbf{X} = \mathbf{x}$ of a set of variables $\mathbf{X} \subseteq \mathbf{V}$. Figure 2 illustrates this. Here, we have replaced the structural equation for the variable S by the value 1. This can be interpreted as an external actual or hypothetical intervention on this variable. An intervention of this kind deletes all functional dependencies on other variables. In our example, this means that when we decide to artificially intervene in the system by turning the light switch on, the position of the switch does not depend on the time of day anymore.

Next, we need to introduce the notion of a *potential response*. A potential response $Y_{\mathbf{X}=\mathbf{x}}(\mathbf{U} = \mathbf{u})$ represents the value that the variable Y takes according to the set of equations $F_{\mathbf{x}}$ and a particular realization $\mathbf{u}$ of the background variables $\mathbf{U}$ (Pearl, 2009, p. 204). In our example, we can for instance think of how the lamp would potentially respond to the intervention of setting the light switch to "on", assuming that the lamp happens to be intact. According to the structural equations above, this would lead to the lamp being on. Consequently, this potential response would be formalized as $L_{S=1}(U_1 = 1) = 1$. For the sake of simplicity, we will henceforth leave the background variables implicit and denote the potential response by $Y_{\mathbf{X}=\mathbf{x}}$. Moreover, where it is unambiguous which variable we refer to, we will abbreviate this by $Y_{\mathbf{x}}$.

The notion of a potential response now allows to define *counterfactual* statements of the form "The value that Y would have obtained, had $\mathbf{X}$ been $\mathbf{x}$" (for $\mathbf{X}, Y \subseteq \mathbf{V}$) as the potential response $Y_{\mathbf{x}}$. Given a causal model M and a probability distribution $P(\mathbf{u})$ over $D_{\mathbf{U}}$, the conditional probability of a counterfactual "If it were the case that $\mathbf{X} = \mathbf{x}$, then it would be the case that $Y = y$" given evidence $\mathbf{e}$ can be evaluated by (1) updating $P(\mathbf{u})$ by conditioning on evidence $\mathbf{e}$ in order to obtain $P^*(\mathbf{u}) = P(\,\mathbf{u} \mid \mathbf{e}\,)$, (2) generating the submodel $M_{\mathbf{x}}$ of M obtained by removing the structural equation for $\mathbf{X}$ from M and replacing it by a constant $\mathbf{x}$, (3) using the submodel $M_{\mathbf{x}}$ and the updated probability distribution $P^*(\mathbf{u})$ to compute the probability of $Y = y$. This probability of the counterfactual statement is denoted by $P(\,Y_{\mathbf{x}} = y \mid \mathbf{e}\,)$.

Let us again illustrate this with our example. Assume we attempt to determine the probability of the counterfactual "If the light switch had been turned on, then the lamp would be on", knowing that the lamp is actually not on. The formalization of this is $P(\,L_{S=1} = 1 \mid L = 0\,)$. Now we simply have to run through the three steps. First, we update the relevant background variables. In general, this would be both, U_1 and U_2, but here only U_1 is relevant[1]. Let us assume that initially, we would think that it 90% likely that the lamp is intact, i.e. $P(\,U_1 = 1) = 0.9$. After learning that the lamp is currently off ($L = 0$), we update the probability assignment to, say, $P(\,U_1 = 1 \mid L = 0\,) = P^*(U_1 = 1) = 0.8$, reflecting the fact that the lamp being off is weak evidence for the lamp being broken. Next, we have to generate a submodel (see Figure 2) by replacing $s = f_2(u_2)$ with $s = 1$. Using the updated probability assignment and the submodel, we can now calculate $P(\,L_{S=1} = 1 \mid L = 0\,) = P^*(min(s, u_1) = 1) = P^*(min(1, u_1) = 1) = P^*(U_1 = 1) = 0.8$. In words, the probability that the lamp would have been on, if the switch had been on, is simply the probability of the lamp being intact, given we observe that actually the lamp is off.

[1] This is because after the intervention on S, L is screened off from U_2.

This is due to the fact that the lamp is only on if both, the switch is on and the lamp is intact.

2.2 Causal structures and the projection theorem

If we strip a causal model of its parameters (i.e. the information on the coefficients of the structural equations in F), we obtain a *causal structure* (Pearl, 2009, p. 203). A causal structure can be represented as a directed acyclic graph in which each node corresponds to a variable in $\mathbf{U} \cup \mathbf{V}$, and in which there is a directed edge pointing toward $V_i \in \mathbf{V}$ from every node corresponding to a variable that occurs in f_i. A directed edge from one node to another consequently represents a direct causal link between the corresponding variables. More intuitively speaking, the causal structure contains purely qualitative information about the causal relations between the variables in the model. Figures 1 and 2 are examples of causal structures: each figure represents the qualitative information about causal dependencies between variables in a qualitative way. It is not apparent from the graph what functional form the causal dependencies between the variables exactly take. To denote the causal structure of a causal model M, we will henceforth write G_M.

In order to establish a connection between a causal structure and an associated probability distribution over the variables represented as nodes, we need to introduce the notion of *d-separation*. Two variables X and Y are said to be d-separated by $\mathbf{Z}$ in a causal structure G_M if and only if each path between the nodes representing X and Y contains either (i) a chain ($i \rightarrow m \rightarrow j$) or a fork ($i \leftarrow m \rightarrow j$), and m is a node representing a variable in $\mathbf{Z}$, or (ii) a collider ($i \rightarrow m \leftarrow j$) and m is a node representing a variable not in $\mathbf{Z}$ (Pearl, 2009, pp. 16-17). In figure 1, for example, the nodes U_1 and S are d-separated by the empty set, as are U_2 and S, while L and U_2 are d-separated by S.

We say that a probability distribution $P(\cdot)$ is *Markov* relative to a causal structure G_M if for any X, Y, and $\mathbf{Z}$, it is the case that if $\mathbf{Z}$ d-separates X and Y, it is also the case that $(X \perp\!\!\!\perp Y \mid \mathbf{Z})$ (Pearl, 2009, p. 26). Conversely, we say that $P(\cdot)$ is *faithful* to G_M if $(X \perp\!\!\!\perp Y \mid \mathbf{Z})$ implies that $\mathbf{Z}$ d-separates X and Y. In a probability distribution that is Markov compatible with the causal structure in Figure 1, the following conditional independencies have to hold: $(U_1 \perp\!\!\!\perp U_2 \mid \emptyset)$, $(U_1 \perp\!\!\!\perp U_2 \mid S)$, $(U_1 \perp\!\!\!\perp S \mid \emptyset)$, and $(L \perp\!\!\!\perp U_2 \mid S)$. For the probability distribution to be faithful to the causal structure, there must not be any other (conditional) independencies between the variables. Henceforth, we will generally assume that a probability distribution $P(\cdot)$ is both *faithful* and *Markov* relative to its associated causal structure G_M. [2]

The above notions now allow us to describe a theorem which will later help us construct an economical proof of the impossibility theorem presented in this paper. We will call this theorem the *projection theorem*. It states the following: For every set of observed variables $\mathbf{O}$, there exists a causal structure with a node U_{ij} for each pair of variables $O_i, O_j \in \mathbf{O}$, representing their (potential) latent common cause, which is (Markov) compatible with the joint probability distribution over $D_{\mathbf{O}}$ (Verma, 1993). More intuitively speaking, whenever we have a set of variables and a joint probability distribution over these variables but no information about their causal dependencies, we are guaranteed to be able to represent the "correct" causal structure of these variables, if, in addition, for each pair of observed variables we assume the existence of a hidden variable which potentially influences both observed variables simultaneously.

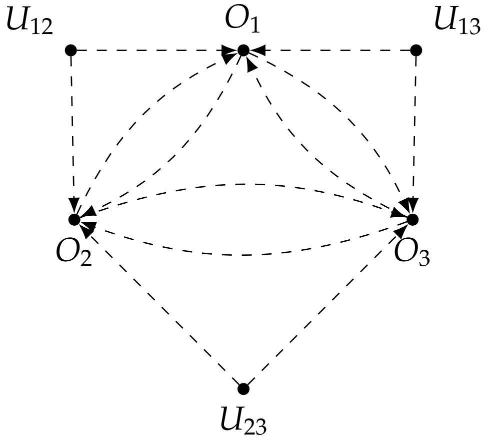

Figure 3: Representation of all the possible causal structures between O_1, O_2, and O_3.

If, for example, we are interested in the variables O_1 (which stands for, say, "diabetes"), O_2 ("sugar consumption") and O_3 ("weekly amount of exercise"), and know their joint probability distributions but not what the causal relations between the variables are, we can assume there

[2] For a discussion and defense of the two assumptions, see Pearl (2009, pp. 61-64) and Zhang & Spirtes (2016).Note, that some authors occassionally use an assumption weaker than faithfulness, namely causal minimality (Zhang & Spirtes, 2011). The argument in this paper, however, relies on inferential steps in both directions: from (conditional) independencies to properties of the causal graph, as well as from properties of the causal graph to (conditional) indepenencies. The minimality condition alone would not suffice to allow for these steps under all possible probabilistic parameters in the causal model (if, say, the effect of one causal path were to exactly undo the effect along another one in terms of a change in the probability distribution).

is a causal structure (i.e. directed acyclic graph) of O_1, O_2, O_3 and three hidden variables U_{12},U_{23}, and U_{13}, to which their probability distribution is Markov compatible. U_{12} would, in this structure, stand for any unobserved background factor which could simultaneously influence whether a person has diabetes and what their level of sugar consumption is - for instance, some genetic disposition. Figure 3 illustrates this: the projection theorem guarantees that, if we choose the right arrows (among the possible, dashed arrows), we obtain a graph that depicts the correct causal structure between O_1, O_2, and O_3.

3 Fairness in predictive models

Let us now turn back to the topic of predictive models in algorithmic decision-making. Predictive models can exhibit discriminatory bias. That is, the predictions of a machine learning model can be such that decisions based on them would constitute cases of discrimination relative to a given protected characteristic. Protected characteristics are typically traits such as ethnicity, gender, religion, or disability. In recent years, several fairness constraints have been proposed with the aim of ensuring that, provided a predictive model satisfies the fairness constraint, the model is guaranteed to not exhibit such discriminatory bias. Each of these different constraints interprets the notion of discriminatory bias in a different way. While different proposals for fairness constraints abound, three constraints are at the center of the debate: counterfactual fairness, equalized odds, and predictive parity.

We will now introduce the three formal definitions and their underlying rationales. In order to define the criteria in a rigorous fashion, let $\mathbf{X} \subseteq \mathbf{U} \cup \mathbf{V}$ be a set of input variables, $Y \in \mathbf{V}$ the target variable, i.e. the variable representing the presence or absence of the property of interest which is unknown at the time of prediction, and $A \in \mathbf{U}$ the protected characteristic relative to which we aim to evaluate or constrain the predictive model. For the sake of simplicity, we will assume Y to be a binary variable taking the values 0 or 1. If, for a given individual, it is the case that $Y = 1$, we will say that the individual belongs to the *positive class*. We will moreover assume that A is a binary variable with values a_1 and a_2, which represent the presence and the absence of the protected characteristic, respectively. When we refer to protected groups, we refer to the groups constituted by individuals with property $A = a_1$ and individuals with property $A = a_2$. Finally, let us denote the causal model representing the mechanisms of the real world situation within which the (sets of) variables $\mathbf{X}, Y$ and A are situated as $M_{base} = (\mathbf{U}_{base}, \mathbf{V}_{base}, F_{base})$.

Let us next turn to the representation of *predictive models*. To this end, let $\hat{Y}$ be a binary variable which is interpreted as an attempt to predict the value of the target variable Y. Whenever $\hat{Y} = 1$, we will speak of a *positive prediction*. Generally, we will take predictive models to be functions of the form $g : D_{\mathbf{X}} \rightarrow D_{\hat{Y}}$, that is, functions from a vector of *input values* $\mathbf{x}$ to a *prediction* $\hat{y}$. This is a simplifying assumption since many predictive models provide a probability estimate of the presence of a property instead of an outright prediction of the property's presence or absence. To keep the discussion simple, however, we will in this paper assume that predictions are binary. This means, we assume that the model either predicts that the property y is present or that it is absent. This simplification does not affect the generality of the result presented here.

For given $\mathbf{X}, Y$ and A in a causal model M_{base}, a predictive model g can be represented within an augmented causal model $M_{aug} = (\mathbf{U}_{base}, \mathbf{V}_{aug}, F_{aug})$, where $\mathbf{V}_{aug} = \mathbf{V}_{base} \cup \hat{Y}$, and where F_{aug} is the extension of F_{base} obtained by adding the function g representing the predictive model as a structural equation to F_{base}. We here interpret the function g as the causal relation between the predictive model's input variables $\mathbf{X}$ and the prediction $\hat{Y}$. For every predictive model, there consequently is a specific augmented causal model representing the causal relations between relevant variables and the prediction. Since g is a deterministic function of $\mathbf{X}$, which is a subset of $\mathbf{U} \cup \mathbf{V}$, the joint probability distribution over the variables in the augmented causal model is readily obtained from the set of structural equations F_{base} and the probability distribution over the exogenous variables $P(\mathbf{u})$. Subsequently, when we speak about causal relations we will always do so relative to a specific predictive model g, hence referring to causal relations within an augmented causal model as outlined above.

3.1 Equalized odds

The first fairness constraint we introduce is *equalized odds* (Hardt et al., 2016). It formalizes the requirement that a predictive model produce equal false positive and false negative error rates across protected groups. The underlying idea here is that a disparity in error rates across protected groups indicates that the model is biased with regard to a group in that it takes the protected characteristic (or proxies thereof) to be more predictive of the target variable than it actually is. If, for example, a predictive model is applied to predict whether a defendant is at risk of reoffending or not, and it has a higher false positive rate for African-American defendants than for white defendants, this means that a greater proportion of low-risk African-American defendants will be falsely predicted to be at high risk than is the case for white defendants. Implicitly, the model seems to overestimate how predictive the trait of being African-American (or information closely linked to it, like for instance living in a certain neighborhood, or having a certain name) is of recidivism. Overestimating how predictive a person's ethnicity is of some other property can clearly be considered a form of bias against (or towards) people of this ethnicity.

In practical terms, different error rates reflect that a different standard is applied to one protected group than to the other, or so the argument goes. In the recidivism example, individuals of the group with a higher false positive rate are held to a higher standard - on average, they have to satisfy stricter conditions (as reflected in the information that serves as input to the model) in order to be deemed to be at low risk of recidivism than individuals of the other group. Equalized odds can be formalized as follows:

Definition 1 (Equalized odds)**.** A predictive model g satisfies *equalized odds* (relative to A) if and only if for all $\hat{y} \in D_{\hat{Y}}$ and $y \in D_Y$

$$P(\, \hat{y} \mid a_1,\, y\,) = P(\, \hat{y} \mid a_2,\, y\,) \tag{1}$$

This formalization can be understood as requiring that the value of the prediction $\hat{Y}$ be independent of the value of the protected characteristic A, once we control for the actual value of the target variable Y. Applied to the above example, it means that the probability of being deemed to be at high risk of recidivism (or low risk, respectively) should be equal across low risk African-American and low risk white defendants (and, analogously, it should be equal across high risk African-American and high risk white defendants).

By the axioms of probability and the definition of conditional independence, equalized odds is equivalent to $(\, \hat{Y} \perp\!\!\!\perp A \mid Y\,)$. This, in turn, is equivalent to $\hat{Y}$ and A being d-separated by Y in the associated causal structure, due to the assumption that $P(\cdot)$ is Markov and faithful.

3.2 Predictive parity

Next, we introduce the fairness constraint called *predictive parity* (Chouldechova, 2017). The central metric used in this constraint is positive (and negative) predictive value. The positive predictive value of a predictive model is the proportion of instances that actually belong in the positive class among those that received a positive prediction. Analogously, the negative predictive value is the proportion of instances that actually do not belong in the positive class among those that did not receive a positive prediction. Predictive parity requires that these two metrics be equal across protected groups.

In our running example, this would mean that the proportion of defendants who go on to reoffend among those who received a high recidivism risk prediction should be equal for African-American and white defendants (and, of course, analogously for negative predictions). The rationale behind this is that predictions should be equally informative and reliable across different protected groups. If the positive predictive value is much lower for one protected group than for another, this means that positive predictions for individuals of this group are less trustworthy, and are less indicative of the individual actually being in the positive class, than for individuals of a different protected group. More intuitively speaking, a prediction of being at high risk of recidivating should mean the same for an African American and a white defendant. This idea can be expressed as the following mathematical constraint:

Definition 2 (Predictive parity)**.** A predictive model g satisfies *predictive parity* (relative to A) if and only if for all $\hat{y} \in D_{\hat{Y}}$ and $y \in D_Y$

$$P(\, y \mid a_1,\, \hat{y}\,) = P(\, y \mid a_2,\, \hat{y}\,) \tag{2}$$

Analogously to equalized odds, predictive parity can be expressed in terms of conditional independence by stating that the value of the target variable Y should be independent of the protected characteristic A, once we control for the value of the prediction $\hat{Y}$. Formally, this can be expressed as $(\, Y \perp\!\!\!\perp A \mid \hat{Y}\,)$. For the associated causal structure, this means that Y and A are d-separated by $\hat{Y}$.

3.3 Counterfactual fairness

The third and most complex fairness constraint we introduce is *counterfactual fairness* (Kusner et al., 2017). It formalizes the requirement that an individual with a given value of a protected characteristic would have received the same prediction as they actually received, had their protected characteristic A taken a different value, while everything else that is not causally downstream of the protected characteristic had stayed the same. In other words, if the predictive model is fair, the change in the value of the protected characteristic does not make a difference to the prediction for an otherwise identical individual. Whether a predictive model is counterfactually fair is not determined by the probability distribution $P(\cdot)$ alone, but requires a fully specified causal model. Otherwise, the probability of the counterfactual statement could not be calculated. Given such a model M, counterfactual fairness can be defined as follows:

Definition 3 (Counterfactual fairness)**.** A predictive model g satisfies *counterfactual fairness* (relative to a_1) if and only if for all $\hat{y} \in D_{\hat{Y}}$ and $\mathbf{x} \in D_{\mathbf{X}}$

$$P(\,\hat{Y}_{a_1} = \hat{y} \mid \mathbf{x}, a_1\,) - P(\,\hat{Y}_{a_2} = \hat{y} \mid \mathbf{x}, a_1\,) = 0 \tag{3}$$

Note that, other than equalized odds and predictive parity, counterfactual fairness is defined relative to a specific trait a_1, rather than the variable A. For example, equalized odds might determine whether error rates are equally distributed among, say, different religious groups, but counterfactual fairness determines whether one specific group's trait, say being Muslim as opposed to being Christian, makes a difference to a given prediction. This, however, does not mean that counterfactual fairness has to be used in a trait-relative way. If a given context calls for it, one could require that counterfactual fairness be satisfied for all $a \in D_A$, rather than just for a specific trait a_1. Here, we will only assume the weaker version of counterfactual fairness, as this will strengthen the impossibility result.

The above definition of counterfactual fairness implies that there is no causal chain from A to $\hat{Y}$ in the causal structure G_M. To see this, note that by the semantics of counterfactuals we need to consider the submodel M_{a_1} (in which the structural equation for A was replaced by the constant a_1) in order to determine the probability of the counterfactual statement. With regard to the graph, this means that all the incoming edges into A are removed. Any outgoing edges from A

remain intact. Counterfactual fairness then requires that (given a specific assignment of a joint probability distribution to the latent variables in $\mathbf{U}$) in the resulting probability distribution P_{a_1} associated with the submodel M_{a_1}, $\hat{Y}$ is independent of A, i.e. $(\hat{Y} \perp\!\!\!\perp_{a_1} A)$. By the assumption of faithfulness, this entails that A and $\hat{Y}$ are d-separated by the empty set in the causal structure $G_{M_{a_1}}$. In particular, this means that there is no causal chain from A to $\hat{Y}$. Since any outgoing edges from A would have remained intact in the submodel and would hence also exist in M_{a_1}, we can conclude that there is also no causal chain from A to $\hat{Y}$ in the causal structure G_M of the original causal model M.

4 An impossibility theorem

As it turns out, there are circumstances under which counterfactual fairness is incompatible with both, equalized odds and predictive parity. In particular, we will show that the following four individually plausible propositions are jointly inconsistent:

(1) If a predictive model is fair, it satisfies equalized odds or predictive parity.

(2) If a predictive model is fair, it satisfies counterfactual fairness.

(3) There are some morally relevant prediction contexts where the protected characteristic has some (possibly mediated) causal relevance to the target variable.

(4) For every morally relevant prediction context there exists a fair predictive model.

We will briefly explain each of the four propositions in turn. Proposition (1) states that it is necessary for a fair predictive model to at least satisfy one of the two fairness constraints equalized odds and predictive parity. While both are prima facie plausible, they were shown to be mutually incompatible whenever the base rate prevalence of the predicted property differs among protected groups (Kleinberg et al., 2016; Chouldechova, 2017). Hence, we cannot require that a fair model generally satisfy both, but it seems like a relatively weak desideratum to require that a fair model satisfy at least one of the two. Proposition (2) simply states that it is necessary for a fair predictive model to satisfy counterfactual fairness.

Proposition (3) contains a number of concepts that require explaining. First, by prediction context we mean a situation in which a specific property is being predicted, for instance, whether a given applicant would be a profitable employee for the hiring company. We say that a prediction context is morally relevant when the prediction and the subsequent decision are subject to moral norms, like for instance non-discrimination or equality of opportunity norms. To make precise what it means that a protected characteristic is causally relevant to the target variable, we have to refer to the causal modeling framework outlined in section 2. Using this framework, we can say that the protected characteristic is causally relevant to the target variable if there is a (hypothetical) intervention on the former that results in a change of the probability distribution of the latter. In other words, it is possible that there is a causal link, direct or indirect, from the protected characteristic to the target variable. Formally, this means that in those contexts there exists a $y \in D_Y$, and $\mathbf{x} \in D_{\mathbf{X}}$ such that

$$P(\, Y_{a_1} = y \mid \mathbf{x}, a_1\,) - P(\, Y_{a_2} = y \mid \mathbf{x}, a_1\,) > 0 \tag{4}$$

With regard to the causal structure, this means that there is a sequence of edges originating in A toward Y in the graph.

Lastly, (4) states that for every prediction context that is subject to moral norms, there is some way of predicting the target variable in question. This means, there always exists some kind of evidence that would warrant a judgment about the target variable.

To show that (1)-(4) are jointly inconsistent, assume (2), (3), and (4). Imagine, as warranted by accepting (3), a prediction context in which the protected characteristic has some causal relevance to the target variable. By (4), there exists a fair predictive model for the given prediction context. By (2), the fair model satisfies counterfactual fairness. The following theorem implies the negation of (1):

Theorem 1. Every counterfactually fair predictive model necessarily violates equalized odds and predictive parity if the protected characteristic A has a (possibly mediated) causal effect on the target variable Y.

Before presenting the formal proof of Theorem 1, we first specify the framework and the assumptions applied in the proof. Generally, the idea is to construct a graphical proof of the theorem which shows that in any causal structure that incorporates our assumptions, A and $\hat{Y}$ are not d-separated by Y, and A and Y are not d-separated by $\hat{Y}$. This entails that equalized odds and predictive parity are violated in any context represented by a causal structure compatible with the assumptions. We will take the following as premises of our argument:

Premise 1 (Predictive model)**.** We assume that a predictive model is a function $g : D_{\mathbf{X}} \rightarrow D_{\hat{Y}}$ that maps a set of input values $\mathbf{x}$ to a prediction $\hat{y}$. This implies that in a causal structure, the only edge into node $\hat{Y}$ is a directed edge from $\mathbf{X}$.

Premise 2 (Relation between target and input variables)**.** We assume that either of three causal relations holds between the target variable Y and the input variables $\mathbf{X}$ on the basis of which a prediction of Y is to be made:

- there is an edge from $\mathbf{X}$ into Y,
- there is an edge from Y into $\mathbf{X}$, or
- there is an unobserved node with outgoing edges into both, Y and $\mathbf{X}$ (i.e. the node represents a latent common cause).

Premise 3 (Protected characteristic)**.** We assume that the protected characteristic A is such that it is not caused by either the target variable Y, the prediction $\hat{Y}$, or the input features $\mathbf{X}$. This implies that in a causal structure there are no outgoing edges (or chains) from Y, $\hat{Y}$, or $\mathbf{X}$ into A.

Premise 4 (Counterfactual fairness)**.** As argued above, counterfactual fairness implies that there is no outgoing edge (or chain) from A into $\hat{Y}$.

Premise 5 (Effect of protected characteristic on target variable)**.** The protected characteristic having a (possibly mediated) causal effect on the target variable implies that there is a directed edge (or chain thereof) from A into Y. For the sake of simplicity, we can ignore the case in which it is a chain without loss of generality.

The proof strategy we pursue here is a proof by cases. We show that in any possible causal structure representing the relations between Y, $\hat{Y}$, $\mathbf{X}$, and A that satisfies the premises, equalized odds and predictive parity are not satisfied. To this end, we can exploit the projection theorem. Recall that the theorem states that any causal structure with unobserved latent variables can

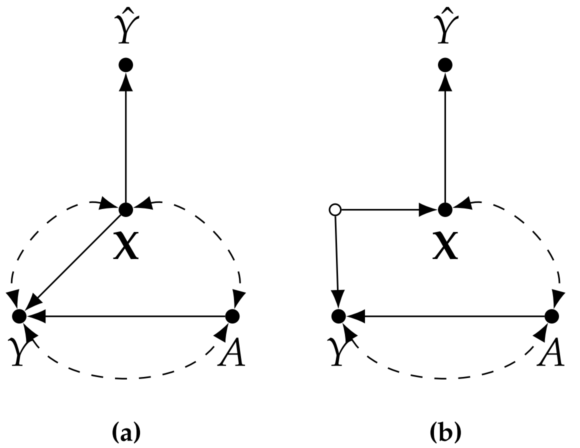

Figure 4: The two possible causal structures involving $Y, \hat{Y}, A$, and **X**.

be represented as a causal structure where the only latent variables are the potential common causes of each pair of observed variables.

This restricts the number of possible causal structures significantly. It leaves us with exactly two classes of structures. Note that we will consider classes of rather than individual causal structures because we can summarize a number of possible causal structures by indicating the possible presence of latent common causes. As we only need to show that in each such class there exists one path which is unblocked for the relevant nodes in order to show that equalized odds and predictive parity are violated, the presence or absence of latent common causes remains irrelevant as long as we find another path that is unblocked. We will represent the possible presence of an unobserved common cause by a dashed bidirectional arrow. Actual but unobserved common causes are depicted by unnamed, hollow circles.

Proof. We consider all the causal structures that represent different possible causal relations among Y, $\hat{Y}$, **X**, and A compatible with premises 1-5. The two resulting classes of graphs are depicted in Figure 4.

Let us first show that equalized odds is violated. Recall that equalized odds is equivalent to $\hat{Y}$ and A being d-separated by Y. This is not the case in either of the two classes of causal structures. In (a), the path $A \rightarrow Y \leftarrow X \rightarrow \hat{Y}$ is not blocked by Y, hence in this class of causal structures, $\hat{Y}$ and A are not d-separated by Y. Whether the potential latent common causes are actually present or not does not matter, since we have already found an unblocked path. It is similar in (b), where the path $A \rightarrow Y \leftarrow \circ \rightarrow X \rightarrow \hat{Y}$ is not blocked by Y, and hence in this class of causal structures $\hat{Y}$ and A are also not d-separated by Y. We conclude that in any possible causal structure compatible with the premises, equalized odds is violated. It follows that equalized odds is not compatible with the premises.

Let us next show that predictive parity is violated as well. Predictive parity is equivalent to Y and A being d-separated by $\hat{Y}$. As in both causal structures, there is, by hypothesis, an edge from A to Y, they cannot be d-separated by $\hat{Y}$. Therefore, in any possible causal structure compatible with the premises, predictive parity is violated as well and is hence itself not compatible with the premises. □

This shows that the four individually plausible propositions about algorithmic fairness introduced at the beginning of this section are jointly inconsistent. Note, that this generalizes to models that provide a score rather than a binary classification in the following way: A model that provides a score can be decomposed into an array of binary classifiers. Say, we consider a scoring model that assigns a score between 1 and 10. Then, for each value, we can derive a binary classifier (e.g. to classify into scores of x or more, and less than x, where x is a number between 2 and 10). Relevant fairness criteria should hold at each level of the score, hence they should hold for each of these binary classifiers. Since we have shown the impossibility for binary classifiers, at each level of the scoring model the impossibility will hold as well.

5 Escaping the impossibility

We will now consider how we can potentially circumvent the impossibility established in the previous section. While propositions (1)-(4) were shown to be jointly inconsistent, it is easy to see that every combination of three of the four propositions is consistent. This means the impossibility can be avoided by giving up or adequately relaxing one of the four propositions. For each of the four propositions, we will consequently explore whether this is a plausible route to take. We will work through the propositions in reverse order, beginning with proposition (4).

5.1 Relaxing proposition (4)

We begin by considering whether it is reasonable to relax the proposition that for every morally relevant prediction context there exists a fair predictive model. In light of the impossibility result, it might be tempting to conclude that in situations in which at best one of the three fairness criteria equalized odds, predictive parity, and counterfactual fairness can be satisfied, there simply exists no (fully) fair predictive model. In these prediction contexts, we have to abstain from making algorithmic predictions.

This, however, has strong counterintuitive consequences. Recall that we defined predictive models as functions from some input features to a prediction of the value of the target variable in question. This is a very general definition that allows representing any systematic procedure of moving from evidence to a prediction of the target variable's value as a predictive model. So, if there is a fair systematic procedure for a human agent to come to a judgment about the target variable's value, then there is a fair predictive model to predict the target variable's value. And, on the other hand, it means that if there is no fair predictive model, there is also no fair systematic way for humans to make such a judgment.

Consequently, if we give up proposition (4), we have to accept that there are some situations in which we have to suspend judgment about a particular proposition on moral grounds, no matter what evidence we have. This seems hard to accept. Intuitively, it seems that for every proposition, there exists some type of evidence that would warrant a judgment on it. It would, for instance, be hard to accept that there are propositions where even in the presence of direct observational evidence the only morally permissible doxastic attitude is to suspend judgment.

Relaxing or giving up proposition (4) hence does not seem to be the most promising way of circumventing the impossibility result. We will next consider whether we can reasonably relax proposition (3) instead.

5.2 Relaxing proposition (3)

Giving up proposition (3) means to accept that there are no morally relevant prediction contexts in which the protected characteristic has some, possibly mediated, causal relevance to the target variable. Different lines of argument can be pursued to defend this claim. First, one could argue that it is conceptually impossible that in a morally relevant prediction context, protected characteristics can be causally relevant to the target variable. One could either do so by arguing that protected characteristics are by definition those that are not causally relevant to a given target variable, or by arguing that when they are, the prediction context is not morally relevant. Secondly, one could argue that empirically this type of case simply never occurs, or is so unlikely to occur that it is not worth considering it morally relevant.

None of the defenses are tenable. Let us consider each in turn. First, we will consider the claim that protected characteristics are by definition irrelevant to a given target variable. Protected characteristics are most commonly defined as socially salient traits that indicate an individual's membership in a specific social group. Social salience of a trait can be understood as the fact that the trait is well perceivable and that the trait plays a role in the structure of social relations (Lippert-Rasmussen, 2014, pp. 30ff). The US law, for instance, considers being of a particular religion, ethnicity, or gender as protected, as well as being disabled or belonging to a certain age group[3]. All of these traits are, to some degree, perceivable - significant age differences are visible, many religious groups are clearly distinguishable by clothing or accessories, as are some physical disabilities. Moreover, they do, to some degree, structure social interaction - some people might act differently towards a woman than they would towards a man, or to someone with a disability than to someone without a disability. So, this definition of protected characteristics seems to indicate that protected characteristics can have causal effects on social interactions. Moreover, the definition does not rule out that protected characteristics have further causal effects. Depending on how the target variable is chosen, it might well be the case that a protected characteristic has a causal effect on it. The claim that by definition protected characteristics are causally inefficacious traits is hence clearly wrong.

Secondly, we will consider the claim that when protected characteristics are causally relevant to the target variable, the prediction context is not morally relevant. In other words, this claim states that in prediction contexts in which there is some causal link from the protected characteristic to the target variable, no moral norms apply. Indeed, there is a family of theories of discrimination according to which the main constitutive component of wrongful discrimination is that people are treated differently on the basis of an irrelevant trait (Halldenius, 2017).

[3] See, e.g., Title VII of the Civil Rights Act of 1964, the Age Discrimination in Employment Act of 1967, the Rehabilitation Act of 1973, and the Americans with Disabilities Act of 1990.

Treating people differently on the basis of irrelevant traits lacks rational justification (see, e.g., Flew, 1993). But acknowledging that in a given situation the protected characteristic is, to some degree, causally relevant to the target variable does certainly not imply that no moral norms apply at all. At best, it implies that the causal relevance of a protected attribute renders a certain, rationally justified, degree of differential treatment morally permissible. It does not imply that it renders arbitrarily differential treatment permissible. So this line of argument fails, too.

Lastly, we will consider the claim that as a contingent matter of empirical fact, these types of cases never occur, or are sufficiently unlikely to occur to be a matter of moral concern. This claim, too, can be easily refuted. To see this, we can consider a number of common, morally relevant examples. One domain that is certainly bound to fairness constraints is hiring. The target variable in a prediction for a hiring decision might be whether an applicant would be productive (in the sense of generating profit for their company) in their role if they were hired. Depending on the role at issue, the productivity might well be influenced by a protected characteristic. Think, for instance, of the role of a salesperson for the Spanish-speaking market - being of Hispanic ethnicity will likely contribute to being productive in this role, simply for the fact that it might explain why someone speaks Spanish fluently. This entails that a Hispanic person who is in fact productive in their role as a Spanish-market salesperson would not have been as productive as they are, had they not been of Hispanic ethnicity. To provide another example, consider the health insurance domain. Imagine an insurer wishes to predict how many claims an applicant will likely make on their health insurance policy. Here, age will certainly have an effect, since age is a factor that influences one's health. Consequently, it might be the case that an older person would not have made as many insurance claims as they actually did, had they been younger. These examples should suffice to refute the claim that cases in which protected characteristics have a causal effect on the target variable are too unlikely to be of moral concern.

So it seems that giving up proposition (3) is no attractive way to circumvent the impossibility result either. Next, we consider whether one or more of the fairness criteria can reasonably be relaxed without allowing for intuitively unfair cases of algorithmic prediction.

5.3 Relaxing proposition (2)

Can we give up or relax counterfactual fairness as a requirement for fair predictive models? To explore this possibility, let us first consider the normative theory that motivates the counterfactual fairness constraint. It is plausible to interpret counterfactual fairness as an anti-discrimination constraint. Discrimination is typically defined as unjustified disadvantageous treatment of one individual as compared with another, that can be (causally) explained by the fact that the former individual possesses a protected characteristic that the latter does not possess (see, e.g., Eidelson, 2015; Lippert-Rasmussen, 2014; Moreau, 2010). In other words, discrimination occurs when a protected characteristic makes an unjustified difference to how someone is treated.

This definition can be applied to predictive models. If an individual unjustifiedly receives a worse prediction than another because the former individual possesses a sensitive characteristic that the latter does not possess, then the prediction exhibits discriminatory bias. And this, conversely, means that in a non-discriminatory prediction, the protected characteristic does not make a difference to the prediction, unless this is justified in some way. Counterfactual fairness formalizes exactly that, except for the proviso that an influence on the prediction is unjustified.

This suggests a straightforward way of relaxing counterfactual fairness, namely to allow for certain conditions under which the protected characteristic can have an influence on the prediction. While there is some disagreement in the philosophical and legal literature about when exactly disadvantageous treatment on the basis of a protected characteristic is unjustified, a widely held view is that such differential treatment is unjustified when the protected characteristic is irrelevant to the goal at hand (Halldenius, 2017; Eidelson, 2015). If, for instance, someone is not granted a loan because of their religion, this constitutes a case of discrimination because religion is irrelevant to whether someone will pay back their loan or not. By the same token, ethnicity and race are irrelevant to an individual's risk of violent crime, as well as gender to hiring decisions for, say, a managerial role. Hence, using these traits in such decisions constitutes discrimination. But there are some situations in which the protected characteristic is relevant, and in which disadvantageous treatment would typically not count as discrimination. For example, when deciding whom to grant a driving license, it seems justified to take into account whether a person is visually impaired because their visual ability is relevant for driving safely. So, we might say that counterfactual fairness is too strict in those cases in which the protected characteristic is relevant to the prediction at issue.

Consequently, we might give up proposition (2) in its universal form. It seems that counterfactual fairness is a necessary requirement for fair predictive models only when the protected attribute is causally irrelevant to the target variable in question. In at least some of the cases in which the protected characteristic is relevant to the target variable, it does not seem reasonable to require that the predictive model satisfy counterfactual fairness in order to be considered fair. A somewhat weaker non-discrimination criterion would suffice. Hence, this provides a promising way of escaping the impossibility.

5.4 Relaxing proposition (1)

Let us now explore whether we can forgo or soften the claim that a model must satisfy either equalized odds or predictive parity to be fair. This means accepting that a model can be fair even without satisfying these conditions. We first argue against predictive parity as a universal fairness criterion, followed by arguments against equalized odds.

Predictive parity means that a predictive model's positive and negative predictive values - the truthfulness of a prediction - are equal for all protected groups. Therefore, predictions are equally informative for different protected groups on average. If the informativeness differs across protected groups, it motivates unequal treatment based on the protected attribute, which is undesirable as it unfairly disadvantages individuals from one protected group.

Consider a company using a hiring algorithm to predict a potential employee's profitability. If the algorithm's positive predictive value is lower for female than male applicants, the employer is incentivized to prefer male applicants predicted as profitable since the probability of this prediction being true is higher. This occurs despite an equal likelihood of male and female applicants being profitable.

If one agrees that situations like these aren't morally problematic, predictive parity isn't universally required for fairness. Hellman (2019, p. 833) argues that while equal predictive values seem desirable, not having them doesn't necessarily constitute discriminatory bias. She questions whether individuals truly have a right to the ”best available decision-making tool,” in which case we can abandon predictive parity as a universal fairness criterion. Moreover, it seems that there can be undeniably fair predictive algorithms which do not satisfy predictive parity, which implies that equalized odds might not be a necessary condition of algorithmic fairness (Eva, 2022).

Could we argue against (or for relaxing) equalized odds as a fairness requirement? From a decision-making perspective, violating equalized odds implies holding different standards for different protected groups, suggesting unfairness. However, in some cases, like when compensating for past injustices or pursuing diversity, it’s morally permissible to have different standards for different protected groups. For example, affirmative action policies for disadvantaged minorities can be justified on the basis of compensatory justice. Similarly, lowering standards for certain groups may be justified to enhance diversity. If we agree that everyone should not be held to the same standards, then these arguments can challenge equalized odds as a universal fairness requirement for predictive models. Similar to Eva’s argument, Hedden (2021) shows that certain intuitively fair prediction algorithms do not satisfy equalized odds, making it an unlikely necessary criterion of fairness.

6 Conclusion

To summarize, we have shown that four intuitively plausible propositions about algorithmic fairness are jointly incompatible. After discussing different ways of escaping this impossibility by giving up one or more of the propositions, we concluded that there are two reasonable ways of doing this. First, it seems plausible to relax counterfactual fairness as a universal requirement of algorithmic fairness and replace it with a weaker criterion - a path implicitly taken by, for instance, Chiappa (2017) and Loftus et al. (2018). Secondly, one could give up predictive parity and either only require equalized odds in specific situations, or replace it with a weaker criterion that captures the intuitive idea of discriminatory bias more adequately.

References

Beigang, F. (2023). Reconciling Algorithmic Fairness Criteria. Philosophy & Public Affairs.

Celis, L. E., Huang, L., Keswani, V., & Vishnoi, N. K. (2019). Classification with fairness constraints: A meta-algorithm with provable guarantees. In Proceedings of the conference on fairness, accountability, and transparency (pp. 319-328).

Chiappa, S. (2019). Path-specific counterfactual fairness. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 7801-7808).

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data, 5(2), 153-163.

Eidelson, B. (2015). Discrimination and disrespect. Oxford University Press.

Eva, B. (2022), Algorithmic Fairness and Base Rate Tracking. Philosophy & Public Affairs., 50: 239-266.

Flew, A. (1993). Three concepts of racism. International social science review, 68(3), 99.

Halldenius, Lena. (2017). Discrimination and Irrelevance.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in neural information processing systems, 29, 3315-3323.

Hedden, B. (2021), On statistical criteria of algorithmic fairness. Philosophy & Public Affairs., 49: 209-231.

Hellman, D (2019). Measuring Algorithmic Fairness. Virginia Public Law and Legal Theory Research Paper No. 2019-39, Virginia Law and Economics Research Paper No. 2019-15, Virginia Law Review, Forthcoming.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. arXiv preprint arXiv:1706.02744.

Kilbertus, N., Ball, P. J., Kusner, M. J., Weller, A., and Silva, R. (2020). The sensitivity of counterfactual fairness to unmeasured confounding. In Uncertainty in Artificial Intelligence (pp. 616-626). PMLR.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores.

Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. (2017). Counterfactual fairness. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 4069–4079.

Lippert-Rasmussen, K. (2014). Born free and equal?: a philosophical inquiry into the nature of discrimination. Oxford University Press.

Loftus, J. R., Russell, C., Kusner, M. J., and Silva, R. (2018). Causal reasoning for algorithmic fairness. arXiv preprint arXiv:1805.05859.

Moreau, S. (2010). What is discrimination?. Philosophy & Public Affairs, 143-179.

Pearl, J. (2009). Causality. Cambridge university press.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., Weinberger, K. Q. (2017). On fairness and calibration. Advances in neural information processing systems, 30.

Russell, C., Kusner, M., Loftus, J., and Silva, R. (2017). When worlds collide: integrating different counterfactual assumptions in fairness. In Advances in Neural Information Processing Systems, pp. 6414-6423. 2017.

Verma, T. (1993). Graphical aspects of causal models. Technical R eport R-191, UCLA.

Wu, Y., Zhang, L., & Wu, X. (2019). Counterfactual Fairness: Unidentification, Bound and Algorithm. In IJCAI (pp. 1438-1444).

Zhang, J., & Spirtes, P. (2016). The three faces of faithfulness. Synthese, 193(4), 1011-1027.

Zhang, J., & Spirtes, P. (2011). Intervention, determinism, and the causal minimality condition. Synthese, 182, 335-347.

Statements and Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

Not applicable.

Competing Interests

The author has no relevant financial or non-financial interests to disclose.

Funding

The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

Authors' contributions

(removed for review)