London School of Economics and Political Science

Doctoral Thesis

Causal Models and Algorithmic Fairness

Author:
Fabian Beigang

Supervisors:
Prof. Christian List
Prof. Richard Bradley
Prof. Liam Kofi Bright

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

at the

Department of Philosophy, Logic and Scientific Method

February 7, 2023

Declaration of Authorship

I, Fabian Beigang, certify that the thesis, titled Causal Models and Algorithmic Fairness, I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others.

I confirm that a version of Chapter 1 has been published in the journal *Minds and Machines* as "On the Advantages of Distinguishing Between Predictive and Allocative Fairness in Algorithmic Decision-Making" (Beigang, 2022).

I declare that this thesis consists of 47,305 words.

Abstract

This thesis aims to clarify a number of conceptual aspects of the debate surrounding algorithmic fairness. The particular focus here is the role of causal modeling in defining criteria of algorithmic fairness. In Chapter 1, I argue that in the discussion of algorithmic fairness, two fundamentally distinct notions of fairness have been conflated. Subsequently, I propose that what is usually taken to be the problem of algorithmic fairness should be divided into two subproblems, the problem of *predictive fairness*, and the problem of *allocative fairness*. At the core of Chapter 2 is the proof of a theorem that establishes that three of the most popular (predictive) fairness criteria are pairwise incompatible. In particular, I show that under certain conditions, a predictive algorithm that satisfies a criterion called *counterfactual fairness* will with logical necessity violate two other popular predictive fairness criteria called *equalized odds* and *predictive parity*. In Chapter 3, a new predictive fairness criterion is developed using a mathematical framework for causal modeling. This fairness criterion, which I call *causal relevance fairness*, is a relaxation of another popular fairness criterion, counterfactual fairness, but turns out to be more closely in line with philosophical theories of discrimination. In Chapter 4, another infamous impossibility result in algorithmic fairness is analyzed through the lens of causality. I argue that by using a causal inference method called *matching*, we can modify the two fairness criteria equalized odds and predictive parity in a way that resolves the impossibility. Lastly, Chapter 5 contains an empirical case study. In it, the fairness of a popular recidivism risk prediction tool is analyzed using the criteria of (predictive) fairness developed earlier.

Acknowledgements

I would, first and foremost, like to thank my supervisors Richard Bradley, Christian List, and Liam Kofi Bright. Without your support and critical feedback, as well as your patience, which was necessary thanks to my frequent changes of mind, this thesis could not have become what it is.

I am grateful to my amazing cohort: Cecily, Sophie, Dmitry, Charles, and Nick. It feels like yesterday that we met for the first time, raising our glasses in front of the *George* to the next four years as PhD students. The time has gone by much quicker than I would ever have expected. I am also grateful to all the LSE PhD students from other cohorts, as well as my crew from King's, and especially Gregor. Thanks also to the Venetians Laura, Lily, and Harry, and to Konrad and Jost for many nightly philosophical discussions in Café Arte and elsewhere.

Moreover, I want to thank everyone at the LSE who has taught me, helped me become a teacher (especially Laurenz Hudetz, Roman Frigg, and Miklós Rédei), or just involved me in conversations on the staircase or in the *White Horse*, many of which sparked relevant and irrelevant philosophical insights. I also want to thank organizers and critical audiences at conferences at the *Universidad Nacional Autónoma de México*, *Copenhagen University*, LSE, and *NeurIPS*; fellow students and teachers at the *Aegina Summer School of Social Cognition*; and the anonymous reviewers for *Minds and Machines* and *Philosophy and Public Affairs*.

My special gratitude goes to my family, who I suspect at times didn't quite understand the usefulness of my philosophical ventures (not always without good reason, as I realize with hindsight), but who nonetheless always supported me without hesitation or doubt.

Finally and most importantly, I want to thank Daniella for the love and support without which my time as a PhD student in London would not have been the same.

Contents

Introduction

"Move fast and break things"

- Mark Zuckerberg

The invention and adoption of new technology are rarely preceded by a comprehensive exploration and evaluation of its likely consequences. While Zuckerberg's famous imperative was intended to inspire technologists to break with conventional approaches and existing paradigms, all too often it was human rights, democratic processes, and laws that ended up being broken instead. The ethical implications of the use of novel technologies are, more often than not, merely an afterthought.

This is, in particular, true for artificial intelligence technology. Especially with the rise of machine learning in the 1990s, the field of artificial intelligence made progress at a dizzying pace, quickly spinning off a wide variety of real-world applications. Phones all of a sudden began to understand our voices, allowing us to transcribe spoken words into text, cars became capable of performing complex maneuvers without human help, and doctors were put in a position where they could get a second opinion on a patient's diagnosis from a virtual colleague. But making mundane tasks more efficient was not the only manner in which artificial intelligence was being utilized. In domains that require making socially sensitive decisions, where stakes for affected individuals are typically high, the use of AI increased rapidly as well.

Examples of this are plentiful. Some American cities, for instance, saw the rise of predictive policing — a machine learning-based technology used to predict in which area the occurrence of a crime is most probable at a given moment. As a consequence, communities in predicted areas experienced great increases in police presence[1]. To mention another example, hiring decisions nowadays commonly rely on applicant tracking systems that scan applicants' resumes in search of patterns that are indicative of professional

[1] See, e.g., Richardson et al. (2019) for a critical analysis of predictive policing practices.

success, discarding those that do not exhibit these patterns. So the list continues.

Social progress is usually hard-won. Several decades had to pass after the first discussion of racial profiling before US police departments started implementing policies to address and prohibit this practice (Harris, 2020). The rise of technologies such as predictive policing threatens to stall or even reverse this progress. Typically, predictive policing software predicts future crime probability on the basis of historical data on the number of past arrests and emergency police calls in a given area. The rate of innocent people being stopped and searched, however, is, at least in the US, much higher for racial minorities than for White citizens (Gelman et al., 2007). In a similar vein, the police are frequently called on the grounds of unjustified, racist suspicions[2]. If these types of historically biased data are used to inform policing strategies, this creates a feedback loop that can perpetuate historical and existing biases. It is not hard to see that this is racial profiling in a technological guise.

Similarly long was the fight women have fought for educational equality and equality in the workplace. While the fight is ongoing, important milestones were achieved when legislation was passed that prohibits hiring decisions based on gender as well as gender-based discrepancies in pay. The use of applicant tracking systems, however, might reintroduce discriminatory practices into the hiring process. Imagine a company decides to use a resume screening tool in the process of hiring for, say, a technical position. Furthermore, imagine the tool is being trained on data about the company's current workforce, which, as is often the case for technical roles, happens to be predominantly male. The algorithm will learn the pattern that people in this technical role are mostly men. It might consequently deem female applicants less qualified for the job solely on the basis of their gender. Far from being a merely hypothetical consideration, the inadvertent and unnoticed introduction of such biases is a real risk. Unsurprisingly, the systematic public engagement with the normative questions raised by such tools only emerged after they were already in use.

[2]Examples of this are occasionally discussed in the media, as was for instance the case with the so-called "Central Park birdwatching incident" (Hackett & Schwarzenbach, 2020).

0.1 Mind the gap

Academia caught on to the topic as well. Researchers from two different branches of academic research began to show interest in the subject: computer scientists on the one hand and philosophers on the other. Computer scientists, for the most part, treated the issue as an optimization problem, trying to define machine learning algorithms that perform maximally well under some mathematical fairness constraint. Philosophers, meanwhile, equipped themselves with moral theories to evaluate the application of AI technology in specific social situations.

While both gave rise to valuable insights, there was a gap between the two approaches that needed filling. Most early computer science papers on the problem of algorithmic fairness motivated their formal fairness constraint using a few cases in which the constraint seemed intuitively plausible, before swiftly turning to the more technical treatment of the problem at stake. Few engaged in depth with moral and political theory to build their algorithms on stable normative grounds. To build machine learning algorithms that can be guaranteed to avoid discriminatory outcomes, however, it seems that a more principled approach would be necessary. Philosophers, in contrast, rarely undertook the (admittedly non-trivial) attempt to translate their conclusions into mathematical language so that they could straightforwardly inform the creation of future machine learning models.

While over the last couple of years this space slowly started to fill[3], it is still in a relatively early stage. Some philosophers have started evaluating fairness constraints proposed by computer scientists; computer scientists have started taking moral theory more seriously and used it to assess the assumptions made in their work. Nonetheless, to this day it seems that most questions have not been answered conclusively.

This is the space in which I want to situate the present work. The central question that motivates the different essays in this thesis is the following: *which constraints can and should we impose on machine learning algorithms in order to prevent discriminatory bias?* In this thesis, I am going to tackle the question from a specific vantage point. In particular, I will assume that (1)

[3] A number of computer science papers engage with political and moral theory, see, e.g. Carey and Wu (2022); Heidari et al. (2019). At the same time, formal fairness constraints are being discussed by a number of philosophers, see, e.g., Eva (2022); Hedden (2021).

discrimination has a causal aspect and that (2) the wrongfulness of discrimination is (in at least some cases) grounded in a failure to treat a person as an individual. I will not argue for these premises here — this has been done by others (see, e.g., Loftus et al., 2018; Eidelson, 2015). Readers not sympathetic to these assumptions risk being disappointed; all others will hopefully find the essays insightful.

0.2 Chapter overview

Each of the chapters is written as a self-contained article. That is, each chapter can be read independently and does typically not presuppose that the reader is familiar with the previous ones. Where ideas from earlier chapters are referenced, this is explicitly stated. Nonetheless, the five chapters are connected by an overarching argumentative structure.

In Chapter 1, I argue that in the discussion of algorithmic fairness, two fundamentally distinct notions of fairness have been conflated. Subsequently, I propose that what is usually taken to be the problem of algorithmic fairness — the problem of finding an adequate formal constraint that, when imposed on predictive algorithms, ensures that they produce fair outcomes — should be divided into two subproblems, the problem of *predictive fairness*, and the problem of *allocative fairness*. This, as I will go on to show, resolves several paradoxes in the discussion of algorithmic fairness. Moreover, it allows for delimiting and focusing the subsequent discussion of predictive algorithmic fairness.

At the core of Chapter 2 is the proof of a theorem that establishes that three of the most popular (predictive) fairness criteria are pairwise incompatible. In particular, I show that under certain conditions, a predictive algorithm that satisfies a criterion called *counterfactual fairness* will with logical necessity violate two other popular predictive fairness criteria called *equalized odds* and *predictive parity*. Different ways to escape the impossibility are subsequently discussed.

In Chapter 3, a new predictive fairness criterion is developed using a mathematical framework for causal modeling. This fairness criterion, which I call *causal relevance fairness*, is a relaxation of the above-mentioned criterion counterfactual fairness. Replacing counterfactual fairness with causal relevance

fairness not only resolves the impossibility established in the previous chapter but is also more closely in line with philosophical theories of discrimination.

In Chapter 4, another infamous impossibility result in algorithmic fairness is analyzed through the lens of causality. I argue that by using a causal inference method called *matching*, we can modify the two fairness criteria equalized odds and predictive parity in a way that resolves the impossibility. Instead of requiring that error rates be equal across protected groups, I argue that we should require that the protected characteristic does not causally influence the error rates of a predictive model. Likewise, instead of requiring that predictive value be equal across protected groups, I argue that we should require that the protected characteristic does not causally influence the predictive value of the predictive model.

Chapter 5 contains an empirical case study. In it, the fairness of a popular recidivism risk prediction tool is analyzed using the criteria of (predictive) fairness developed in Chapters 3 and 4.

0.3 Mathematical formalism

We will end with a brief introduction of the central mathematical concepts that will be of importance in this thesis, as well as the notational conventions I am following.

0.3.1 Random variables and probabilities

A *random variable* describes a state of affairs that depends (to some degree) on a random or random-seeming process. We can, for instance, describe the outcome of the roll of a die using a random variable, since this outcome depends on a physical process that seems random to a human observer. To rigorously define what a random variable is, we need to introduce the notion of a *sample space*. A sample space, denoted by Ω, is a non-empty, finite set of possible states — for example, the state that the number of eyes that comes up in the roll of a die is four. Precisely speaking, a random variable X is a function from the sample space to the real numbers, i.e. $X : \Omega \to \mathbb{R}$. It hence assigns a numerical value to each of the elements of the sample space. We denote random variables by capital letters, e.g. X, their values by lower-case letters, e.g. x, and the domain of a random variable X by D_X. As an example, consider

the following random variable which describes whether the outcome of a die roll results in an even or an odd number of eyes:

$$R = \begin{cases} 1 & \text{if number of eyes } 2, 4, \text{or } 6 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

A vector 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 vector of n variables is defined as the Cartesian product of the domains of the individual variables in the vector, i.e. $D_{X_1} \times D_{X_2} \times ... \times D_{X_n}$. Apart from this notational difference, vectors of variables can be treated just the same as individual variables. This will be particularly useful when speaking about the vector of features in a dataset, which we will typically denote by the compound variable $\mathbf{X}$, rather than the list of each individual feature.

Next, we can define probabilities of random variables. To this end, let us introduce the notion of a *probability space*. A probability space is a triple $(\Omega, \mathcal{F}, P)$. As stated above, Ω is the *sample space* of possible states. Subsets of Ω are called *propositions* or *events*. Events describe what is the case and can be represented as random variables taking specific values. For example, the event $A := \{R = 1\}$ describes that in the roll of a die, the number of eyes is even, i.e. that random variable R takes the value 1.

The *event space* $\mathcal{F}$ is a subset of the powerset $\mathcal{P}(\Omega)$ of the sample space that is closed under Boolean operations, that is, under conjunction $\wedge$, disjunction $\vee$, and negation $\neg$. These three logical operators are understood in set-theoretic terms: $\wedge$ as the intersection of the sets of outcomes that constitute the events in question, $\vee$ as their union, and $\neg$ as the complement of the set constituting the event the operator is applied to.

Finally, P is a *probability function* on $\mathcal{F}$, that is, P assigns real numbers to all events $A, B \in \mathcal{F}$, such that $0 \leq P(A) \leq 1$, $P(\Omega) = 1$, and $P(A \vee B) = P(A) + P(B)$ whenever A and B are mutually incompatible, i.e. the intersection of the sets of outcomes constituting A and B is empty. The probability that a variable X takes value x, $P(X = x)$, will be abbreviated by $P(x)$ when this is unambiguous. The probability of the conjunction of two events $X = x$ and $Y = y$ will occasionally be abbreviated as $P(x, y)$.

The *conditional probability* $P(x \mid y)$ is the probability of an event $X = x$ given

that event $Y = y$ has occurred. For example, the probability that the number of eyes after the roll of a die is 2, given that the number is even. The concept of conditional probability is related to unconditional probability in the following way:

$$P(x \mid y) = \frac{P(x, y)}{P(y)} \tag{2}$$

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 $z \in D_Z$, and $P(y, z) > 0$. We denote conditional independence by $(X \perp\!\!\!\perp Y \mid Z)$.

0.3.2 Causal models and counterfactuals

We next introduce a number of central concepts from the mathematical framework of causal modeling. While a number of different people, such as Lauritzen (1996) and Spirtes et al. (2000), have been central in the development of the mathematical theory of graphical causal models, we will here follow the theory as presented by Pearl (2009). A *causal model* is defined as a triple $(\mathbf{U}, \mathbf{V}, F)$ such that (i) $\mathbf{U}$ is a set $\{U_1, U_2, ..., U_n\}$ of exogenous variables whose values are determined by factors outside the present model, (ii) $\mathbf{V}$ is a set $\{V_1, V_2, ..., V_n\}$ of endogenous 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.

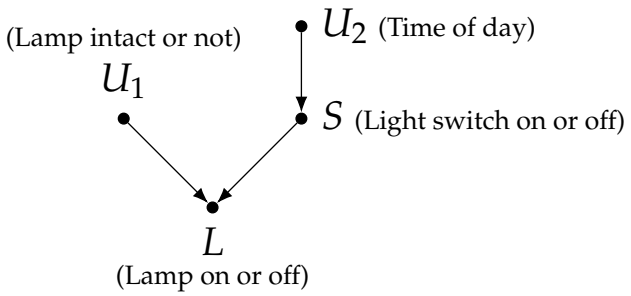


FIGURE 0.1: An example of a causal model (represented by its causal structure).

To illustrate this definition, take for example the causal model $(\mathbf{U}, \mathbf{V}, F)$ depicted in Figure 0.1, 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.

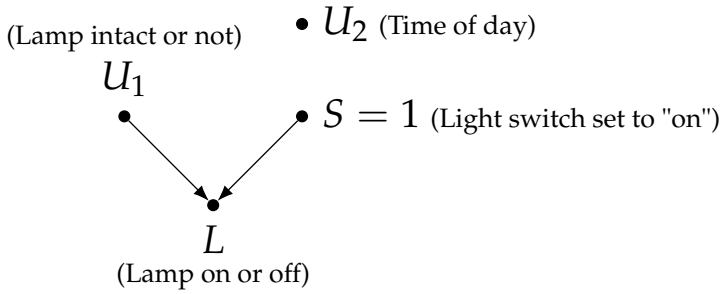


FIGURE 0.2: Submodel of the original causal model. By setting $S = 1$, the initial link from U_2 to S is deleted.

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 $M_{\mathbf{X}=\mathbf{x}} = (\mathbf{U}, \mathbf{V}, F_{\mathbf{X}=\mathbf{x}})$ where $F_{\mathbf{X}=\mathbf{x}} = \{f_i : V_i \notin \mathbf{X}\} \cup \{\mathbf{X} = \mathbf{x}\}$ for a particular realization $\mathbf{X} = \mathbf{x}$ of a set of variables $\mathbf{X} \subseteq \mathbf{V}$. Figure 0.2 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 us 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[4]. Let us assume that initially, we would think that it is 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,

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

$P(U_1 = 1 \mid L = 0) = P^*(U_1 = 1) = 0.8$, 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 0.2) 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. This is due to the fact that the lamp is only on if both, the switch is on and the lamp is intact.

Chapter 1

Two Concepts of Fairness in Algorithmic Decision-Making

1.1 Introduction

Machine learning algorithms generate models that attempt to predict or estimate unobserved properties on the basis of historical data. Commonly, these predictions are used to inform a decision-making process to which the prediction is relevant. Automating decision-making processes in this manner, however, runs the risk of systematizing morally problematic decision patterns. This is in particular a cause for concern when minority groups are the ones who could experience disproportionate negative consequences of algorithmic decision-making, as it could potentially reinforce existing biases and structural inequalities. The recognition of this problem has led to a wide-ranging discussion about algorithmic fairness.

Typically, the problem of algorithmic fairness is presented as the problem of defining a unique formal criterion that guarantees that a given algorithmic decision-making procedure is morally permissible. In this chapter, I argue that this is conceptually misguided and that we should replace the problem thus formulated with two more specific sub-problems. An algorithmic decision system can be conceptualized as operating in two stages: first, it predicts a relevant property, and second, it recommends a decision based (at least partly) on this prediction. It is important to notice that predictions are subject to different normative constraints than decisions. While predictions ought to be unbiased with regard to certain protected characteristics, decision-making based on these predictions ought to ensure that the resulting allocation of goods and opportunities is in line with the relevant principles of distributive

justice. Current approaches to algorithmic fairness have failed to make this distinction. This chapter provides a formal framework to address both ethical issues and argues that this way of conceptualizing them resolves some of the paradoxes present in the discussion of algorithmic fairness.

The chapter is organized as follows. In Section 1.2, I introduce the problem of algorithmic fairness and explain why all of the proposed solutions to it are unsatisfactory. In Section 1.3, I explicate the concept of algorithmic decision systems and argue for a model of algorithmic decision systems which explicitly distinguishes between the predictive and the decision component of such systems. In Section 1.4, I turn to the ethical aspects of algorithmic decision-making, first examining the ethics of public decision-making more generally, before applying the conclusions of this analysis to algorithmic decision systems. In Section 1.5, I provide a formal framework for addressing the sub-problems obtained in the foregoing analysis, which I call the problem of predictive fairness and the problem of allocative fairness. In Section 1.6, I demonstrate how this bifurcation of algorithmic fairness problems can help to resolve several issues and paradoxes that beset the original approach to algorithmic fairness. Lastly, Section 1.7 discusses a number of potential objections to this proposal.

1.2 The problem of algorithmic fairness

The topic of algorithmic fairness gained traction when in 2016 the non-profit newsroom ProPublica published an article that analyzed the results produced by a software tool called COMPAS, which supports bail and sentencing decisions in some US courts by calculating the risk that a defendant will commit crimes in the near future (Angwin et al., 2016). ProPublica's journalists were able to show that for the data they obtained for Broward County in Florida, the false positive rates of COMPAS' predictions were much higher for defendants identified as African American than for those identified as Caucasian, and on the other hand, that false negative rates were much higher for Caucasian than for African American defendants. This means African Americans were much more often falsely accused of committing future crimes, while Caucasians were much more often falsely deemed innocent. They thus concluded that COMPAS is racially biased. A discussion ensued about the question of whether disparities in error rates do indeed indicate bias, or whether there is a more appropriate criterion by which algorithmic decision systems

such as COMPAS could be assessed (Flores et al., 2016). This marked the beginning of the field of *fair machine learning*.

While it is rarely made explicit, the problem addressed in much of the literature on fair machine learning is in fact a demarcation problem. The aim is to provide a precise criterion that constitutes a necessary and sufficient condition for the moral permissibility of an algorithmic decision-making process. This means, a criterion that, given a specific state of the world, allows us to rigorously distinguish algorithmic decision systems that are morally problematic from those that are unproblematic. Such a criterion needs to be formulated in terms of properties that can meaningfully and unambiguously be applied to algorithmic decision systems. Since these systems are, at some level of abstraction, mathematical objects, a fairness criterion needs to be formulated as a mathematical constraint. The problem of algorithmic fairness can, preliminarily, be stated as follows:

The problem of algorithmic fairness. *For which formal criterion ϕ is it the case that the application of algorithmic decision system S in world W is morally permissible if and only if ϕ is satisfied?*

Proposals for ϕ abound. Verma and Rubin (2018) list and describe more than 20 different fairness criteria in their survey paper. Typically, proposals are formulated as conditions involving the following variables: the *input features* $\mathbf{X}$ that are fed into the algorithmic system in order for it to arrive at a decision; the relevant *protected characteristic* A, which typically denotes a trait such as ethnicity, gender, or religion; the *target variable* Y, that is, the relevant property that is being estimated by the algorithm, and which is unknown at the time of application; and lastly, the *outcome* C, which denotes the value the algorithm returns after execution.

To illustrate with an example what these variables could stand for, think of a bank that uses an algorithmic decision system to determine whom to grant a loan to. $\mathbf{X}$ could here represent a vector of variables containing a person's income level (X_1), credit repayment history (X_2), and the like. The variable A could represent the applicant's religion, while Y would most likely stand for whether the applicant would pay back their loan. The variable C represents the categories that the algorithm can assign to an applicant: creditworthy or not creditworthy.

Fairness criterion	Description
Statistical parity	Algorithmic decisions are fair iff[1] the probability of receiving outcome $c \in D_C$ is equal across all protected groups $a_i \in D_A$.
Equalized odds (Hardt et al., 2016)	Algorithmic decisions are fair iff the probability of receiving outcome $c \in D_C$ conditional on being in class $y \in D_Y$ is equal across all protected groups $a_i \in D_A$.
Predictive parity (Cleary, 1966)	Algorithmic decisions are fair iff the probability of being in class $y \in D_Y$ conditional on receiving outcome $c \in D_C$ is equal across all protected groups $a_i \in D_A$.
Fairness through awareness (Dwork et al., 2012)	Algorithmic decisions are fair iff any two individuals i and j with similar input features $\mathbf{x}^{(\mathbf{i})}, \mathbf{x}^{(\mathbf{j})} \in D_{\mathbf{X}}$ receive similar outcomes $c^{(i)}, c^{(j)} \in D_C$.
Counterfactual fairness (Kusner et al., 2017)	Algorithmic decisions are fair iff for each decision it is the case that the outcome $c \in D_C$ would have been the same had the individual's protected characteristic $a_i \in D_A$ been different.

TABLE 1.1: Five of the most popular fairness criteria.

Table 1.1 contains brief descriptions of five of the most widely discussed fairness criteria. For the sake of simplicity, the criteria are presented as prose descriptions instead of mathematical definitions. I will later, where necessary, introduce their precise mathematical formalizations. For the moment, however, the prose descriptions should suffice to provide a conceptual exposition of the criteria.

Despite the initial plausibility of each of these criteria, they come with a number of problems. First, none of the criteria seems to adequately capture the moral permissibility of the application of an algorithmic decision-making process. Often, the criteria are motivated by a handful of hypothetical or actual scenarios of algorithmic decision-making for which they give the right verdict but are not shown to generally guarantee the absence of a particular moral wrong. For each criterion, forceful counterexamples can be constructed which demonstrate that moral permissibility and the satisfaction of

[1]Note that here, "iff" is used to abbreviate "if and only if".

the formal criterion can come apart. A counterexample can be a clearly permissible case of algorithmic decision-making that fails to satisfy the criterion or a clearly impermissible case that does satisfy it. This means that none of the criteria provides both, a necessary and sufficient condition, and, as a matter of fact, many provide neither.

Second, the three so-called "statistical criteria" — *statistical parity*, *equalized odds*, and *predictive parity* — were shown to be pairwise incompatible when the target variable Y is correlated with the protected characteristic A (Kleinberg et al., 2016; Chouldechova, 2017). This means that, in most realistic scenarios, whenever one of the three criteria is satisfied, the other two criteria will be violated. This is an unfortunate result for a set of individually plausible fairness criteria.

Third, some of the criteria are constraints on individual algorithmic decisions (*fairness through awareness*, *counterfactual fairness*), while others are constraints on the population-level patterns of decision outcomes (statistical parity, equalized odds, predictive parity). This raises the question of whether the moral wrongs inherent in certain algorithmic decision procedures are constituted at the individual or at the collective level, and if on both, how they relate to each other.

These three problems cast doubt on the possibility of solving the problem of algorithmic fairness as formulated above. A potential candidate for ϕ would have to (1) guarantee that whenever the application of a given algorithmic decision system in a given context is wrongful, ϕ is violated, and vice versa, that whenever the application is permissible, ϕ is satisfied; (2) be grounded in a moral theory that explains away the mutual incompatibility of statistical parity, equalized odds, and predictive parity, by specifying the conditions under which the more fundamental fairness criterion ϕ implies statistical parity, equalized odds, or predictive parity, respectively, and showing that under given conditions, it implies at most one of the three; and (3) said theory either shows that, fundamentally, the objects of algorithmic fairness are individuals, or that they are groups and explains away intuitions to the contrary. Altogether, this is much to ask of a single fairness criterion.

It seems that the best explanation for the occurrence of the three problems is that there are different types of moral wrongs that can occur in applications of algorithmic decision systems, even though the unified use of the term *algorithmic fairness* suggests the opposite. This, in turn, implies that different

moral norms are relevant to algorithmic decision-making.

The apparent inability to specify universally applicable necessary and sufficient conditions for the absence of moral wrongs in algorithmic decision-making suggests that whether a given moral norm applies might depend on factors outside the mere specification of how the algorithm moves from input data to the resulting output. It might, first, depend on which aspect of the algorithmic decision-making process one is concerned with, and, second, on contextual factors that have a moral bearing on a given decision. If this is right, it entails that it is impossible to define a single, universally applying formal criterion of algorithmic fairness.

Now, if one accepts this explanation, this calls for a principled way of fine-graining the problem of algorithmic fairness, such that for each aspect of algorithmic decision-making that is bound to different normative constraints, we separately look for a (possibly context-relative) formal fairness criterion. This will be the task for the remainder of the chapter.

1.3 Algorithmic decision systems

I begin by specifying what I take *algorithmic decision systems* to be. This will, first, help identify what the relevant normative questions about such systems are, and, secondly, delimit the scope of application of our framework. While these days algorithms are used in a variety of different ways, I am here concerned with one specific, but commonly used type of algorithmic system: a system, deployed in the public or semi-public sphere, that recommends or autonomously takes decisions affecting individuals, where these decisions are made based on predictions from available information about these individuals.

Algorithmic decision systems of this sort are becoming increasingly popular in areas such as credit lending, criminal justice, hiring, and fraud detection. Returning to the example from the previous section, a bank could, for instance, use such a system to make a decision about whether and at what conditions to offer a loan to a loan applicant. The decision would (at least partly) be based on a prediction about the probability that the applicant would if granted, default on the loan. To generate the prediction, the system might, as mentioned above, take as input data information about the applicant's repayment history, education, and employment (see, e.g., Lee and Floridi, 2020).

Many algorithmic decision systems deployed in other fields work similarly.

Before outlining my proposal for a model of algorithmic decision systems, it is necessary to highlight an important conceptual distinction. In the discussion of algorithmic fairness, it is rarely acknowledged that there is a morally relevant difference between algorithmic *predictions* and algorithmic *decisions*. Even though some authors explicitly distinguish between predictions and decisions (see, e.g., Hedden, 2021; Kleinberg et al., 2018; Corbett-Davies et al., 2018), the terms are, especially with regard to their moral aspects, often used interchangeably[2]. A plausible explanation for this is that algorithmic decisions are, as a matter of fact, almost always closely tied to predictions. One might therefore be misled to conclude that there is no need to distinguish between them. This, however, is a faulty line of reasoning. A prediction — broadly understood as an inference of an unknown proposition from a body of evidence — and a decision — understood as a choice of an act from a set of alternatives — differ in which properties can meaningfully be applied to each. While we can, for instance, speak of the *accuracy of a prediction*, it would be a category mistake to speak of the *accuracy of a decision*. By the same token, we can speak of the *expected utility of a decision*, but it would be a category mistake to speak of the *expected utility of a prediction*. The same, I contend, is true for moral properties. Consequently, we need to apply a model of algorithmic decision systems that is sensitive to this distinction in order to consistently discuss ethical aspects of algorithmic decision-making.

Algorithmic decision systems, according to the model proposed here, have two components (see Figure 1.1): a *predictive model* and a *decision function*. The predictive model takes the feature values $\mathbf{x}$ as input, and, given a vector of learned model parameters $\boldsymbol{\theta}$, outputs a probability assignment to the prediction $\hat{y}$. The decision function, on the other hand, takes this probability assignment $\hat{f}_{\boldsymbol{\theta}}(\mathbf{x})$ as an input (and possibly the input values $\mathbf{x}$ as well, as the dashed arrow indicates), and, given a cardinal utility function u over

[2]This is, for instance, evidenced by the following quotes: "It is always possible to construct a trivial *predictor* satisfying equalized odds by making *decisions* independent of X, A, and R" (Hardt et al., 2016, p. 6), "If we think of the *decision* as a *binary prediction* of the outcome, then b_{00} and b_{11} are the values of true negatives and true positives, respectively." (Corbett-Davies et al., 2018, p. 7), "we use the following notations: [...] d: *predicted decision* (category) for the individual (here, *predicted credit score* for an applicant – good or bad)" (Verma and Rubin, 2018, p. 2), and Kusner et al. (2017), who first write "*predictor* $\hat{Y}$ is counterfactually fair if (...)" (p. 3) but then "while $\hat{Y}$ *is the actual decision* of giving the loan" (p. 5) [italics in quotes are my own].

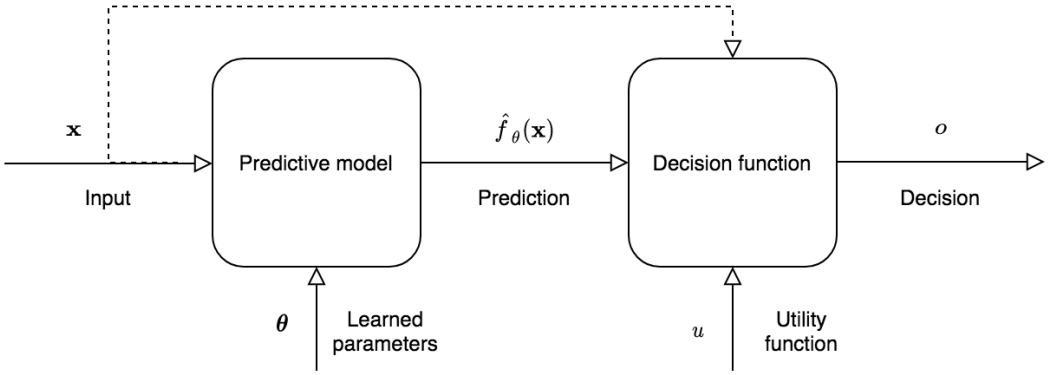


FIGURE 1.1: Schematic model of an algorithmic decision system

different possible outcomes, determines a decision o. This can be made mathematically precise. Let $\mathbf{X}$ be a vector of input variables (with domain $D_{\mathbf{X}}$), $\hat{Y}$ a random variable (with domain $D_{\hat{Y}}$) representing predictions of a target variable Y, and O a non-empty set of decision options. We can then define the notion of an algorithmic decision system as follows:

Definition 1.3.1 (Algorithmic decision system)**.** An *algorithmic decision system* is an ordered pair $S = (\hat{f}_{\boldsymbol{\theta}}, d_u)$, consisting of a *predictive model* $\hat{f}_{\boldsymbol{\theta}} : \mathbf{X} \to [0, 1]$, where $\hat{f}_{\boldsymbol{\theta}}(\mathbf{x})$ is interpreted as the conditional probability of $\hat{y}$ given $\mathbf{x}$, and a *decision function* $d_u : [0, 1] \times D_{\mathbf{x}} \to O$, where $d_u(\hat{f}_{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{x})$ is interpreted as the decision option assigned to the combination of prediction $\hat{f}_{\boldsymbol{\theta}}(\mathbf{x})$ and input $\mathbf{x}$.

A few remarks are in order. The predictive model is defined as a function from the input features $\mathbf{x}$ to a real number in the interval $[0, 1]$. The output of the function represents an estimation of how likely it is that feature values $\mathbf{x}$ make the prediction $\hat{y}$ true, and it coincides with the conditional probability of the prediction $\hat{y}$ given input features $\mathbf{x}$. The reason for introducing a new function symbol $\hat{f}$ is to highlight that we consider the predictive model to be a function of $\mathbf{x}$ for fixed $\hat{y}$. This is conceptually distinct from a probability function $P_{\boldsymbol{\theta}}(\hat{y} \mid \mathbf{x})$, which is a function of $\hat{y}$ for fixed $\mathbf{x}$. This definition of a predictive model is conceptually in line with common practice in machine learning, where models are typically conceptualized as functions of the input values together with a quantification of the uncertainty of a given prediction (see Deisenroth et al., 2020, Ch. 8.2). On our definition, the predictive model encompasses simple models, such as logistic regression, but also more complex ones, such as deep neural networks (see Goodfellow et al., 2016, p. 174).

The decision function $d_u(\cdot)$ is defined in analogy to choice functions in decision and game theory (see, e.g, Bradley, 2017, p. 247; Sen, 1971, p.2, Suzumura, 2009, pp. 20ff). However, it differs from them in that the set of available decision options O is held fixed, as we assume that a given algorithmic decision system will only be applied to one specific type of decision situation. Generally, the decision function is a function of the probabilistic prediction $\hat{f}_{\boldsymbol{\theta}}(\mathbf{x})$, and possibly further information encoded in the input vector $\mathbf{x}$. The output is a decision option o from the set of available decision options O. The decision function can embody principles such as *maximize expected utility* or the *maximin rule*, relative to a fixed cardinal utility function u which assigns numerical utilities to outcomes. As is standard in decision theory, outcomes are defined as combinations of a given decision option o, input values $\mathbf{x}$, and a value y of the target variable (the latter two representing mutually exclusive and jointly exhaustive possible states of the world).

To illustrate this with our previous example, imagine an algorithmic system for lending decisions. The system will proceed as follows: it will take as input data on the applicant's income (x_1) and their repayment history (x_2), on the basis of which the predictive model estimates how probable it is that the applicant defaults on the loan. On the basis of this probabilistic prediction $\hat{f}_{\boldsymbol{\theta}}(\mathbf{x})$, the decision function then outputs a decision, namely whether to grant the applicant the loan or not. Obviously, this is an unrealistically simplified model for making lending decisions, but is helps clarify the concept of an algorithmic decision system.

More formally, the vector of input variables $\mathbf{X} = \{X_1, X_2\}$ contains the two variables X_1 (annual income in thousands of dollars), and X_2 (repayment history), with respective domains

- $D_{X_1} = \mathbb{N}$,
- $D_{X_2} = \{0, 1, 2\}$, where 0 stands for "No late payments", 1 for "Some late payments", and 2 for "Many late payments"

Assume that the predictive model is a logistic regression model, which estimates the probability that a given applicant will default on their loan according to the following equation[3]:

[3]The function $S(\cdot)$ stands for the logistic function. This detail is of no importance to the subsequent arguments in this chapter, and only serves the purpose of illustration.

$$\hat{f}_{\boldsymbol{\theta}}(\mathbf{x}) = S(-0.05x_1 + 1.5x_2) \tag{1.1}$$

Now imagine an applicant, Alice (A), who earns \$35,000 annually, and whose repayment history contains no records of any late payments. That is, her input values are $x_1^{(A)} = 35$ and $x_2^{(A)} = 0$. We can hence calculate her probability of defaulting on the loan according to the predictive model as follows:

$$\hat{f}_{\boldsymbol{\theta}}(\langle 35, 0\rangle) = S(-0.05 * 35 + 1.5 * 0) = 0.148 \tag{1.2}$$

According to the predictive model, Alice has a 14.8% probability of defaulting. In the next step, this prediction is used to inform the decision on whether to grant Alice a loan. To this end, we have to specify the decision function $d_u(\cdot)$. We will assume that there is only one type of loan in terms of credit amount and conditions. The set of decision options O hence contains exactly two possible decisions: to reject an applicant ("Reject"), or to grant them a loan ("Grant"). The decision function could then take the following form:

$$d_u(p, \mathbf{x}) = \begin{cases} \text{Grant} & \text{if } p < 0.3 \\ \text{Reject} & \text{if } p \geq 0.3 \end{cases} \tag{1.3}$$

Recall that by the definition of algorithmic decision systems, the first argument of the decision function is the output of the predictive model, that is, $p = \hat{f}_{\boldsymbol{\theta}}(\mathbf{x})$. This means, a loan is granted if the applicant has less than 30% probability of defaulting. Since in our example, Alice's estimated probability of defaulting is 14.8%, the decision function's output is "Grant". To sum it up, the algorithmic decision system would make the decision to grant her a loan, based on the information that she earns \$35,000 per year and that her repayment history contains no records of late payments.

Note that the second argument of the decision function, $\mathbf{x}$, does not influence the decision in this example (other than through its influence on the probabilistic prediction). We can, however, imagine a different decision function according to which the decision to grant the loan is made only if the probability of a default is less than 30% *and* the applicant earns above \$100,000 annually. The decision function would look as follows:

$$d'_u(p, \mathbf{x}) = \begin{cases} \text{Grant} & \text{if } p < 0.3 \text{ and } x_1 > 100 \\ \text{Reject} & \text{otherwise} \end{cases} \tag{1.4}$$

The reason for including this further condition (that the applicant needs to earn more than $\$100,000$) might be completely unrelated to the epistemic consideration of whether the applicant will default on the loan. It might, for instance, be that the bank is only interested in acquiring high-earners as customers, or some other non-epistemic, business-related reason.

The model introduced in this section is an idealized representation of an algorithmic decision system intended to be general enough to subsume most of the impactful systems that are used in the public and semi-public sphere, and yet specific enough to allow for a sufficiently deep analysis that does justice to the complexity of the ethical questions we attempt to address. I will now turn to the ethical questions that arise when a system of the above form is applied to contexts where it is used for making decisions about individuals.

1.4 Ethical aspects of algorithmic decision-making

In order to examine the relevant ethical aspects of algorithmic decision-making, it will be useful to take a step back and think about the ethical aspects of public decision-making more generally. I will use the term *public decision* in a relatively loose sense, denoting two different types of decisions. First, any act or policy implemented by a public body, such as central and local governments, courts, or police departments, which allocates certain benefits or incurs certain harms on individual persons. Secondly, acts by private actors that involve access to goods which can reasonably be expected to be regulated by the government, such as education, housing, employment, or transport. For the purposes of this analysis, we can ignore the difference between the two.

There are two ethical concerns about decisions in the public sphere, which persist even if we assume that the decisions are taken without objectionable intentions. First, the decisions might be based on biased beliefs[4], which can result in discriminatory decisions. Secondly, the decisions might produce

[4]Note that the term *bias* is here used in the sense of *cognitive bias* (as opposed to behavioral or emotional bias), and refers to a systematic error in forming propositional attitudes.

unjust distributions of benefits and burdens among different groups in society[5]. While discrimination is closely connected to distributive injustice, it is important to distinguish between the two concepts, for they are neither co-extensive, nor is it obvious whether their wrongfulness is grounded in the same fundamental moral principles (see, e.g, Eidelson, 2015, pp. 51-58). I will discuss each in turn.

Discrimination can broadly be understood as wrongfully disadvantaging someone because they belong to a certain salient social group (see, e.g, Moreau, 2010; Eidelson, 2015; Lippert-Rasmussen, 2014). The property of belonging to such a group is what is called a *protected characteristic*, the group constituted by this shared property a *protected group*. Whether an individual is treated disadvantageously is determined relative to some other (actual or hypothetical) individual, who is not a member of that group, and who is, by some standard, suitable for comparison. When a decision-maker takes an individual's social group membership as a reason for intentionally treating them in a disadvantageous way, we speak of *direct* discrimination. However, not all forms of discrimination require an intention to discriminate. When rules and policies are set up in a way such that, despite the absence of any intentions to this effect, being a member of the group results in experiencing certain disadvantages, we speak of *structural* discrimination. Under which conditions exactly disadvantageous treatment of the above form is wrong, and why it is when it is, is widely debated (see, e.g., Alexander, 1992; Eidelson, 2015). I will, for now, set this issue aside.

Unintentional discrimination can come about when decisions are informed by beliefs that are defective in particular ways (see, e.g, Eidelson, 2015; Ch. 5, Lippert-Rasmussen, 2014, pp. 41 ff). This is the case when beliefs are biased, either in that they are inferred from inaccurate generalizations about the properties or behaviors of individuals who belong to a specific social group (i.e. stereotyping), or in that they are grounded in, for instance, a decision-maker's emotional reaction to members of a specific group, rather than in adequate evidence (i.e. prejudice). When decisions in the public sphere are taken, it is hence obligatory to ensure that beliefs which inform the decision at hand are arrived at in an appropriate way.

[5]This is sometimes (e.g. in legal texts) called *indirect discrimination*, even though, as some have argued (see, e.g., Eidelson, 2015, Ch. 1.2), this is a misleading use of the term *discrimination*. For this reason, we will give preference to the term *distributive injustice*.

On the other hand, we can say that a decision contributes to creating or amplifying distributive injustice, when the decisions, which typically allocate certain benefits or burdens, do so in a way that disrespects the distributive principle relevant in a given context. A strict egalitarian principle, for instance, would require that certain goods[6] be distributed equally among different groups, while an equality of opportunity principle would require that economic and educational opportunities be equally distributed among those with the same level of talent and diligence. It is plausible to think that different goods ought to be distributed according to different distributive principles. Which principle applies to the distribution of a given good depends on the social meaning attributed to the good in question (see, e.g, Walzer, 1983).

So, while discrimination refers to the procedure by which a decision is determined, distributive injustice refers to the resulting distribution of goods. To make this distinction more tangible, consider the following two scenarios. In both, we assume that a company is looking to hire a suitable employee. In the first scenario, we assume that in order to decide between two applicants, the employer estimates how much profit an applicant would generate for the company, were they employed. One applicant is female, has a relevant degree from a renowned university, and has a track record of prestigious jobs which evidence her willingness to work hard. The other applicant is male, has no university degree, and has an employment record of rather unimpressive jobs. In estimating their profitability, the employer considers the first applicant's gender to be a point against employing her, as the employer thinks that women are generally not capable of hard work. Nonetheless, due to the male applicant's lack of relevant education and work experience, the female applicant is estimated to be slightly more profitable for the company, and is hence offered the job. While in this scenario, the employer's decision is clearly informed by a stereotypical and hence wrongful belief about women, this does not result in an unjust distribution of employment opportunities.

To contrast the previous example, consider the second scenario (inspired by Eidelson, 2015, p. 53). In this scenario, we assume the employer knows that if an employee has a parent who has herself been a long-term employee of the company, this has a positive effect on the new employee's productivity, and hence the profitability for the company. Assume this is due to the fact

[6]For the sake of brevity, I will use the term *good* to denote any material object or service that is assumed to have a (positive or negative) utility to individual persons. This includes what is sometimes called *economic bads* (see, e.g, Varian, 2006, p. 41).

that having a parent who is a senior employee facilitates certain things for new employees — it might, for instance, allow them to get acquainted with the processes within the company more quickly, or to get to know people in important roles at a more personal level, and so on. For this reason, the employer prefers, all else being equal, applicants who have a parent who has been working for their company. Now assume further that non-Christian applicants are less likely to have a parent who has been working in the employer's company — possibly because many of the non-Christian applicants happen to be children of recent immigrants. This means that the employer's hiring policy disproportionately denies non-Christians the opportunity to work for the company, even if they are, on average, equally talented and diligent. Hence, this constitutes a case of distributive injustice against the group of non-Christians, despite the fact that the decision is not informed by biased beliefs about non-Christians.

In both scenarios, we can criticize the employer's decision-making procedure as wrongful. However, we do so on different grounds. In the first case we can criticize the decision as being made on the basis of a belief which is, in a morally relevant way, defective. We cannot, however, criticize the outcome of the decision. In the second case, we can criticize the decision as producing an unfair distribution of economic opportunities among different social groups. We cannot, however, criticize that the employer's belief about the profitability of potential employees is defective, since, by assumption, the belief is true.

Let us now transfer the above analysis to algorithmic decision systems. When algorithmic decision systems are deployed in order to make or recommend decisions in the public sphere, they are bound to the same normative constraints as public decisions taken by human decision-makers. Hence, it is necessary to ensure that they do not make decisions on the basis of biased beliefs, and that they do not make decisions that allocate goods in a way that violates the relevant distributive principle.

While algorithmic decision systems do not have beliefs in any literal sense, they do possess representations of real world properties. Those are encoded in the input features $\mathbf{x}$, and the estimation of the probability that the unobserved property y is present. Consequently, the first normative constraint on algorithmic decision systems is that the probabilistic estimation of y on the basis of $\mathbf{x}$ must not be biased. This, unsurprisingly, is a constraint on the first

component of an algorithmic decision system, the predictive model.

When an algorithmic decision system makes a decision that allocates goods, allocating these goods according to the probabilistic prediction of property y and background information $\mathbf{x}$ must be compatible with the relevant distributive principle. This means that, for a given decision, the variable Y has to be chosen such that distributing a good according to it (possibly together with some of the input variables in $\mathbf{X}$) is permissible in the light of the principle. Think, for instance, of the second scenario discussed above. Assume we deem an equality of opportunity principle the right principle for allocating job opportunities. This principle demands that everyone with the same talent and diligence should have the same chance to be offered a given job. If we accept this principle, then in the scenario above, hiring decisions cannot (merely) be based on a prediction of the profitability of an applicant, because profitability is influenced by factors beyond talent and diligence — namely having a parent who also works for the employer's company. Put differently, in the above case predicted profitability alone does not provide a permissible reason for a hiring decision. So, the second normative constraint on algorithmic decision systems is that a decision must be determined on the basis of properties (or predictions thereof) which are permissible for a given allocation of goods. This, on the other hand, is a constraint on the second component of an algorithmic decision system, the decision function.

We can conclude by summarizing that there are two aspects of algorithmic fairness, which are both necessary but individually insufficient for guaranteeing that the application of an algorithmic decision system is morally permissible. Consequently, we are confronted with two problems of algorithmic fairness: (1) finding a constraint on predictive models that ensures that probabilistic predictions are generated in an unbiased way, and (2) finding a constraint on decision functions that ensures that decisions about the allocation of a given good are based on information and estimations of adequate properties. These two problems will be made more precise in the next section.

1.5 Two concepts of algorithmic fairness: a formal framework

The above analysis suggests a way to replace the *problem of algorithmic fairness* presented in Section 1.2 with two separate subproblems. Rather than finding

a single formal criterion which guarantees that, if satisfied, the application of a given algorithmic decision system is morally permissible, we should turn our attention to finding two different criteria: one criterion that guarantees the absence of biased predictions, and another criterion that guarantees that decisions are made in a way such that no unjust distribution of goods results. While it seemed infeasible to find a single criterion that guarantees the intuitive permissibility of algorithmic decision systems, the bifurcation of the problem into two sub-problems aligns well with moral theory and allows us to explain away seeming contradictions.

Let us begin with the problem of finding a constraint on an algorithmic decision system's predictive model that guarantees the absence of discriminatory bias. We say that predictive models are biased when their predictions exhibit specific patterns of errors. In other words, biased predictions deviate from the truth in systematic ways. To determine whether a predictive model is biased, it is thus necessary to not only take the probabilistic predictions themselves into consideration, but moreover what is actually the case in (some relevant aspect of) the world. The constraint on the algorithmic decision system must hence be formulated relative to a specification of the relevant aspects of the world. How informationally rich this specification needs to be depends on how exactly one defines the notion of bias. In order to make the present framework compatible with as many different approaches as possible, I will here not take a stance on which technical notion of bias is to be chosen. To provide two examples, however, note that the world could simply be specified as the set of all the (relevant) true propositions[7], or as a causal model which not only specifies what is true, but which also represents the relevant underlying mechanisms and processes[8]. We can now formulate the first subproblem as follows:

The problem of predictive fairness. *For which formal criterion ϕ is it the case that the predictive model $\hat{f}_{\boldsymbol{\theta}}(\cdot)$ is unbiased if and only if $\hat{f}_{\boldsymbol{\theta}}(\cdot)$ satisfies ϕ relative to world W and protected characteristic A?*

Next, consider the problem of allocative algorithmic fairness. The task here is to find a constraint that ensures that the distribution of goods resulting from the application of the algorithmic decision system is in line with the relevant

[7] Examples of criteria that take into account whether certain propositions are true are *equalized odds* and *predictive parity*.

[8] Examples of criteria for the absence of bias that take the relevant causal mechanisms into account are *counterfactual fairness* and *no-proxy discrimination* (Kilbertus et al., 2017).

distributive principle. More technically speaking, this means that we want to constrain which properties are allowed or need to be correlated with receiving the specific good. For example, a strict gender parity principle would require that there be no correlation between an individual's gender and receiving the good in question. Applied to, say, a hiring context, this would ensure that the proportion of female applicants offered a job is equal to the overall proportion of female applicants. An equality of opportunity principle, on the other hand, would require that receiving the good be perfectly correlated with talent and diligence, even if this means that receiving the good is to some degree correlated with a protected characteristic. Applied to the hiring example, equality of opportunity would ensure that the applicants which score the highest on features such as education, professional experience, or the performance on relevant tests, are the ones who are offered the job.

As argued above, we can assume that how a good ought to be distributed depends on the type of good in question. One might argue, for example, that certain government jobs should be allocated in accordance with a gender parity principle to ensure equal representation of men and women in policy making. On the other hand, this principle certainly doesn't hold for criminal justice decisions: whether someone receives a jail sentence should only depend on factors such as the seriousness of the crime and the degree to which a defendant can be held liable for the crime. Hence, whether the decision function of an algorithmic decision system is fair can only be determined relative to the specific good in question.

In order to define a formal framework for determining whether a decision function produces unfair allocations of a given good G, we hence have to specify two sets of properties. The first, $\mathbf{I}_G$, denotes the set of properties for which it is *impermissible* to be correlated with the decision outcome d_u. The second, $\mathbf{O}_G$, denotes the set of properties for which it is *obligatory* that they be correlated with the decision outcome d_u. Since an impermissible property cannot be obligatory, we can assume that the set of impermissible properties and the set of obligatory properties are disjoint, i.e. $\mathbf{I}_G \cap \mathbf{O}_G = \emptyset$. We can now formulate the second subproblem of algorithmic fairness as follows[9]:

[9]Note that for full generality, we would need to frame the problem of allocative fairness in terms of partial/conditional correlations, rather than unconditional correlations. This would mean that $\mathbf{I}_G$ and $\mathbf{O}_G$ would contain tuples of properties rather than individual properties. This would allow us to express conditional requirements, as for instance that it is impermissible that the decision outcome d_u is correlated with V_i given U_i. Formally, this requirement

The problem of allocative fairness. *For which pair of property sets* $(\mathbf{I}_G, \mathbf{O}_G)$ *is it the case that the decision function* $d_u(\cdot)$ *is allocatively fair with regards to a given good* G *if and only if, under the assumption of perfectly accurate predictions, the outcomes of* $d_u(\cdot)$ *are sufficiently correlated with all variables* $V_i \in \mathbf{O}_G$*, and are sufficiently uncorrelated with all variables* $V_j \in \mathbf{I}_G$*?*

Operationalizing any specific definition of allocative fairness requires making precise what is meant by saying that two variables are sufficiently correlated or uncorrelated. One natural way of doing this would be to define two variables to be sufficiently correlated whenever the absolute value of some correlation coefficient, such as Pearson's correlation coefficient (see, e.g, Lee Rodgers and Nicewander, 1988), is above a certain threshold. Conversely, we could define two variables to be sufficiently uncorrelated whenever the absolute value of their correlation coefficient is below a certain threshold. There are, however, many different ways in which these notions could be explicated, and I will here leave open how the question which of them is the most adequate one is to be answered.

Note that we always evaluate predictive and allocative fairness relative to a specific protected characteristic. This means that we decide on the protected characteristic relative to which we want to evaluate a given algorithmic decision system beforehand, and then check whether the chosen criteria of predictive and allocative fairness hold for this specific characteristic. The framework presented here is agnostic about what counts as a protected characteristic and how to choose which protected characteristics are of special importance in a given situation. Those are complex questions in their own right, in particular in light of the fact that sometimes we care about intersectional characteristics, like for instance being a woman of a particular ethnicity. While the present framework cannot provide answers to these questions, it is general enough to be compatible with different theories about protected characteristics.

Let us now illustrate the distinction between the two problems with two hypothetical examples. As a first example, consider an algorithmic decision system which estimates how probable it is that a given, previously criminal individual, will commit another crime within some specified time frame in the future. On the basis of this prediction, the system then recommends

would be expressed by stating that $(V_i, U_i) \in \mathbf{I}_G$. For the sake of conceptual clarity, however, and due to the fact that most distributive principles can be expressed as constraints on unconditional correlations, we will restrict the discussion to the latter.

Ethnicity	Area	Crime	ADS	
			$\hat{f}_{\theta}(\cdot)$	$d_u(\cdot)$
White	1	No	0.2	✗
White	1	No	0.2	✗
White	1	Yes	0.2	✗
White	2	Yes	0.8	✓
Non-White	1	No	0.2	✗
Non-White	2	Yes	0.8	✓
Non-White	2	Yes	0.8	✓
Non-White	2	No	0.8	✓

TABLE 1.2: This table contains information on the protected characteristic (Ethnicity), the input variable (Area), the "ground truth" of the target value (Crime), and the predictions and decision recommendations of the algorithmic decision system.

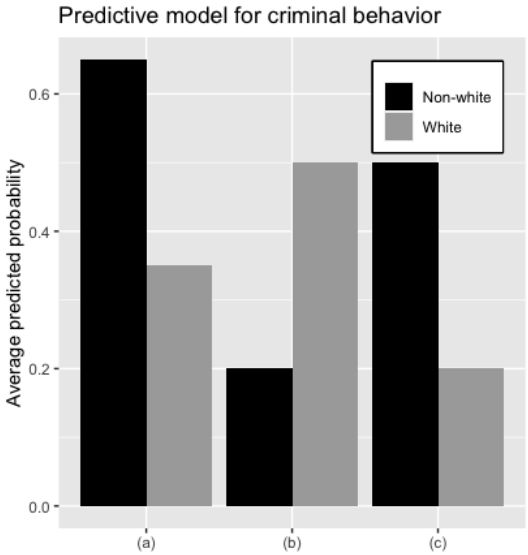


FIGURE 1.2: Average probabilistic predictions of (a) criminal behaviour, (b) absence of criminal behaviour among individuals who actually go on to commit a crime, (c) criminal behaviour among individuals who actually do not go on to commit a crime.

whether to subject the individual to increased monitoring measures.

In this example, we can assume that the only input the *predictive model* of the algorithmic decision system takes is information about which neighborhood a given individual lives in. We can further assume that it is known that there is a correlation between living in a given neighborhood and exhibiting criminal behavior, so that this choice of input data has, at least at first glance, some plausibility. The predictive model assigns a 0.2 probability of criminal behavior if the individual lives in neighborhood 1, and 0.8 probability if the individual lives in neighborhood 2. The *decision function* of the system is equally simple: it outputs the decision to increase monitoring ("✓") of an individual whenever the prediction is greater than 0.7, and the decision to stay with a regular level of monitoring ("✗") otherwise. While this is certainly an unrealistically simplistic algorithmic decision system, its simplicity allows us to focus on those aspects that I aim to illustrate without getting caught up in technical details.

Table 1.2 contains information on eight fictitious individuals for whom predictions of criminal behaviour were generated. In particular, we have information on each individual's ethnicity, which is the protected characteristic

relative to which we will assess the fairness of the system; on the neighborhood an individual lives in, which is the (only) input feature to the predictive model in this example; and on whether an individual actually exhibited criminal behavior, which is the target variable for which the predictive model estimates a probability. Note that, from the perspective of the predictive model, the value of the target variable is not known. Additionally, the table describes the probabilistic prediction $\hat{f}_{\boldsymbol{\theta}}(\cdot)$ of the algorithmic decision system and the decision output generated by the decision function $d_u(\cdot)$.

In order to assess whether the algorithmic decision system is fair according to our proposed framework, we have to fill in the variables in the two fairness schemata to obtain concrete fairness criteria. Begin with predictive fairness. The protected characteristic relative to which we evaluate whether the predictive model is biased is *ethnicity* (denoted by variable A). The relevant aspect of the world W, relative to which we check whether the predictions make systematic errors, is whether an individual does in fact commit a crime (denoted by variable Y). As the criterion that ensures that the predictive model is not biased with regards to ethnicity, we choose *equalized odds* (Hardt et al., 2016). This means we require that the average predicted probability that an individual will not commit a crime, given she does in fact commit a crime (and, likewise, the average predicted probability that an individual will commit a crime, given that she does not, in fact, commit a crime) be equal among White and non-White individuals. These metrics can be considered the analogues of the false positive and the false negative rates for probabilistic predictions. Hence, we substitute ϕ in the schema with the condition that for all $\hat{y} \in D_{\hat{Y}}, y \in D_Y$, and $a_1, a_2 \in D_A$:

$$P(\hat{y} \mid y, a_1) = P(\hat{y} \mid y, a_2) \tag{1.5}$$

Having specified a concrete predictive fairness criterion, we can now assess whether the predictive model is biased. A quick look at the dataset in Table 1.2 shows that of the four White individuals, two turned out to commit criminal offenses (row 3 and 4), as did two of the four non-White individuals (rows 6 and 7). This means, the prevalence of criminal behavior is equal among the two groups according to our data. If, however, we look at the summary statistics of the predictive model $\hat{f}_{\boldsymbol{\theta}}(\cdot)$, we can see that on average, the non-White individuals received a probabilistic prediction of crime above 0.6, while the White individuals received on average predictions below 0.4

(Figure 1.2(a)). More specifically, we note that the average predicted probability of absence of criminal behavior among those who did in fact commit a crime is much higher for White individuals than for non-White individuals (Figure 1.2(b)). At the same time, the average predicted probability of criminal behavior among those who do in fact not commit a crime is much higher for non-White individuals than for White individuals (Figure 1.2(c)). Intuitively speaking, this means that for White individuals it is much more probable to be deemed innocent while actually going on to commit a crime, whereas for non-White individuals it is much more probable to be deemed criminal while actually being innocent. This clearly violates the fairness criterion specified above — the predictive model is biased on our definition.

Next, we have to choose a criterion according to which we can examine whether the decision function allocates the good in question in a fair way. The "good" at issue is in fact an "economic bad" — a burden that comes with negative utility for the individual — namely, to be subjected to an increased level of monitoring. Presumably, no egalitarian principles apply here — it seems implausible to think that increased monitoring should necessarily be equal among different ethnicities, men and women, and so on. Rather, it seems, a burden such as increased monitoring should be allocated according to a desert-based principle — in other words, the individuals subjected to increased levels of monitoring should be those who deserve so due to their inclination towards criminal behavior. In accordance with our formal framework, this can be formalized as the requirement that $Crime \in \mathbf{O}_{Monitoring}$. This means that, assuming that the predictive model generates perfectly accurate predictions, the target variable *Crime* (that is, whether an individual did actually exhibit criminal behavior) ought to be correlated with the outcome of the decision function $d_u(\cdot)$. Apart from this, there are no further constraints.

If the predictions were perfectly accurate, then every individual who will in fact go on to commit a crime would have received a predicted probability of 1, and every individual who will not would have received a predicted probability of 0. Since the decision rule $d_u(\cdot)$ recommends increased monitoring for those individuals who have a predicted probability of criminal behavior above 0.7, every criminal would be subjected to increased monitoring, whereas no innocent individual would. Hence, the decision outcomes would be perfectly correlated with criminal behavior, and we can conclude

that the decision function satisfies our criterion of allocative fairness.[10]

To summarize, the algorithmic decision system in this example produces unfair decisions. As our analysis has shown, this is due to a biased predictive model. Hence, in this case the predictive model should be adjusted so as to not produce such biased predictions. There is, however, no reason to change the decision function.

Let us now turn to the second example. Here, we are considering university admission decisions. The predictive model of the algorithmic decision system estimates how likely it is that a given individual would be successful at the university they apply to, where success will be defined as achieving a grade average above a specific threshold. The decision function then recommends an admission decision on the basis of this estimation. Similar to the previous example, the predictive model $\hat{f}_{\boldsymbol{\theta}}(\cdot)$ assigns a predicted probability of *university success* of 0.8 whenever the individual had a high school grade of A or B, and 0.4 whenever the high school grades were below that. The decision function $d_u(\cdot)$ recommends the decision to admit an individual whenever the predicted probability of success if greater than 0.7 ("✓"). Table 1.3 depicts a fictitious dataset with information on whether an individual has dyslexia (the protected characteristic in this example), their high school grades (the input data), whether they actually turned out to be successful at university (the target variable), and what the algorithmic decision system would have predicted and decided for each individual.

Examining Figure 1.3, we notice that individuals without dyslexia have, on average, a higher predicted probability of academic success. Yet, the average predicted probability of not being successful, given that the student would actually have been successful, as well as the average predicted probability of being successful, given that the student would actually not succeed, are equal across the two groups. By our notion of predictive fairness (*equalized odds*), the predictive model would hence count as fair. Yet, if we think about how education opportunities should be distributed, we might want to say that a learning difficulty such as dyslexia should not affect one's chances of being accepted to a university programme. Dyslexic students, on this view, should have the same overall admission rate as students without dyslexia.

[10] We could, for instance, use the Pearson correlation coefficient to measure the degree of correlation. As we here have a perfect correlation between criminal behaviour and increased monitoring, the coefficient would take the maximum value $+1$. This would trivially be considered a sufficiently strong correlation.

Dys-lexia	High school grades	Uni-versity success	ADS	
			$\hat{f}_{\boldsymbol{\theta}}(\cdot)$	$d_u(\cdot)$
No	A	Yes	0.8	✓
No	B	Yes	0.8	✓
No	C	No	0.4	✗
No	D	No	0.4	✗
Yes	B	Yes	0.8	✓
Yes	C	No	0.4	✗
Yes	D	No	0.4	✗
Yes	E	No	0.4	✗

TABLE 1.3: This table contains information on the protected characteristic (Dyslexia), the input variable (High school grades), the "ground truth" of the target value (University success), and the predictions and decision recommendations of the algorithmic decision system.

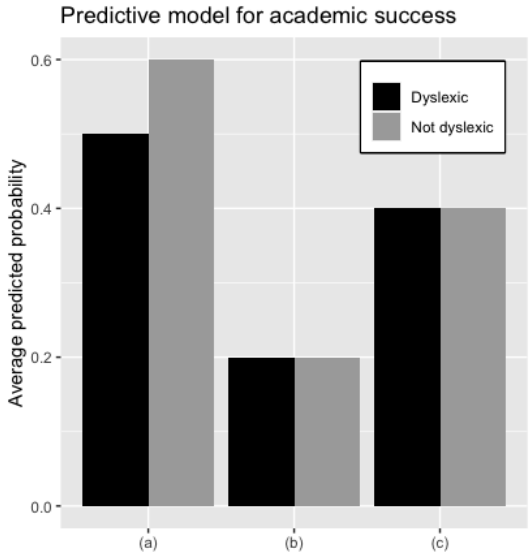


FIGURE 1.3: Average probabilistic predictions of (a) university success, (b) university failure among individuals who would actually succeed, (c) university success among individuals who actually would actually not succeed at university.

More precisely, the decision outcomes should not be correlated with the variable *Dyslexia*, i.e. $Dyslexia \in \mathbf{I}_{Admission}$. This allocative fairness criterion is clearly violated by the decision function. Out of four students with dyslexia, only one is admitted to the university, as compared to two out of the four students without dyslexia. This means, the decision outcomes are not statistically independent of the variable *Dyslexia* [11]. Moreover, this would still be the case if the probabilistic predictions were perfectly certain and accurate.

What these two examples show is that two intuitively unfair algorithmic decision systems can suffer from fundamentally different flaws, and hence require different approaches to rectify these flaws. In the first example, the predictive model is biased against non-White individuals, and consequently, the appropriate response to this assessment would be to put effort into increasing the predictive accuracy for data points of non-White individuals. Changing the decision function would not help in any way, and would presumably lead to further unforeseen and undesirable consequences. In contrast, in the

[11] More precisely speaking, the absolute value of the Pearson correlation coefficient of admission and dyslexia is 0.26, while we would expect it to be 0 (or close to 0) in a fair algorithmic decision system.

second example, the predictive model is not biased against dyslexic individuals. Yet, basing decisions solely on the predictions of success at a given university creates a distribution of admissions which conflicts with our principle of allocative fairness. This, however, calls for a very different approach than the first example. Here, increasing predictive accuracy would not help to make the system fair. What would potentially help, in contrast, would be to change the decision function such that it takes not only an individual's predicted probability of success into account, but moreover whether the individual has a learning difficulty. A fair decision function could, for instance, implement different cut-off thresholds for individuals with dyslexia and for individuals without learning disorder[12]. This would counteract the unequal initial conditions for individuals with dyslexia and those without.

1.6 The three issues revisited

In Section 1.2, I introduced three problematic issues for current approaches to algorithmic fairness, namely that (1) none of them adequately capture the notion of moral permissibility of deploying an algorithmic decision system, (2) three of them are, from a mathematical point of view, pairwise incompatible whenever Y is correlated with A, and (3) it is unclear whether algorithmic fairness is an individual-level or a collective-level property. In this section I will discuss to which extent the proposed bifurcation of the problem of algorithmic fairness into the two sub-problems of predictive and allocative fairness allows us to resolve or explain away these issues.

1.6.1 Inadequacy of current criteria

Recall that by stating that none of the proposed constraints capture the notion of moral permissibility adequately, it was meant that none of the constraints provides a necessary and sufficient condition for the moral permissibility of a given algorithmic decision system. More precisely, this means that for each of the criteria, we can provide a counterexample of an algorithmic decision system that is either morally permissible but does not satisfy the constraint in question, or is not morally permissible but satisfies the constraint. In the case of some of the proposed fairness criteria, both types of counterexamples can be constructed.

[12] The idea of implementing different cut-off thresholds for different protected groups was explored in more detail by Kleinberg et al. (2018).

While the bifurcation of fairness notions will certainly not be able to completely resolve the issue that many fairness constraints face plausible counterexamples, it may provide an explanation for why, despite the existence of a multitude of plausible fairness constraints, it seems that it is relatively easy to construct tenacious counterexamples to each of them. This is so because the constraints were implicitly intended to simultaneously play two distinct and incompatible roles, namely to act as a fairness constraint on predictions and as a fairness constraint on decisions. As argued above, predictions and decisions are subject to different normative constraints. Consequently, applying a fairness constraint that is plausible for predictions to decisions, or vice versa, will in most cases conflict with our moral judgment. This, in turn, means that we have a simple recipe for constructing counterexamples. We only need to figure out which realm (predictions or decisions) a given constraint is intuitively plausible for and then construct an example in which we apply the constraint to the other realm.

A second potential explanation can be made with regards to allocative fairness. As argued above, allocative algorithmic fairness constraints should be indexed by goods, since for different goods different distributive principles hold. This means that an allocative fairness constraint that is plausible for one particular good might not be plausible for another, different good. So, a second recipe for constructing counterexamples to fairness constraints is to apply an allocative fairness constraint to an algorithmic decision system which is used for a good that is subject to a different distributive principle than the one corresponding to the fairness constraint.

My claim is that if the scope of a given fairness constraint is being restricted according to the bifurcation of fairness problems proposed above, many counterexamples will lose their argumentative force. It would be tedious to check for every alleged counterexample whether the above pair of explanations can in fact rebut it, and it would be impossible to show more generally that we can do so for every conceivable counterexample. To illustrate the point, however, we can look at a number of prominent counterexamples in order to see whether the explanations are any good.

Let us first consider the fairness criterion *statistical parity*, which requires that the members of different protected groups be equally likely to receive a certain algorithmic outcome, or, in other words, that the algorithmic outcome be statistically independent of the protected characteristic. A typical case in

which this seems to be a reasonable normative constraint is university admissions, where the protected characteristic in question is gender (see, e.g., Bickel et al., 1975): arguably, it would be morally problematic if the probability of being admitted to the relevant university were significantly lower for a randomly picked female applicant than for a randomly picked male applicant.

Statistical parity was criticized as a formal algorithmic fairness criterion in a number of ways. Hardt et al. (2016), for instance, argue that statistical parity is too strict a requirement for fairness. Their argument is based on the observation that whenever there is a correlation between the target variable and the protected characteristic, a perfect predictor, that is, a predictive model which predicts the target variable with perfect accuracy, will not satisfy statistical parity. If one assumes that perfectly accurate predictions are always morally permissible, it follows that statistical parity is not a necessary condition for fairness. The example they mention to illustrate this argument is credit lending. Imagine a predictive model which predicts with perfect accuracy whether an applicant will default on a loan or not. It would not be reasonable, they contend, to consider this model discriminatory and hence unfair, even if the proportion of positive predictions were different for loan applicants of different ethnicities.

Another counterexample was put forward by Corbett-Davies et al. (2017), who argue that applying statistical parity to decision-making in an area such as criminal justice is not morally optimal. In their example, which is based on the COMPAS dataset[13], statistical parity is applied to an algorithmic decision system for pretrial release decisions, that is, for decisions as to whether to detain or release a defendant for the time leading up to the trial. Corbett-Davies et al. compare two different decision functions: one that maximizes expected social utility without any fairness constraints, and one that maximizes expected social utility subject to statistical parity with regards to ethnicity. In this scenario, it is assumed that positive utility is assigned to detaining defendants who would otherwise commit violent crimes, while negative utility is assigned to the social and economic costs incurred through detention. It can be shown that in this specific case a decision function which satisfies statistical parity yields a lower expected overall utility: such a function can be expected to lead to a higher number of violent crimes committed by released

[13]The dataset can be found here: https://github.com/propublica/compas-analysis (Accessed: 21 April 2022).

defendants as well as a higher rate of detentions of individuals who would not have committed violent crimes had they been released. If we assume, as Corbett-Davies et al. seem to do, that in the domain of criminal justice the expected social utility of a decision has a bearing on its moral evaluation, it follows that ensuring statistical parity alone is not sufficient for the moral permissibility of an algorithmic decision system.

We can make sense of these two counterexamples with our conceptual distinction between predictive and allocative algorithmic fairness. Statistical parity clearly only makes sense as an allocative fairness criterion. It only takes into account whether the protected characteristic is correlated with the algorithmic outcome. This would not be plausible for a constraint on predictions. As argued above, we have to check whether predictions deviate from the truth in systematic ways in order to determine whether they are biased. To do so, we obviously have to take information about the relevant aspect of the world (that is, at least the individual truth values of the target variable) into account. Statistical parity does not do this — it merely considers whether outcomes are uniformly distributed across protected groups. It is hence misguided to interpret statistical parity as a criterion of predictive fairness. But this is exactly what Hardt et al. did: they argued against statistical parity on grounds that it possibly prohibits the perfect predictor. This, however, is wrong — statistical parity can at best constrain how to move from perfectly accurate predictions to decisions. So, the first counterexample loses its force when viewed through the lens of our conceptual distinction.

In order to address the second counterexample, we have to keep in mind that criteria of allocative fairness are indexed by goods. Statistical parity can be represented as the pair of property sets $(\mathbf{I}_G, \mathbf{O}_G) = (\{A\}, \emptyset)$. This means that for a certain class of goods G, it is impermissible that the decision outcomes are correlated with the protected characteristic A, but that there are no requirements as to which variables the outcomes *must* be correlated with. The class of goods G for which statistical parity encodes the relevant distributive principle may contain goods such as access to education, healthcare, or political offices, and in general all goods that should be uniformly distributed among protected groups. But it is important to note that different types of goods are subject to different distributive principles, as we ascribe different social meanings to different goods. Legal punishment certainly does not fall into the same distributive category as education or healthcare, as it does not seem to be morally required that legal punishment be distributed uniformly

among groups, but rather according to desert. So, the counterexample of Corbett-Davies et al. cannot be taken as an argument against statistical parity per se, but at best as an argument that, if statistical parity is interpreted as an allocative fairness criterion for a certain class of goods, legal punishment does not fall into this category of goods.

Let us now consider an alleged counterexample to *equalized odds*. Recall that equalized odds is the fairness criterion that requires that the probability of a prediction of some target variable, given the actual value of the target variable, be equal for all protected groups. This is a generalization of the requirement that the false positive and false negative rates of the algorithmic decision system be equal for all protected groups. An example of a context in which equalized odds can plausibly be applied is criminal sentencing. The criterion was, for example, used to evaluate whether algorithmic assignments of risk scores, measuring a defendant's risk of violent reoffence, are biased in discriminatory ways. The intuition behind this criterion is that if one protected group has a higher false positive rate than another, meaning that it is more likely for members of one group to actually be innocent and yet be deemed to be at high risk of violent reoffence by the algorithm, this reflects a discriminatory bias on part of the model underlying the algorithm[14].

Gölz et al. (2019) argue against equalized odds as a criterion of algorithmic fairness on grounds that under some circumstances, equalized odds conflicts with certain game-theoretic axioms of fair division. Most strikingly, they contend, equalized odds is largely incompatible with a principle called *population monotonicity*. This principle states that when a finite amount of goods is to be distributed among a number of individuals, removing one individual (for instance because the individual ceases to be interested in the goods to be allocated) should not negatively affect the allocation of goods to the remaining individuals. This means, any individual who would previously have received the good in question should, after the removal of the other individual, still receive the good. Gölz et al. put forward an example along the following lines: imagine a number of student loans can be given out to applicants of a given university. Assume further that the algorithmic decision system which recommends whether to grant a loan to a student or not satisfies equalized odds. That is, for each protected group it is the case that of those students in

[14] A higher false negative rate, on the other hand, reflects a reverse bias: it means that it is more likely to actually be a violent reoffender and yet be deemed to be at low risk of reoffending.

that group who are in fact capable of paying back their loan, an equal proportion are granted a loan. Analogously, for each protected group, of those students in that group who would in fact default on the loan, an equal proportion are denied the loan. Now, if a student from one group, who was granted a loan and is in fact capable of paying it back, decides to reject the loan — maybe because the student decided to enrol at a different university -, this might require withdrawing the initially granted offer of a loan from students of the other protected groups in order to restore equalized odds. It is counterintuitive to think that this would be morally permissible, let alone morally required. In other words, it seems that this shows that equalized odds is not a necessary condition for moral permissibility.

This counterexample, too, can be explained away using the bifurcation of fairness problems. While the cited axiom of fair division, population monotonicity, is concerned with the fair allocation of goods, equalized odds must — contrary to what is done in the example — be interpreted as a criterion of predictive rather than allocative fairness[15]. Since equalized odds is a criterion of which one parameter takes into account what is actually the case in the world (by considering the target variable Y), it nicely fits the schema of the problem of predictive fairness. Interpreting it as an allocative fairness criterion, on the other hand, is implausible: the very notion of a true or false positive can not be meaningfully applied to decision settings. Predictions can turn out to be true or false (or, in the probabilistic case, accurate), but decisions can not. What gives rise to the counterintuitive consequence in the example is the mistaken assumption that equalized odds can act as a fairness constraint on decisions to allocate goods.

Hence, the purported counterintuitive consequence in the example does not actually follow. When applying a predictive model to determine whether a student would pay back their loan, equalized odds can be used to ensure that predictions are not biased. The predictions then act as an input to the decision function in order to determine whom to grant a loan. If one of the students who is initially granted a loan rejects the offer, this has an effect on

[15]Note that a number of articles, contrary to what I propose, explicitly categorize equalized odds as a criterion of allocative fairness. Heidari et al. (2019), for example, understand equalized odds as analogous to a Rawlsian conception of equality of opportunity. This, however, rests on the implausible assumption that the predicted value of the target variable directly corresponds to some measure of utility for the affected individual. This is clearly not the case: the prediction that a person will pay back their loan, for example, will, even if directly tied to a decision, have very different utility values for individuals from, say, different socioeconomic backgrounds.

the distribution of loans, but not on the predictions made by the predictive model. So, it does not affect whether the predictive model satisfies equalized odds or not. Once again, the counterexample emerged due to a failure to distinguish between normative constraints on predictions on the one hand, and normative constraints on the allocation of goods on the other.

These three examples should suffice to show that the conceptual distinction between predictive fairness and allocative fairness can help to rebut many of the arguments put forward against specific notions of algorithmic fairness. Many of the counterexamples arise simply because the scope of proposed fairness criteria is not appropriately delineated. The above examples should count as evidence for the claim that at least part of the difficulty of defining adequate criteria of algorithmic fairness can be explained by the inappropriate framing of the problem of algorithmic fairness as the problem of finding a unique formal criterion for the moral permissibility of an algorithmic decision system.

1.6.2 Pairwise incompatibility

Next, we consider the issue that three of the fairness criteria — statistical parity, equalized odds, and predictive parity — are pairwise incompatible except under highly constrained circumstances. Let us make this more precise. In (Kleinberg et al., 2016), it is shown that statistical parity is inconsistent with both, equalized odds and predictive parity, whenever there is a correlation between the protected characteristic A and the target variable Y. That means, whenever the base rate of the property of interest differs between protected groups, an algorithm whose outcomes satisfy statistical parity will satisfy none of the other two fairness criteria. What is more, if in addition to this condition the predictor is imperfect — which means that it is not the case that $\hat{f}_{\boldsymbol{\theta}}(\cdot)$ assigns probability 1 to all individuals for whom Y takes value 1, and probability 0 to all individuals for whom Y takes value 0 -, then equalized odds and predictive parity cannot be satisfied simultaneously. The latter was independently shown by Kleinberg et al. (2016) and Chouldechova (2017).

While the conceptual bifurcation of fairness problems will certainly not be able to completely resolve this impossibility result, it can offer a partial explanation as to why the impossibility emerges. In particular, it can offer an explanation as to why statistical parity is inconsistent with equalized odds and predictive parity. Again, this is due to a conflation of the predictive model

and the decision function. Statistical parity is an allocative fairness principle whose aim it is to ensure an equitable distribution of goods. Equalized odds and predictive parity, on the other hand, must both be interpreted as criteria of predictive fairness. Equalized odds ensures that individuals have equal probabilities of receiving a false positive (and false negative, respectively) prediction across all protected groups. Predictive parity, on the other hand, ensures that individuals have equal positive and negative predictive values across all protected groups. The positive predictive value is the conditional probability of the target variable taking a specific value, given the prediction that it would take this value. Both criteria are based on metrics that compare predictions with some ground truth — the false positive/negative rate in the case of equalized odds, and the positive/negative predictive value in the case of predictive parity. So, they both fit the schema of predictive fairness, while it would require somewhat of an interpretive stretch to understand them as allocative criteria.

We can think of situations in which it is possible and desirable that the predictive model satisfies either equalized odds or predictive parity, and where, at the same time, the decision function satisfies statistical parity. Think, for instance, of the earlier example in which an algorithmic decision system is applied in order to make university admission decisions. Assume the protected characteristic we care about is whether a prospective student has dyslexia. We might think that students with dyslexia and students without should have equal admission rates — that is, the distribution resulting from the admission decision algorithm should satisfy statistical parity. To this end, we might deploy an algorithmic system that has a predictive model which satisfies either equalized odds or predictive parity, and can hence be assumed to provide unbiased predictions of how well a prospective student will perform. Based on these performance predictions for each student, we might define two separate cutoff thresholds for admittance, one for the students with dyslexia and one for the students without — this is the decision function of the algorithmic decision system. The thresholds are chosen such that the best, say, fifty percent of each group are admitted. This creates outcomes which satisfy the allocative criterion statistical parity, and these outcomes result from decisions which are partly based on predictions which satisfy one of the predictive fairness criteria equalized odds or predictive parity.

The above example shows that the incompatibiltiy of statistical parity on the one hand, and equalized odds and predictive parity on the other, is due to

a conceptual conflation of predictions and decisions and can be explained away using our bifurcation of fairness criteria. The incompatibility of the two predictive fairness criteria equalized odds and predictive parity, however, cannot be explained away so easily[16]. Arguably, this is the more worrying incompatibility because, as opposed to statistical parity, equalized odds and predictive parity both seem, at first glance, to be universally desirable properties of predictive models. Yet, the bifurcation can at least shed some light on the incompatibility of fairness criteria, even though it cannot completely resolve the problem.

This, at the same time, highlights the limits of the approach presented here. While many incompatibilities and counterintuitive situations can be resolved by recognizing the difference between moral requirements on the predictive model and the decision function, of course not all problems in the field emerge only from this conflation. Some incompatibilities remain, and quite possibly new ones will be discovered in the future. Resolving these will require further analysis of the moral norms which underlie the criteria involved, and of how conflicts between them can be resolved. Nonetheless, the distinction between predictive and allocative fairness criteria is an important step in the direction towards fully resolving the question under which circumstances which criterion of fairness is to be used.

1.6.3 Level of description

Lastly, let us turn to the problem of different levels of description. The problem here is that it is unclear whether fairness should be described as a group-level property or as an individual-level property. This is particularly problematic because it is often the case that an algorithmic decision system that satisfies a criterion of individual fairness will lead to outcomes that fail to satisfy any of the group fairness criteria.

Before we try to analyze this issue through the lens of the conceptual distinction between predictive and allocative fairness, let us first make the distinction between group and individual fairness more precise. The term *group fairness* describes all those criteria of fairness which compare some summary statistic across protected groups. This category encompasses the criteria statistical parity, equalized odds, and predictive parity. Each of those criteria

[16] This incompatibility is addressed in more detail in Chapter 4.

considers a different summary statistic — the average probability of a positive outcome, for instance, or the average probability of a false positive or false negative — for each protected group and compares them. If the summary statistic is equal (or sufficiently similar) across protected groups, the fairness criterion is satisfied. The term *individual fairness*, on the other hand, describes all those criteria of fairness which take into account individual cases of algorithmic decision-making and determine whether the outcome is fair. Whether an individual algorithmic outcome is fair is, for instance, determined by comparing the case to similar actual cases, as implemented in the criterion *fairness through awareness*, or by comparing it to a similar hypothetical case, as implemented in the criterion *counterfactual fairness*. Being able to categorize fairness criteria in this way has lead many researchers to adopt the following widely held idea: while group fairness ensures that protected groups are treated fairly, individual fairness ensures that individuals are not treated unfairly due to having a certain protected characteristic.

The most thorough philosophical treatment of the distinction between individual and group fairness and their potential conflict can be found in an article by Binns (2020). Binns argues that the apparent conflict between individual and group fairness criteria can be resolved by acknowledging that whether criteria of individual or group fairness are called for depends on what is believed about the data generating process underlying the observed data. If disparities between protected groups in the data are assumed to be due to unjust social structures (for instance in that disparities came about through discriminatory institutional practices), group fairness criteria are appropriate. These criteria help to counteract unfair inequalities. If, on the other hand, observed disparities can be assumed to be due to the free choices of the individuals constituting these groups, individual fairness criteria are most appropriate. In this case, there is no need to ensure that outcomes are equalized, only that everyone's outcomes are arrived at in the same way.

To some extent this explanation is plausible, and in some situations it can help to guide the choice of an appropriate fairness criterion. It is, however, important to acknowledge that Binns' explanatory hypothesis rests on two strong assumptions. First, the assumption that any observed disparity in algorithmic outcomes between two protected groups which is not due to the free choices of the individuals within the group must be due to unjust social structures. Secondly, the assumption that luck egalitarianism is the correct theory of distributive justice. This means, it is assumed that disparities in the

distribution of goods are only justified if they are the result of the free choices of individuals. When the properties determining the outcome are not within the control of the individuals themselves, disparities in the distribution of goods are not justified. For Binns' explanation to count as a general account of the relation between group and individual fairness rather than as a partial guide to choosing an appropriate criterion, these two conditions have to be taken to hold universally. While I do not aim to scrutinize these two assumptions here, it is important to note that both assumptions are contentious.

Rather than formulating a criticism of Binns' thesis, we can try to offer a related but more general account of the relation between individual and group fairness. To this end, let us begin with a general consideration of the potential wrongs at play in public decision-making. As outlined in Section 1.4, public decisions can be morally wrong in two different ways: they can either be wrongful due to being discriminatory, or because they produce unjust distributions of benefits and burdens among protected groups. These two wrongs arise at different levels of description. While the moral wrong involved in discrimination[17] occurs at the level of the individual, the moral wrong involved in distributive injustice (unless it is caused by discrimination) occurs at the level of the collective. To illustrate this, consider the following example, which was put forward by Eidelson (2015, pp 55-56). Imagine a society which is, by and large, friendly and accepting towards homosexual people, but in which a specific situation occurs where an individual homophobe treats a gay person in a discriminatory way. Here, it seems clear that it is the individual person who is wronged (because of their sexuality), and not (or at least to a lesser extent) the gay community as a whole. On the other hand, we can imagine a school whose admission criteria are such that the most talented students are admitted. As it turns out, however, this results in children of working class parents being admitted at lower rates than children of academics. Insofar as this can be considered morally wrong, it is a wrong that emerges only at the collective level. As every student is held to the exact same standard, it would be hard to argue that individuals are being wronged in this scenario.

Turning back to the topic of algorithmic fairness, we recall that algorithmic decision-making can be discriminatory due to predictions which are biased in discriminatory ways, and that they can result in unjust distributions when

[17] I here refer only to direct discrimination, and exclude what is sometimes called indirect discrimination.

the decision function allocates goods on the basis of inadequately chosen variables. Now, if we accept the above line of reasoning, this would suggest that criteria of predictive fairness are concerned with avoiding wrongs at the individual level (as they prohibit bias and hence prevent discrimination), while criteria of allocative fairness are concerned with avoiding wrongs at the collective level (as they prohibit unjustly distributed decision outcomes). So, we would expect that criteria of individual fairness are in fact criteria of predictive fairness, and that criteria of group fairness are in fact criteria of allocative fairness.

This, however, is not generally the case. Some of the criteria fall into one, but not the other category. Equalized odds, for example, is considered a criterion of group fairness, since whether it is satisfied or not depends on summary statistics across groups. At the same time, as was argued earlier, equalized odds fits the schema of predictive fairness, and would be implausible as an allocative fairness criterion. An analogous argument can be made for predictive parity.

In order to make sense of this mismatch, one has to consider the interpretation of equalized odds. First, it is important to note that equalized odds should not be understood as a definition of predictive fairness. Rather, the violation of equalized odds should be understood as an indication that a predictive model is biased. Bias here is to be understood as an over- or underestimation of the relevance of a protected characteristic for predicting the property of interest. Imagine, for instance, that a prediction of the expected profitability of a potential future employee is to be made. We can assume that for this prediction, gender is an irrelevant property. In this scenario, a prediction would count as gender-biased if the information that an applicant is female were to lead to a lower prediction of the expected profitability. If such bias were present, this would, on the whole, lead to a higher rate of false negative predictions for women than for men.

Underestimating an applicant's expected profitability for the company due to their gender is a wrong that occurs at the individual level. Yet, due to the opacity of the workings of many predictive algorithms, it is often hard to detect what exactly leads to a given prediction. It is much easier to analyze summary statistics to detect systematic patterns in the predictions. This is what equalized odds is intended for: it is a metric that uses summary statistics to infer whether a predictive model is systematically biased and hence

wrongs individuals with a given protected characteristic by over- or underestimating the relevance of certain traits.

This analysis shows that the two categories individual fairness and group fairness, as used in the literature, do not map neatly onto the underlying space of moral concepts. The conceptual distinction between predictive and allocative fairness, in contrast, is more helpful in categorizing criteria in a way that corresponds to the relevant underlying moral concepts. It would hence seem sensible to give up the distinction between individual and group fairness and replace this distinction with the conceptual distinction between predictive and allocative fairness.

Under certain assumptions, the predictive/allocative fairness distinction entails Binns' hypothesis that the choice of fairness criterion should depend on the underlying data generating mechanism. Recall that predictive fairness criteria should hold universally — a predictive model should under no circumstances be biased with regards to a particular protected group. Two individuals with relevantly similar input features, but of different protected groups, should consequently receive similar (or similarly accurate) predictions. Allocative criteria, on the other hand, are indexed by goods. Following Walzer (1983), we can assume that which allocative criterion holds for a given good is determined by the social meaning ascribed to this good. If we now, as Binns does, assume that goods are to be allocated according to some distributive parity criterion whenever the properties relevant for the allocation decision are outside a person's control, and there are disparities in the predictions between different protected groups, then it will be necessary to take further variables beyond the predicted variable into account to ensure that despite disparities in the predictions, the decision outcomes are distributed equitably.

1.7 Potential objections

I will now address a number of potential objections to the proposed framework and the assumptions on which it is built. The first objection is that the problem of algorithmic fairness is not presented in an adequate form. The second objection is that one central premise, namely that we can clearly distinguish predictions from decisions, is false. Let us discuss both potential objections in turn.

1.7.1 Misrepresentation of the problem of algorithmic fairness

One could argue that the way the problem of algorithmic fairness is presented here — namely as an attempt to find a single formal criterion that is a necessary and sufficient condition for the moral permissibility of an algorithmic decision system — does not correspond to the reality of what researchers in the field of algorithmic fairness are actually doing. Instead of trying to find a single formal criterion which provides a necessary and sufficient condition for fairness, they aim to identify individually necessary conditions of moral permissibility, with the greater goal of being able to find the list of all those individually necessary conditions which are jointly sufficient for the moral permissibility of algorithmic decision systems. Formally represented, we could say that on this alternative view, researchers are trying to find some ϕ_i, such that $\phi = \phi_1 \wedge ... \wedge \phi_n$, with $i \in 1, ..., n$.

The first thing to be said about this is that, clearly, many of the seminal papers in the field of algorithmic fairness can be understood as attempts to formulate a *definition* of fairness[18]. Giving a definition typically means providing necessary and sufficient conditions. But granted that indeed the goal of most authors is to provide only necessary conditions for fairness, would this invalidate our argument?

The central point this chapter is trying to establish is that when considering criteria of algorithmic fairness, be they intended as necessary and sufficient, or as necessary conditions only, one has to take into account whether these criteria are reasonable constraints on the predictive model or on the decision function. This determines whether in evaluating the algorithmic decision system, we take the output to be the probabilistic prediction $\hat{f}_{\boldsymbol{\theta}}(\mathbf{x})$ or the decision option o. Given an algorithmic decision system and a criterion of algorithmic fairness, we might come to different conclusions about whether it satisfies the criterion depending on whether we take $\hat{f}_{\boldsymbol{\theta}}(\mathbf{x})$ or o to be the relevant output. The aim of proposing a framework for distinguishing between predictive and allocative fairness criteria is to eliminate this kind of ambiguity.

While the assumption that the problem of algorithmic fairness is the search

[18] See, e.g., Dwork et al. (2012, p. 2), who speak about "our definition of fairness", or Kusner et al. (2017, p. 16), who speak about giving a "causal definition of fairness".

for a single formal necessary and sufficient condition of moral permissibility provides a motivation for the present project, the value of the proposed framework does not hinge on this assumption.

1.7.2 The distinction between predictions and decisions

The argument outlined in this chapter builds on the assumption that we can, at least in most cases, clearly distinguish between predictions — interpreted as forming an epistemic attitude towards an unobserved event or property — and a decision — interpreted as the choice to pursue one specific course of action. But while in theory the distinction can be upheld, there are some arguments to the effect that this distinction is less strict. This involves some major philosophical projects such as epistemic utility theory (see, e.g., Pettigrew, 2016), or the theory of epistemic democracy (see, e.g., List & Goodin, 2001; Goodin & Spiekermann, 2018) . Let us discuss both in turn.

The central idea of epistemic utility theory is to apply the mathematical machinery of decision theory to the evaluation of epistemic norms. At its foundation sits the assumption that, from an epistemic point of view, all we care about is coming to believe true (and only true) propositions. Epistemologists are hence concerned with finding norms of belief formation that are optimal with regard to this goal. The gist of epistemic utility theory is that the structure of the epistemic problem — forming beliefs in a way that is optimal with regard to the goal of accuracy — is similar to the problem of practical rationality — taking decisions in a way that is optimal with regard to one's personal preferences or values. Since the structure is similar, the methods used to evaluate decision strategies can also be used to evaluate epistemic norms. Nonetheless, epistemic utility theory is about norms of rational belief formation, not about rational decision-making, even though it applies the formal framework of the latter. On our more orthodox interpretation of what a decision is, making predictions cannot be seen as a species of decision-making, since, as Pettigrew (2016, p. 207) himself puts it, "we don't choose our doxastic states". Moreover, adopting a doxastic state does not allocate any goods — and this is, at least in the present context, the central type of decision from which we wish to distinguish predictions. The project of epistemic utility theory, then, does not seem to put into doubt the feasibility of the distinction between predictions and allocative decisions in the context of algorithmic decision-making.

Another philosophical project which seems to blur the lines between decision-making and belief formation is the theory of epistemic democracy. Here, the central notion is that in collective decision-making, there is some fact of the matter about which choice can be considered to be correct. This, however, has to be understood in the following way. For each of the options available to the collective, it is possible to assign an objective utility. On the basis of this objective utility assignment, we can say that it is true (or false) that a given option is the best option available. Choosing the best option can be considered the correct decision, choosing any other option an incorrect decision. While this view introduces some epistemic aspects into collective decision-making, it would be an overstatement to say that the view implies that we cannot clearly distinguish between (purely) epistemic practices (like making predictions) and the act of making a decision to allocate some good.

Now, even if one were to concede that we can understand belief formation as a species of decision-making, or that one can call some decisions (in some epistemic sense) correct and others incorrect, this would still not necessarily invalidate our thesis. The minimal premise needed for the argument outlined here is that in the context of algorithmic decision-making, it is clear whether at a given moment we are concerned with predicting an event or a property, or whether we are concerned with allocating a good. This does not seem to be put into doubt by either of the two projects described above.

1.8 Conclusion

I have argued that the way the problem of algorithmic fairness is commonly presented is misleading and unlikely to be solvable. This, as I have argued, is due to the fact that it conflates two different realms of ethical consideration, namely predictions and decisions. An algorithmic decision system typically makes (or recommends) decisions on the basis of predictions of some variable of interest. Here, two distinct morally problematic phenomena can occur: first, the predictions can exhibit discriminatory bias, and secondly, the decisions can lead to unfair distributions of goods or opportunities. I have provided a general formal schema that helps to individually diagnose and address each of these two problems — the problem of predictive algorithmic fairness, and the problem of allocative algorithmic fairness. I then demonstrated how this bifurcation of fairness criteria enables us to (at least partially) resolve many of the paradoxes that beset the original problem of

algorithmic fairness. I concluded by considering two potential objections to the framework.

Chapter 2

Yet Another Impossibility Theorem in Algorithmic Fairness

2.1 Introduction

In this chapter, I will consider the relation between three of the most popular criteria of predictive algorithmic fairness: *counterfactual fairness*, *equalized odds*, and *predictive parity*. Recall the ideas underlying these criteria: 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 chapter 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 and widely held 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.

Some earlier works have discussed 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). No previous work discusses the incompatibilities presented in this chapter in depth.

The remainder of the chapter is organized as follows. In Section 2.2, I introduce the concept of a causal structure and present a theorem used in the proof of the impossibility result. In Section 2.3, I introduce formal definitions of the three fairness criteria counterfactual fairness, equalized odds, and predictive parity. In Section 2.4, I state and prove the impossibility theorem before then discussing ways to circumvent it in Section 2.5.

2.2 Causal structures and the projection theorem

Recall that in Section 0.3, we defined causal models as triples $(\mathbf{U}, \mathbf{V}, F)$, with $\mathbf{U}$ and $\mathbf{V}$ being sets of variables, and F a set of structural equations. 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 0.1 and 0.2 (from Section 0.3) 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 neither m nor any of its descendants is a node representing a variable in $\mathbf{Z}$ (Pearl, 2009, pp. 16-17). In Figure 0.1, 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 0.1, 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[1] that a probability distribution $P(\cdot)$ is both *faithful* and *Markov* relative to its associated causal structure G_M.

The above notions now allow us to describe a theorem which will later help

[1]For a discussion and defense of the two assumptions, see Pearl (2009, pp. 61-64) and Zhang & Spirtes (2016).

us construct an economical proof of the impossibility theorem presented in this chapter. I 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 and Rubin, 2018). 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.

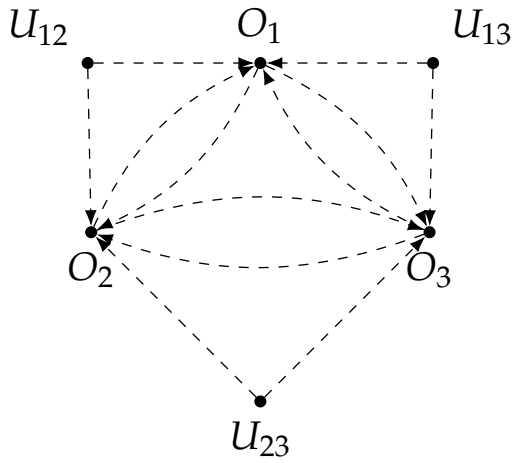


FIGURE 2.1: Representation of all the possible causal structures between O_1, O_2, and O_3.

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 distribution but not what the causal relations between the variables are, we can assume there 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 2.1 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.

2.3 Fairness in predictive models

Let us now turn back to the topic of predictive models in algorithmic decision-making. As argued in the previous chapter, 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. The different fairness constraints that have been proposed in recent years are aimed at guaranteeing that, provided a model satisfies the fairness constraint, it does not exhibit such bias. Each of these 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.

The intuitive motivations for the three constraints have been mentioned in the previous chapter. I will now introduce the formal definitions and the rationales underlying the constraints. 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*, and of a *negative prediction* whenever $\hat{Y} = 0$. Analogously, we will, for lack of better terminology, call the individuals for which it is the case that $Y = 1$ and $Y = 0$ the *positive* and *negative class*, respectively. Generally, we will take predictive models to be functions of the form $\hat{f}_{\boldsymbol{\theta}} : D_{\mathbf{X}} \rightarrow D_{\hat{Y}}$, 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[2]. To keep the discussion simple, however, we will in this chapter 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 $\hat{f}_{\boldsymbol{\theta}}$ 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 $\hat{f}_{\boldsymbol{\theta}}$ representing the predictive model as a structural equation to F_{base}. We here interpret the function $\hat{f}_{\boldsymbol{\theta}}$ 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 $\hat{f}_{\boldsymbol{\theta}}$ 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 $\hat{f}_{\boldsymbol{\theta}}$, hence referring to causal relations within an augmented causal model as outlined above.

2.3.1 Equalized odds

The first fairness constraint I will 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

[2]Note, that with regards to the model of algorithmic decision systems introduced in Section 1.3, this can be interpreted as a boundary case of a predictive model, as $D_{\hat{Y}} = \{0, 1\}$, and $\{0, 1\} \subset [0, 1]$. That is, it can be considered a predictive model that only estimates extreme probabilities, namely 0 and 1, where a prediction of 0 is considered a negative, a prediction of 1 a positive prediction.

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 2.3.1 (Equalized odds)**.** A predictive model $\hat{f}_{\boldsymbol{\theta}}$ satisfies *equalized odds* (relative to A) if and only if for all $\hat{y} \in D_{\hat{Y}}$, $y \in D_Y$, and the constants $a_1, a_2 \in D_A$

$$P(\hat{y} \mid a_1, y) = P(\hat{y} \mid a_2, y) \tag{2.1}$$

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.

2.3.2 Predictive parity

Next, I will 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 2.3.2 (Predictive parity)**.** A predictive model $\hat{f}_{\boldsymbol{\theta}}$ satisfies *predictive parity* (relative to A) if and only if for all $\hat{y} \in D_{\hat{Y}}$, $y \in D_Y$, and the constants $a_1, a_2 \in D_A$

$$P(y \mid a_1, \hat{y}) = P(y \mid a_2, \hat{y}) \tag{2.2}$$

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 $\hat{Y}$.

2.3.3 Counterfactual fairness

The third and most complex fairness constraint to be introduced 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 2.3.3 (Counterfactual fairness)**.** A predictive model $\hat{f}_{\boldsymbol{\theta}}$ satisfies *counterfactual fairness* (relative to constant a_1) if and only if for all $\hat{y} \in D_{\hat{Y}}$ and $\mathbf{x} \in D_{\mathbf{X}}$

$$P(\hat{Y}_{a_1} = \hat{y} \mid \mathbf{x}, a_1) - P(\hat{Y}_{a_2} = \hat{y} \mid \mathbf{x}, a_1) = 0 \tag{2.3}$$

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.

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.

2.4 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, I 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.

I will explain 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 0.3. 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

$$P(Y_{a_1} = y \mid \mathbf{x}, a_1) - P(Y_{a_2} = y \mid \mathbf{x}, a_1) \neq 0 \tag{2.4}$$

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, I will first specify the framework and the assumptions applied in the proof. Generally, the idea is to construct a graphical proof of the theorem that 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. The following are the premises of the argument:

Premise 1 (Predictive model)**.** We assume that a predictive model is a function $\hat{f}_{\boldsymbol{\theta}} : D_{\mathbf{X}} \rightarrow D_{\hat{Y}}$ 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:

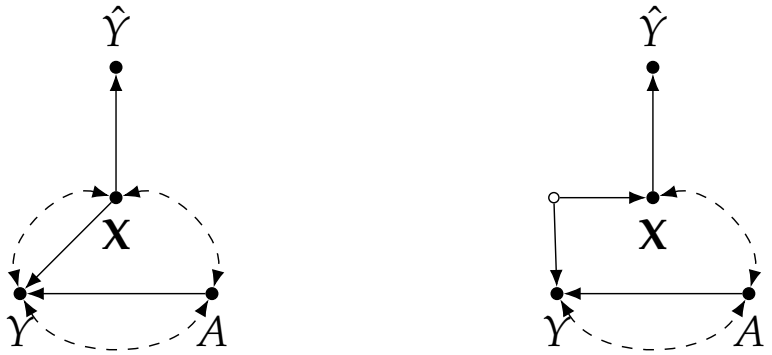
- there is an edge from $\mathbf{X}$ into Y,
- there is an edge from Y into $\mathbf{X}$, or
- there is an unobserved node with outgoing edges into both, Y and $\mathbf{X}$ (i.e. the node represents a latent common cause).

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 proof by cases. We can 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 be represented as a causal structure where the only latent variables are



(A) (B)

FIGURE 2.2: The two possible causal structures involving Y, $\hat{Y}$, A, and $\mathbf{X}$.

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 causal structures 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 (as this implies that the relevant conditional independence holds), 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, $\hat{Y}$, $\mathbf{X}$, and A compatible with premises 1-5. The two resulting classes of graphs are depicted in Figure 2.2.

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 2.2a, 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 2.2b, 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. □

To summarize, I have shown that the four individually plausible propositions about algorithmic fairness introduced at the beginning of this section are jointly inconsistent. In the next section, we explore how we can circumvent this impossibility result.

2.5 Escaping the impossibility

I will now consider how the impossibility established in the previous section can potentially be circumvented. 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).

2.5.1 Relaxing proposition (4)

Let us 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.

2.5.2 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 these 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. The 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 toward a woman than they would toward a man, or to someone with disability than to someone without 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 therefore 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). 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 can be easily refuted, too. To see this, we

[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.

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, 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 (which is generally considered to be a protected characteristic) 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 occur too infrequently 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.

2.5.3 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 the unjustified disadvantageous treatment of an individual (as compared to another individual) where this treatment can be (causally) explained by the fact that the former possesses a protected characteristic that the latter does not possess (see, e.g., Eidelson, 2015; Lippert-Rasmussen, 2014; Moreau, 2010). More simply put, discrimination

occurs when a protected characteristic makes an unjustified difference in 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 an idea along those lines, except for the fact that it does not take into account that under some circumstances the influence of a protected characteristic on the prediction might be justified.

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[4].

2.5.4 Relaxing proposition (1)

Let us now consider whether we can give up or relax the claim that for a model to be fair, it is necessary that it satisfy either equalized odds or predictive parity. Giving up this claim means accepting that even if neither predictive parity nor equalized odds is satisfied, the predictions of a model can still be considered fair. We will first consider an argument for giving up predictive parity as a universal criterion of fairness. We will then consider an argument for giving up equalized odds as a universal criterion of fairness.

Predictive parity ensures that the positive and negative predictive values of a predictive model — the probabilities of positive/negative predictions being true — are equal for all protected groups. If the predictive values are equal, this means that on average predictions are equally informative for members of different protected groups. The argument for requiring that predictions be equally informative is based on the idea that a difference in informativeness across protected groups' predictions incentivizes treating individuals differently on the basis of their protected characteristics. From a normative perspective, this seems undesirable as it rationalizes disadvantaging individuals from one protected group through no fault or shortcoming of their own.

We can, for example, imagine a company using a hiring algorithm to predict whether a potential employee would be profitable for the company if hired. We can assume that the proportion of potentially profitable employees is equal among the subpopulation of female applicants and the subpopulation of male applicants. If we now assume that the positive predictive value is lower for female applicants than for male applicants, then the employer has an incentive to prefer male applicants that were predicted to be profitable, because the employer knows that the probability of them actually being profitable for the company is higher for male applicants who were predicted to be profitable than for female applicants predicted to be profitable. This is so even though the employer knows that overall, the male and female applicants are equally likely to be well qualified and hence profitable for the company.

[4]This is approach is motivated and pursued in more detail in Chapter 3.

Giving up proposition (1) entails that one accepts that there are circumstances under which a situation like the above is not morally problematic. An argument to this effect has been offered by Hellman (2019). She argues that although it seems intuitively undesirable that predictions are not equally reliable for all protected groups, this is not a case of discriminatory bias. When it seems that equal predictive values are required from a normative point of view, then this is so on the basis of a presumed equal entitlement to accurate predictive instruments (Hellman calls this the "right to the best available decision-making tool" (Hellman, 2019, p. 833)). Hellman argues, however, that in many cases it is doubtful whether one really is entitled to such a right. If we accept this argument, we can accept to give up one disjunct of proposition (1).

Let us now turn to an argument for giving up the second disjunct — namely that fair predictive models satisfy equalized odds. The argument we will consider aims to establish that, by focusing on implications for decision-making, it is, under specific circumstances, morally permissible to give up equalized odds. Recall that, from a decision-making perspective, a violation of equalized odds can be interpreted as not holding individuals from different protected groups to the same standards. Requiring that fair predictive models satisfy equalized odds hence implicitly presupposes that it is unfair to hold individuals from different protected groups to different standards.

There are, however, situations in which this does not seem to be the case. In particular, when due to past injustices, opportunities are not distributed equitably, it might be morally permissible to hold individuals to different standards based on their membership in a protected group. In a society, for instance, in which for a long period of time a certain minority was systematically discriminated against, and where, consequently, members of this minority group have lower average levels of education and socioeconomic status, it might be morally permissible or even required that when it comes to, say, university admission decisions, members of this minority be held to lower standards. Such affirmative action policies can help achieve what is sometimes called compensatory justice — making up for the past injustices a minority has suffered. Such policies can also be justified from other perspectives. First, one might argue that sometimes brute luck leads to unequal opportunities. This is for instance the case when children are born with disabilities. Luck egalitarian theories of justice would have it that it is unfair that children who, through no fault of their own, were born with a disability, face

reduced opportunities in life as a consequence. This, so the argument goes, ought to be compensated by holding them to different standards than children without disabilities. A third argument for such policies can be made by appealing to the intrinsic value of diversity. According to it, there is intrinsic value in having, say, students of a diverse range of different backgrounds in a cohort that reflect the diversity of backgrounds in the general population. This includes, for example, different ethnic and socioeconomic backgrounds. If the university's standard admission procedures do not achieve a proportional representation of the different backgrounds present in the general population, lowering the standards for some groups seems to be a morally permissible way to increase diversity. If we accept that equalized odds encodes the rationale that everyone should be held to the same standards, these arguments can be taken as arguments against equalized odds as a universal requirement of fairness for predictive models[5].

2.6 Conclusion

To summarize, I have shown that four intuitively plausible propositions about predictive algorithmic fairness are jointly incompatible. After discussing different ways of escaping this impossibility by giving up one or more of the propositions, I 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. Secondly, one could give up predictive parity and only require equalized odds in specific situations.

[5]Note, however, that in the light of the distinction between predictive and allocative fairness introduced in Chapter 1, this argument can easily be refuted. It seems that different standards for different protected groups ought to be implemented as a constraint on the decision function, not the predictive model.

Chapter 3

Causal Relevance Fairness

3.1 Introduction

Much of the early discussion on algorithmic fairness was centered around statistical criteria of fairness. These criteria define fairness in terms of joint probability distributions over a set of relevant variables. A criticism of these criteria, however, is that they are unable to distinguish between certain cases of which some intuitively seem to be cases of discrimination, whereas others do not. Whether an algorithmic prediction is discriminatory or not seems to be determined by facts that go beyond mere population-level correlations.

This implies that in order to adequately determine whether an algorithmic prediction should count as discriminatory, one has to, first, evaluate individual instances of algorithmic predictions instead of population-level patterns, and, secondly, consider the underlying mechanisms by which the prediction came about. Causal modeling has become an increasingly popular framework for achieving this in developing criteria of algorithmic fairness.

The most popular and widely discussed fairness criterion that makes use of causal modeling is *counterfactual fairness*. As outlined in the previous chapters, counterfactual fairness formalizes the idea that an algorithmic prediction is fair if and only if it is the case that the prediction would have been the same, had the relevant protected characteristic (e.g. gender or ethnicity) been different. The central idea is thus that in order to determine whether a prediction was fair, we have to compare it with a prediction in a counterfactual world, where the individual about whom the prediction is made, has, say, a different gender or is of different ethnicity, but is otherwise the same. If the prediction is the same in both, the actual and the counterfactual world, the

prediction counts as fair.

Yet, counterfactual fairness, too, suffers from problems. Besides the result from the previous chapter, a number of recent papers have formulated further criticisms of the criterion. While the main focus of these criticisms is problems of applying counterfactual fairness in practice (see, e.g., Kilbertus et al., 2020; Wu et al., 2019), some have argued that there is something conceptually amiss with counterfactual fairness. It seems that there are certain cases in which counterfactual fairness is violated, but which, according to most theories of discrimination, nonetheless would not constitute wrongful discrimination. This point is addressed in a recent paper by Chiappa (2017), who acknowledges that counterfactual fairness seems like an overly strong requirement in certain situations. To mitigate this problem, Chiappa proposes to weaken counterfactual fairness so that only certain causal links between the protected characteristic and the algorithmic prediction are considered unfair, but others are not. This results in a fairness criterion that allows for certain differences between the prediction in the actual and the counterfactual world. However, this approach fails to provide a principled way of distinguishing between fair and unfair causal links. A fairness criterion that requires specifying what is fair a priori is conceptually circular and of little practical use.

In this chapter, I aim to provide an alternative causal definition of fairness that allows to distinguish between fair and unfair causal effects on the prediction in a principled way. The new criterion I propose, *causal relevance fairness*, formalizes the idea that a prediction is fair only if the protected characteristic's effect on the prediction does not exceed its (causal) relevance for the prediction. In other words, the effect of the protected characteristic on the prediction should, at most, be as great as its actual effect on the target variable that is to be predicted. In contrast to counterfactual fairness, this criterion is not susceptible to the counterexamples above, but, unlike Chiappa's path-specific refinement of counterfactual fairness, it is firmly grounded in ethical theories of discrimination.

The plan for the rest of the chapter is as follows. In Section 3.2, I will present two challenges for predictive algorithmic fairness criteria, each consisting of two cases that any reasonable criterion should be able to distinguish. In Section 3.3 I then address the question of how wrongful discrimination can be defined for the context of algorithmic predictions, before I introduce the new

causal fairness criterion, causal relevance fairness, in Section 3.4. In Section 3.5, I revisit the two challenges outlined earlier and show that causal relevance fairness meets these challenges. Section 3.6 discusses a number of noteworthy points and addresses two potential objections to the criterion.

3.2 Two challenges for predictive fairness criteria

The move from statistical to causal criteria of algorithmic fairness was, at least in part, motivated by the observation that a specific class of cases poses problems for statistical criteria of fairness. These are cases in which observational information alone does not suffice to detect morally relevant differences. Additional information about the underlying mechanisms is required in order to distinguish cases that constitute wrongful discrimination from those that do not. This is where causal criteria of fairness, and in particular counterfactual fairness, come in handy. Fairness criteria based on causal assumptions about the underlying structures allow us to differentiate between cases which, despite identical joint probability distributions over the variables of interest, differ in their moral evaluation.

To illustrate this problem, consider the following case[1]. A university applies a predictive algorithm to predict whether a given applicant is adequately qualified to succeed in their degree. We can, as before, take the predictive algorithm to simply be a function from a number of input variables to a variable that represents the algorithm's prediction of a student's success in their academic studies. To keep things simple, I will, for the moment, leave probabilistic predictions aside and only consider binary predictions. Imagine the actual outcomes and the algorithmic predictions are as represented in the confusion matrix in Table 3.1. The columns show the actual outcomes, where $Y = 1$ and $Y = 0$ stand for *adequately qualified* and *unqualified*, respectively, while the rows show the predicted outcomes, where $\hat{Y} = 1$ and $\hat{Y} = 0$ stand for *predicted to be adequately qualified* and *predicted to be unqualified*, respectively. The top left cell, for example, shows the number 80. This cell represents the number of adequately qualified ($Y = 1$) male applicants who were correctly predicted to be adequately qualified ($\hat{Y} = 1$).

[1]The case outlined here is similar to a real-world case in which the University of California, Berkeley was suspected of discriminating against female applicants (Bickel et al., 1975). Kusner et al. (2017) use a similar example to illustrate their counterfactual fairness criterion.

	Male		Female		
	$Y=1$	$Y=0$	$Y=1$	$Y=0$	Total
$\hat{Y}=1$	80	0	72	0	152
$\hat{Y}=0$	20	100	28	100	248
Total	100	100	100	100	

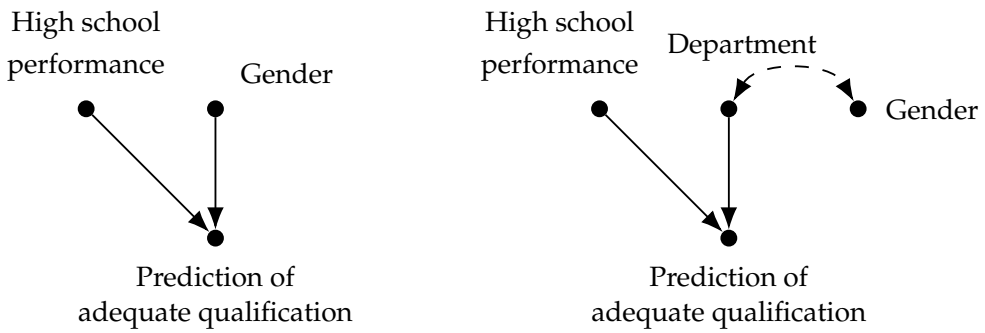
TABLE 3.1: Confusion matrix representing the performance of a hypothetical predictive algorithm.

We observe that there are 200 male and 200 female applicants. The bottom row tells us that 100 of the 200 male and, likewise, 100 of the 200 female applicants are adequately qualified for a university degree. In other words, there is no observable difference in qualification between men and women. Yet, according to the first row, 80 men were predicted to be adequately qualified, while only 72 women were. Now compare the following two scenarios which could have generated the data:

Scenario 1: All applicants in our dataset applied for a business degree. The predictive algorithm used by the university takes as input data an applicant's high school performance (X_1) and their gender (A). As it turns out, both, high school performance and gender, are used by the algorithm as predictors[2]. This means, a female applicant for the business degree is less likely to receive the prediction that she is adequately qualified than a male applicant with identical high school grades. This is represented in the causal graph in Figure 3.1a. Intuitively, it seems that this predictor is biased against women.

Scenario 2: Applicants in our dataset applied for either a degree in physics or a degree in business. The predictive algorithm takes as input data only the applicant's high school performance (X_1) and the degree the applicant applies for (X_2). As it happens, women have a different tendency in choosing their degree than men. A large percentage of women decide to apply for the physics degree (90%) rather than the business degree (10%), while the male applicants are equally divided between business (50%) and physics (50%). Since the physics department has higher entry requirements than the business department, the average rate at which adequately qualified applicants

[2]Note that the presence of a variable in the dataset does not necessitate that it is used by the predictive model in predicting the target variable. Often, feature selection methods which narrow down the number of used variables are part of the machine learning pipeline, for instance in the form of regularization (e.g. Lasso regression) or as part of the pre-processing of the data (e.g. principal component analysis).



(A) Scenario 1 (B) Scenario 2

FIGURE 3.1: Two different causal models potentially producing the same joint probability distribution. The solid arrows indicate causal relations, while the bidirected dashed arrow indicates a correlation due to some unobserved factor which acts as a common cause of both variables.

are predicted to be adequately qualified is only 0.7 for the physics degree, while for the business degree it is 0.9. In other words, the false negative rate is 0.3 for the physics department and 0.1 for the business department. Male and female applicants are equally likely to be predicted to be adequately qualified if we look at each department individually. The lower overall rate at which female applicants are predicted to be adequately qualified can be explained by their tendency towards choosing physics over business[3]. This is represented in the causal graph in Figure 3.1b. Intuitively, it does not seem that this predictor is biased against women.

A criterion of fairness should be able to distinguish between Scenarios 1 and 2. This means, in Scenario 1, the criterion should be violated, whereas in Scenario 2, it should be satisfied. We can consider the ability to make this type of distinction between cases a first challenge for any reasonable candidate fairness criterion:

Challenge 1 Distinguish between Scenario 1 and Scenario 2. That is, Scenario 1 should be categorized as *unfair*, and Scenario 2 as *fair*.

Let us now examine how different fairness criteria handle these two scenarios. We will begin with the statistical fairness criterion *equalized odds*, which we already discussed in previous chapters. Equalized odds requires that the prediction be independent of the protected characteristic conditional on the actual value of the target variable. For a binary predictor, this means that the false positive and the false negative error rates should be equal across all

[3]For a more detailed description of this scenario, see Appendix A.

protected groups. Now, in both above scenarios, the false negative rate is 0.2 for male, and 0.28 for female applicants, while the false positive rate is 0 for both[4]. According to equalized odds, both scenarios therefore count as unfair. This is so because the error rates are determined solely by the frequencies of different outcomes represented in Table 3.1. Since these frequencies are identical in Scenarios 1 and 2, and the only difference is in the causal mechanism generating them, the error rates in the scenarios are necessarily identical. Consequently, equalized odds cannot distinguish between Scenario 1 and 2 and hence fails on Challenge 1.

To contrast this statistical fairness criterion with a causal criterion, we will next consider counterfactual fairness. Recall that, informally stated, the idea behind counterfactual fairness can be described as follows. In a fair prediction, the prediction would still have been the same, had the protected characteristic been different. Put differently, to assess whether a prediction is fair, we need to compare an actual prediction with the prediction the same individual would have received in a counterfactual world in which their protected characteristic is different. Applying this to the above examples, we can look at, say, a prediction for a female applicant. Assume the prediction was negative, predicting that the applicant would not be suitable to study at the university. Now, to check whether this prediction is fair, we have to consider the counterfactual world in which the same individual was a man, and check whether the prediction would be different[5]. If the prediction remained the same, it is fair, if not, it is unfair. If every prediction is fair, the predictive model can be considered fair.

It is easy to see that this criterion will not give identical verdicts in Scenarios 1 and 2. In Scenario 1, we know that gender has a direct causal influence on the prediction. This implies that there are cases where the negative prediction a female applicant received would have been positive if the applicant had been male. According to counterfactual fairness, the predictive model in Scenario 1 would hence be considered unfair.

In Scenario 2, on the other hand, the situation is different. The only variables that influence the prediction are an applicant's high school performance and the department they apply to. So, whatever the prediction is that a given applicant receives, it is clear that it would be the same even if their gender

[4]See the calculation in appendix A.

[5]Precisely speaking, we have to check whether the *probability* of the prediction would be different.

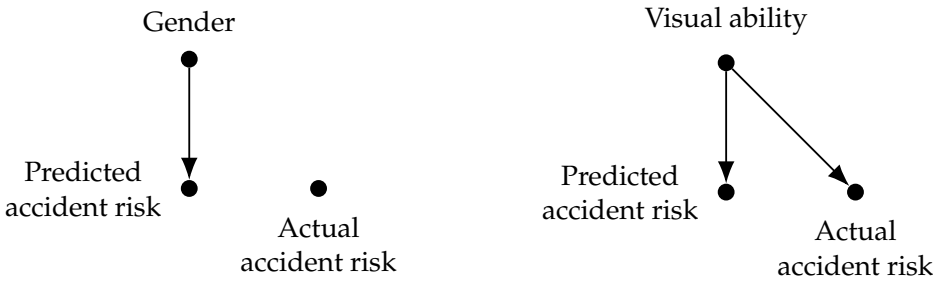
had been different. That is because even though gender is correlated with the department the applicant chooses, gender does not causally influence the choice of department. Hypothetically changing an applicant's gender can thus not influence the prediction. The predictive model in this scenario would hence be considered fair. It is of course important to emphasize that this evaluation is based on the assumption that the causal graph in Figure 3.1b represents the true causal structure of the situation. But once this assumption is granted, we know with certainty that the predictor in Scenario 2 cannot violate counterfactual fairness.

Contrary to equalized odds, counterfactual fairness can hence distinguish between scenario 1 and 2, and consequently meets Challenge 1. However, there is another, different pair of cases that a reasonable candidate fairness criterion should be able to distinguish. These are cases in which the protected characteristic makes a difference to the outcome that is to be predicted. In one of the cases, it intuitively seems that it is morally problematic to take the protected characteristic into account because the protected characteristic is irrelevant to the prediction. In the other case, however, the protected characteristic is relevant, and it seems intuitively permissible for the protected characteristic to make a difference to the outcome. The following two examples help to illustrate this.

Scenario 3: A predictive algorithm is used to estimate a given driver's risk of a car accident. The prediction is based (among other variables, which we will ignore for the moment) on the driver's gender. On average, being female results in an increased risk estimation. As being female does not increase the actual risk of an accident, it intuitively seems that this predictor is biased against women. This scenario is represented in the causal graph in Figure 3.2a, and the confusion matrix in Table 3.2.

Scenario 4: A predictive algorithm is used to estimate a given driver's risk of a car accident. The prediction is based (among other variables, which we will ignore for the moment) on whether the driver is visually impaired or not. On average, being visually impaired results in an increased risk estimation. As we can assume that being visually impaired does in fact increase the actual risk of an accident[6], it intuitively seems that this predictor should not be deemed to be biased against visually impaired people. This scenario is represented in the causal graph in Figure 3.2b and in Table 3.3.

[6]See, e.g., Anstey et al. (2012).



(A) Scenario 3 (B) Scenario 4

FIGURE 3.2: Two different scenarios in which the protected characteristic (gender in (A), visual impairment in (B)) has a causal influence on the prediction.

	Male		Female		
	$Y=1$	$Y=0$	$Y=1$	$Y=0$	Total
$\hat{Y}=1$	50	0	50	50	150
$\hat{Y}=0$	0	50	0	0	50
Total	50	50	50	50	

TABLE 3.2: Confusion matrix representing the performance of a hypothetical predictive algorithm in Scenario 3.

Looking at the two causal models, we see that in Scenario 3, the protected characteristic, gender, does causally influence the predicted risk of a car accident, but it does not causally influence the actual risk of a car accident[7]. In contrast, in Scenario 4, the protected characteristic, visual impairment, does causally influence both, the predicted as well as the actual risk of a car accident.

	Good eyesight		Visually impaired		
	$Y=1$	$Y=0$	$Y=1$	$Y=0$	Total
$\hat{Y}=1$	50	0	100	0	150
$\hat{Y}=0$	0	50	0	0	50
Total	50	50	100	0	

TABLE 3.3: Confusion matrix representing the performance of a hypothetical predictive algorithm in Scenario 4.

We can imagine an experimental setting in which this is empirically investigated. In the first experiment, a group of 100 men and 100 women are asked

[7]Note, that this is a hypothetical assumption. A number of studies seem to indicate that male drivers have a systematically higher risk of car accidents, see, e.g, Arnett (2002) and Simon and Corbett (1996). However, the evidence is ambiguous (Classen et al., 2012).

to drive a car for a set amount of time. Whenever a driver has an accident, this is recorded. For simplicity, we will assume that $Y = 1$ stands for a driver having at least one car accident, while $Y = 0$ stands for no accidents. In the second experiment, a group of 100 people with good eyesight and 100 visually impaired people are asked to do the same. Again, accidents are recorded. The outcomes of these two hypothetical experiments are depicted in the two confusion matrices in Tables 3.2 and 3.3.

The difference in underlying causal mechanisms shows up in the hypothetical confusion matrices corresponding to the two experiments. The last row of the confusion matrix for Scenario 3 (Table 3.2) indicates that there are equally many men and women having at least one accident, namely 50 each, while 50 do not have any accidents. Nonetheless, all women are predicted to have at least one accident, while only 50 men are predicted to have at least one accident. The confusion matrix for Scenario 4 (Table 3.3), on the other hand, indicates that among the people with good eyesight, 50 have at least one accident, while 50 have no accidents. Among the people with visual impairments, all 100 have an accident. All outcomes are correctly predicted. Scenarios 3 and 4 allow to formulate the second challenge for any reasonable candidate fairness criterion:

Challenge 2 Distinguish between Scenario 3 and Scenario 4. That is, Scenario 3 should be categorized as *unfair*, and Scenario 4 as *fair*.

Let us now examine how different fairness criteria deal with this second challenge. Again, we will begin with the statistical fairness criterion equalized odds. To do so, we again have to compare false positive and false negative error rates for the different protected groups, as the predictions are binary. In Scenario 3, the false positive as well as the false negative rate for male drivers is 0. This can be directly read off from Table 3.2: all male drivers that do actually have at least one accident ($Y = 1$) are predicted to do so ($\hat{Y} = 1$), while all who do not have an accident ($Y = 0$) are predicted to not have accidents ($\hat{Y} = 0$). The false positive rate for female drivers, however, is 1: all female drivers who do not actually have accidents ($Y = 0$) are nonetheless predicted to have at least one accident ($\hat{Y} = 1$). Consequently, the criterion equalized odds would evaluate Scenario 3 as unfair. Scenario 4, on the contrary, would count as fair according to equalized odds: the false positive as well as the false negative rate is 0 for both protected groups, the people with good eyesight and the visually impaired. This is easily read off from Table 3.3. So,

equalized odds meets Challenge 2.

How about the causal fairness criterion counterfactual fairness? From the causal graph for Scenario 3, we can infer that there are at least some cases for which a change in gender will effect a change in predicted accident risk. This follows from our definition of causation (Pearl, 2009, p. 46). Hence, Scenario 3 would, as required, be categorized as unfair. However, the same is true for Scenario 4: since there is a causal link from visual ability to predicted accident risk, there are at least some cases for which a change in visual ability will effect a change in predicted accident risk. Even though this seems intuitively acceptable, it violates counterfactual fairness. Hence, Scenario 4 would also be categorized as unfair. Counterfactual fairness fails on Challenge 2, as it cannot distinguish between Scenarios 3 and 4, even though they seem intuitively different in morally relevant ways.

It seems that neither equalized odds nor counterfactual fairness can meet both Challenges 1 and 2. And in fact, this result can be extended to all the currently popular statistical as well as causal fairness criteria. The reason for this is obvious: since statistical criteria[8] only take the probability distributions over certain variables into account, they necessarily fail at distinguishing scenarios which are alike in terms of their relevant variables' probability distributions, but differ in underlying causal mechanisms. Therefore, any statistical fairness criterion will fail on Challenge 1. Current causal criteria, on the other hand, typically only take into account whether there is an (indirect) causal link from the protected characteristic to the prediction of the target variable[9]. They often ignore, however, the relationship between the prediction of the target variable and its actual value. Any such causal fairness criterion will necessarily fail on Challenge 2.

It seems that in order to meet both challenges, a fairness criterion needs to be sensitive to two things, namely (1) the nature of the causal mechanism linking target variable and prediction, and (2) the relationship between prediction and actual value of the target variable. As we will see, the fairness criterion which I motivate and define in this chapter does this.

[8] Examples of such statistical criteria include *demographic parity* and its variants (Dwork et al., 2012; Darlington, 1971; Feldman et al., 2015), *predictive parity* and its variants (Cleary, 1966; Berk et al., 2021; Chouldechova, 2017), *equalized odds* and its variants (Hardt et al., 2016; Zafar et al., 2017).

[9] Examples of such causal criteria include *counterfactual fairness* (Kusner et al., 2017) and the *"no proxy discrimination"-criterion* (Kilbertus et al., 2017).

3.3 What is wrongful discrimination, anyway?

Before providing a formal definition of the new fairness criterion, let us begin by explicating the notion of wrongful discrimination. This is important, since the aim of predictive algorithmic fairness criteria is to ensure that by using a given predictive model one does not wrongfully discriminate against specific individuals.

I will here broadly follow the conceptual analysis provided by Eidelson (2015, pp. 13ff). On Eidelson's account, discrimination occurs when a person is treated differently from another person in some respect, where this differential treatment constitutes a comparative disadvantage for the person, and occurs on the grounds of a perceived difference between the discriminatee and some other person. Eidelson does not treat the concept of discrimination as a normative concept. This means that the above description is agnostic with regard to the question when and why discrimination is morally wrong. Eidelson does, however, offer a separate theory regarding this latter question. What distinguishes wrongful discrimination from permissible discrimination is, in his view, that in an act of wrongful discrimination, the discriminator *fails to respect the discriminatee's standing as a person*. Let us discuss the different aspects of wrongful discrimination in some more detail.

The first aspect of wrongful discrimination is what Eidelson calls the *differential treatment condition*. In an act of wrongful discrimination an individual is treated differently from some other relevant comparison individual. This differential treatment imposes a relative disadvantage on the individual discriminated against. This can mean that some harm is imposed on the discriminatee but not on the comparison individual, that the discriminatee is denied some basic right that the other individual enjoys, or that access to a good that everyone is equally entitled to is not granted. The comparison individual should be from within the same organizational structure that is governed by a unified set of normative principles (e.g. society, country, or company). If, for example, a particular woman is denied a right to vote because she is a woman, then the comparison individual should be a male person from the same country, rather than, for instance, another woman from a different country. The comparison individual need not be an actual individual. In some cases, differential treatment in comparison with a hypothetical individual is sufficient for an act to constitute discrimination.

The second aspect of wrongful discrimination is an *explanatory relation* between a perceived difference (between discriminatee and the comparison individual) and the differential treatment. In other words, the fact that the discriminatee is perceived to be different in some respect from the comparison individual explains that the discriminatee is treated differently. This explanatory relation can take different forms. The most blatant form would be if the perceived difference, for instance in gender, provided a motivating reason for the discriminator to treat one person worse than another. However, the explanatory relation can also be more subtle. There need not be a conscious, malevolent motive behind the differential treatment. Sometimes, an implicit bias held by a decision-maker might lead to differential treatment, or institutions might be set up in a way such that having a specific trait results in some disadvantage. At the most general level, we can say that discrimination involves a causal link from the trait in question (which the discriminatee possesses while the comparison individual does not) and the differential treatment. The existence of this causal link implies that the differential treatment can be at least partially explained by the difference between discriminatee and comparison individual.

Let us return to the example mentioned above, namely that a specific woman is denied the possibility to cast a vote while some suitable comparison individual is allowed to vote. This can only be considered an act of gender discrimination if this differential treatment can be explained by the fact that she is a woman. In 19th century Britain, for instance, we find such cases. During that time, women were by law excluded from voting. This means, in 19th century Britain the fact that a person was a woman explained why they were not allowed to vote. Contrary to this example, if a 16-year-old woman in contemporary Britain is denied the possibility to vote because of her young age, while the 19-year-old male comparison individual is allowed to vote, this would not constitute a case of gender discrimination. In the latter case, there is no causal link from the fact that the person in question is a woman (while the comparison individual is not) and the difference in voting rights (between discriminatee and comparison individual). The difference in voting rights is fully explained by the age difference.

While Eidelson does not take it to be necessary that the trait on the grounds of which a person is treated differently be a so-called protected characteristic, a majority of theories of discrimination do (see, e.g., Holroyd, 2017, p. 384; Lippert-Rasmussen, 2014, p. 25;Fredman, 2011, p. 154). We will here diverge

from Eidelson in that we will also presuppose that in cases of wrongful discrimination, the differential treatment is explained by a difference in some protected characteristic, such as gender, disability, ethnicity, age, and so on. How exactly protected characteristics are to be defined will be left open.

The last aspect of wrongful discrimination is what we will call the *wrongfulness condition*. Not all acts of differential treatment constitute wrongful discrimination, even if a perceived difference in protected characteristics explains the differential treatment. Think, for instance, of a blind person who is denied a driver's license. Here, a protected characteristic — being visually disabled — explains the fact that the person is treated disadvantageously as compared to someone without this disability (in that they are excluded from the opportunity to drive a car). Nonetheless, I assume most people would be hesitant to consider this a case of wrongful discrimination. But, what is it that makes some cases of discrimination wrong and others not? Indeed, it is a much debated question under which conditions exactly an act of discrimination is wrong, and a wide variety of views on it have been expressed (see, e.g., Moreau, 2010; Alexander, 1992; Lippert-Rasmussen, 2014; Halldenius, 2017).

The condition under which discrimination is wrongful, according to Eidelson, is that it constitutes a failure to respect the discriminatee's standing as a person. Respecting someone's standing as a person, in turn, involves two aspects: (1) recognizing that everyone is of equal moral worth, and (2) treating a person as an individual. We will here only focus on the latter, as this seems to be well suited to the domain of algorithmic predictions. While it is possible to conceive of a way in which an algorithmic prediction fails to treat the person about whom a prediction is made as an individual, it is harder to imagine how a prediction could succeed or fail to recognize a person's moral worth. By definition, predictions do not involve any such value judgments.

What does it mean to treat a person as an individual? Here, too, different answers can be given. Eidelson thinks that the intention to treat people as individuals restricts the ways in which one can form generalizations about groups of people, and hence draw inferences about individual members of the group (Eidelson, 2015, p. 142). More specifically, treating someone as an individual imposes two requirements on generalization-based judgments about a person. First, it requires that in the process of arriving at a judgment, adequate weight is given to "evidence about the ways [the person] has

exercised her autonomy in giving shape to her life, where this evidence is reasonably available and relevant to the determination at hand" (Eidelson, 2015, p. 144). Secondly, it requires that no judgment is "made in a way that disparages [the person's] capacity to make [...] choices as an autonomous agent" (Eidelson, 2015, p. 144) in the case that judgments are concerned with the person's choices.

To illustrate this, imagine a landlord who judges on the basis of an applicant's religion that the applicant will not pay their rent reliably, despite the fact that the applicant provides evidence of a secure job and references from previous landlords. The landlord generalizes from the supposition that people who have the applicant's religion do not pay their rent reliably. In doing so, the landlord clearly fails to acknowledge evidence about how the applicant has autonomously shaped their life, for instance in terms of what kind of career to pursue, which evidences reliability. Furthermore, the landlord ignores evidence of the choices the applicant made in the past, in particular with regard to paying the rent reliably, which was readily available in the form of references from previous landlords. Therefore, this act is an instance of a decision that is based on a wrongful generalization.

Generalizations thus fail to respect a person's individuality when weight is given to evidence in inadequate ways. This is, more specifically, the case when not enough weight is given to evidence about a person's character traits, where these character traits are relevant to the property that is to be predicted (e.g. reliability), or not enough weight is given to the person's choices, where these choices are relevant to the predicted property (e.g. criminal behaviour). Instead, too much weight is given to a person's membership in a demographic group (i.e. a protected characteristic), despite this not being directly (or only to a lesser extent) relevant to the predicted property.

Consequently, treating someone as an individual imposes a duty on the decision-maker to give the right weight to all the relevant factors in making a judgment about a person. This, in particular, entails that a person's protected characteristics should only influence a judgment to the degree to which they are actually relevant. If a protected characteristic influences the judgment to a higher degree, then this means that not enough weight is attributed to the person's relevant character traits or the person's relevant choices as an autonomous agent.

We hence arrive at a definition of *wrongful discrimination* against an individual i_1 which involves the conjunction of the following three conditions:

- **Differential Treatment Condition:**
  Individual i_1 is treated less favourably in respect of some dimension W than some other actual or counterfactual individual i_2
- **Explanatory Condition:**
  A (perceived) difference between i_1 and i_2 with regard to the protected characteristic A figures in the explanation of this differential treatment
- **Wrongfulness Condition:**
  This differential treatment on the basis of A constitutes a failure to treat i_1 as an individual. In particular, the differential treatment is based on a judgment where the influence of i_1's protected characteristic exceeds its relevance

We will now turn to the question how this definition can be formalized so that it can be applied to predictive models.

3.4 Causal relevance fairness

The new causal fairness criterion I will now introduce is called *causal relevance fairness*. It can be thought of as a modification of the popular counterfactual fairness criterion that is more closely in line with the informal definition of wrongful discrimination outlined in the previous section. The central task in this section is to formalize the different aspects of this definition in order to render it applicable to algorithmic systems. To do so, I will make use of Pearl's causal modelling framework (Pearl, 2009, Ch. 7).

As previously, we shall, for the sake of simplicity, assume that the target variable Y, the prediction $\hat{Y}$, as well as the protected characteristic A are all binary variables. While there are, of course, variables representing protected characteristics that have more than two values (like for instance *ethnicity*), it is always possible to represent a multi-valued variable as a set of binary variables. Hence, this modelling assumption can be made without loss of generality.

Let us begin with the *differential treatment* aspect. The first question we need

to answer is which type of disadvantageous treatment we are actually concerned with. Since we are here thinking about predictive algorithmic models, the straightforward answer is that the disadvantageous treatment consists in the prediction or classification the discriminatee received in some context. More formally speaking, this means that for some prediction $\hat{Y} = \hat{y}$, the probability of receiving this prediction is different for the discriminatee i_1 (who is characterized by the values of the input variables $\mathbf{X}^{(1)}$) than for the comparison individual i_2 (who is characterized by $\mathbf{X}^{(2)}$). Hence, the formal version of the differential treatment conditions is the following:

$$P(\hat{Y}^{(1)} = \hat{y} \mid \mathbf{X}^{(1)} = \mathbf{x}_1) \neq P(\hat{Y}^{(2)} = \hat{y} \mid \mathbf{X}^{(2)} = \mathbf{x}_2), \tag{3.1}$$

for some $\hat{y}$ in the domain of $\hat{Y}$. Note that the variables $\hat{Y}^{(1)}$ and $\hat{Y}^{(2)}$ represent the predictions for individuals i_1 and i_2, respectively. We will only use indices when comparing probabilities for different individuals, but drop them in equations that are only concerned with one individual.

Next, we have to think about how the *explanatory condition* can be formalized. Recall that we concluded that in order for there to be an explanatory relation between protected characteristic $A = a$ and the differential treatment (with regard to prediction $\hat{Y} = \hat{y}$), there must be a causal link from A to $\hat{Y}$. More precisely speaking, in the specific prediction we're concerned with, it needs to be the case that the fact that $A = a$ (rather than $A = \neg a$, where $\neg a$ indicates the only other possible value of binary variable A) is the *actual cause* of $\hat{Y} = \hat{y}$. Following Kusner et al. (2017), the actual causal effect can be defined as the difference between two quantities: first, the probability that the prediction takes the value that it actually takes (conditional on the fact that the protected characteristic takes its actual value, together with some circumstantial background conditions), and second, the probability that the prediction would have been the same had the protected characteristic been different from its actual value (conditional on the same circumstantial background conditions). Formally, the explanatory condition can hence be expressed as stating that for given $a \in D_A$, $\hat{y} \in D_{\hat{Y}}$ and $\mathbf{x} \in D_{\mathbf{X}}$:

$$P(\hat{Y}_{A=a} = \hat{y} \mid \mathbf{X} = \mathbf{x}, A = a) - P(\hat{Y}_{A=\neg a} = \hat{y} \mid \mathbf{X} = \mathbf{x}, A = a) > 0 \tag{3.2}$$

A few remarks are in order. The left expression of the subtraction is the probability that the discriminatee receives the prediction $\hat{y}$ when the value of the protected characteristic A is set (via intervention) to its actual value a in the causal model, conditional on the input values $\mathbf{X} = \mathbf{x}$. While this, for conceptual continuity with the definition of counterfactual fairness, is expressed as an interventional probability, its value coincides with the conditional probability of receiving prediction $\hat{y}$ given $\mathbf{X} = \mathbf{x}$ and $A = a$. The expression on the right-hand side of the subtraction, on the other hand, is a counterfactual probability, namely the probability that the prediction would have been $\hat{y}$, had the same individual (characterized by $\mathbf{X} = \mathbf{x}$ and $A = a$) had a different value for their protected characteristic, namely $\neg a$ instead of a. To illustrate this with an example, we could, for instance, consider the probability that a specific female loan applicant would receive a prediction that she will default, and subtract from it the probability that she would have received a prediction to default had she been male. The difference between the two probabilities can be interpreted as the actual causal effect of the individual's gender on the prediction[10]. If this value is greater than 0, this means there exists such an actual causal effect, and that, in turn, being female (at least partially) explains why the individual received the prediction they in fact received.

To express the foregoing more concisely, let us introduce the formal concept of a variable's *influence* on a prediction. This should be understood as the actual causal effect of the variable taking a specific value on the prediction $\hat{y}$. We can hence define the influence of protected characteristic $A = a$ on the prediction $\hat{Y} = \hat{y}$ as follows:

$$I(a, \hat{y}, \mathbf{x}) := P(\hat{Y}_{A=a} = \hat{y} \mid \mathbf{X} = \mathbf{x}, A = a) - P(\hat{Y}_{A=\neg a} = \hat{y} \mid \mathbf{X} = \mathbf{x}, A = a) \tag{3.3}$$

Using this concept, we can now express the explanatory condition as the protected characteristic having some influence on the prediction. This means,

[10]Note that we here only consider *positive* causal effects, that is causal effects that are greater than (or equal to) zero. To exhaust the logical space, we would, of course, also have to consider the possibility of negative causal effects. In the present example, this would be a case where being female reduces the probability of a default prediction. While the present account could easily be extended to such cases as well, it would make the formalism significantly more complicated. As the central point of this chapter is a conceptual one, I chose to give priority to simplicity over completeness and ignore these cases.

the influence of the protected characteristic on the prediction is greater than 0:

$$I(a, \hat{y}, \mathbf{x}) > 0 \tag{3.4}$$

Let us now turn to the *wrongfulness condition*. We here need to formalize the notion that an individual's protected characteristic has a greater influence on the prediction than relevant for the prediction. We already formalized the concept of *influence*. In order to formally express the wrongfulness condition, we in addition need to formalize the concept of *relevance*. Similarly to the former, a variable's relevance for a prediction can be interpreted in terms of causal relations. Pearl himself defines a variable's *irrelevance* to some other variable as the absence of any causal effect of the former on the latter (Pearl, 2009, p. 235). Broadly inspired by this definition, I will define the degree of relevance of the protected characteristic to the prediction as the actual causal effect of the protected characteristic on the value the target variable takes (that is, the variable whose value is to be predicted). Note, however, that, strictly speaking, we're here concerned with a different relation than Pearl, since instead of defining relevance as a relation between variables, we are here defining it as a relation between propositions (expressed as variables taking specific values). Formally, the degree of relevance of protected characteristic $A = a$ on the prediction $\hat{Y} = \hat{y}$ can be defined as follows:

$$R(a, \hat{y}, \mathbf{x}) := P(Y_{A=a} = y \mid \mathbf{X} = \mathbf{x}, A = a) - P(Y_{A=\neg a} = y \mid \mathbf{X} = \mathbf{x}, A = a) \tag{3.5}$$

As above, this is the difference between two interventional probabilities: the probability of the target variable taking the value that it actually takes if the protected characteristic is set to the value it actually has, and the probability that the target variable would take this value were the protected characteristic different. To illustrate this, we can consider the degree of relevance of a person's gender for predictions as to whether the person will default on their loan. To calculate this degree of relevance, we have to compare (1) the probability that a person is predicted to default with (2) the probability that they would be predicted to default had their gender been different. Presumably, the difference will be close to zero, as it seems implausible to think that a

person's gender has a strong effect on whether they will pay back their loans or not.

With these two formalizations at hand, we can formally define the wrongfulness condition for algorithmic predictions. In accordance with the outline in the previous section, we can define the wrongfulness condition as the circumstance that a person's protected characteristic has a greater influence on a prediction than would be warranted by its relevance:

$$I(a, \hat{y}, \mathbf{x}) > R(a, \hat{y}, \mathbf{x}) \tag{3.6}$$

On the basis of what was outlined above, we can summarize that a prediction can be considered to constitute wrongful discrimination if and only if the following set of conditions is satisfied:

- **Differential Treatment Condition:**
  $P(\hat{Y}^{(1)} = \hat{y} \mid \mathbf{X}^{(1)} = \mathbf{x}_1) \neq P(\hat{Y}^{(2)} = \hat{y} \mid \mathbf{X}^{(2)} = \mathbf{x}_2)$
- **Explanatory Condition:**
  $I(a, \hat{y}, \mathbf{x}) > 0$
- **Wrongfulness Condition:**
  $I(a, \hat{y}, \mathbf{x}) > R(a, \hat{y}, \mathbf{x})$

It is interesting to note that due to the way we formalized the three conditions, the following logical relations hold between them. Since by assumption we only consider positive causal effects, the degree of relevance $R(a, \hat{y}, \mathbf{x})$ can never be below zero. This, in turn, means that the wrongfulness condition implies the explanatory condition. Whenever the influence of a protected characteristic on a prediction is greater than its relevance, it must be the case that the protected characteristic partially explains the prediction. This is so because we defined that a prediction is (at least partially) explained by some proposition if the proposition has some influence on the prediction. If we assume that in the differential treatment condition we compare the discriminatee to a hypothetical individual which is identical in every aspect except the protected characteristic (and what is causally influenced by the protected characteristic), then the explanatory condition implies the differential treatment condition. This simply follows from our definition of actual causal effects, which forms the basis of the concept of explainability.

The upshot of this is that the wrongfulness condition cannot be satisfied without the differential treatment and the explanatory condition being satisfied as well. So, whenever either the differential treatment condition or the explanatory condition is violated, the wrongfulness condition will also not be satisfied. It can, however, be the case that the differential treatment and the explanatory condition are satisfied without the wrongfulness condition being satisfied. In order to ensure that a prediction does not constitute wrongful discrimination according to our relevance-based definition, it is consequently enough to ensure that the wrongfulness condition is not satisfied. This insight suggests a straightforward way of defining a new causal criterion for predictive algorithmic fairness, which formalizes the idea that for a prediction to be fair, the influence of the protected characteristic should not exceed its relevance:

Definition 3.4.1 (Causal relevance fairness)**.** A predictive model satisfies *causal relevance fairness* (relative to $A = a$) if (and only if) for all $\hat{y} \in D_{\hat{Y}}$, all $\mathbf{x} \in D_{\mathbf{X}}$ and fixed $a \in D_A$, it is the case that

$$I(a, \hat{y}, \mathbf{x}) \leq R(a, \hat{y}, \mathbf{x}) \tag{3.7}$$

It is easy to see that this criterion is a relaxation of counterfactual fairness. Using the above conceptualization, we can express the criterion of counterfactual fairness as follows: a prediction $\hat{y}$ is counterfactually fair (relative to protected characteristic a) if for all $\mathbf{x} \in \mathbf{X}$, it is the case that $I(a, \hat{y}, \mathbf{x}) = 0$. This means, in order for a prediction to be counterfactually fair, the protected characteristic cannot have any influence on it. This condition is relaxed in causal relevance fairness in that it is only the case that the protected characteristic is not allowed to have any influence on the prediction if the protected characteristic is irrelevant to the prediction (i.e. when $R(a, \hat{y}, \mathbf{x}) = 0$). This constitutes a limiting case in which counterfactual fairness is equivalent to causal relevance fairness. If we were to assume that protected characteristics can never have any relevance to predictions, then counterfactual fairness and causal relevance fairness would agree on every case. Whenever the degree of relevance of the protected characteristic for the prediction is greater than zero, however, counterfactual fairness and causal relevance fairness can provide differing fairness evaluations.

In the next section, I will return to the scenarios from Section 3.2 to see

whether causal relevance fairness can meet the two challenges for predictive fairness criteria.

3.5 The two challenges revisited

In Section 3.2 I presented two challenges, which, I argued, any reasonable criterion of predictive fairness ought to meet. On the first challenge, equalized odds fails because it is not able to distinguish between different underlying causal mechanisms that produce identical observational data. On the second challenge, counterfactual fairness fails because it is not able to distinguish between a protected characteristic's legitimate and illegitimate causal influence on a prediction. Let us now see how our new criterion, causal relevance fairness, performs on the two challenges.

Let us begin with Challenge 1. In order to assess the two scenarios, we have to consider the causal relations not only between prediction, input variables, and the protected characteristic, but also the target variable. In this specific case, this means we have to include information on the causal relations between gender, high school performance, and whether an applicant is qualified for a university degree. It seems reasonable to assume that whether someone is adequately qualified is causally influenced by one's high school performance, but not by one's gender. Consequently, the scenario will be assessed according to the causal model depicted in Figure 3.3a.

We now need to frame the situation in Scenario 1 in terms of the concepts coined in Section 3.4. We can read off a number of things from the causal graph directly. Since there is a causal link between the protected characteristic gender and the predictions of whether an applicant is adequately qualified, we know that there can be (in particular negative) predictions on which the protected characteristic (being female) has an influence. Formally expressed, this means that

$$I(A = female, \hat{Y} = 0, X = x) > 0 \tag{3.8}$$

(the variable X here represents the only input variable *high school performance*). But, since there is no causal link from gender to the target variable, the relevance of gender on whether the person is adequately qualified is 0, i.e.

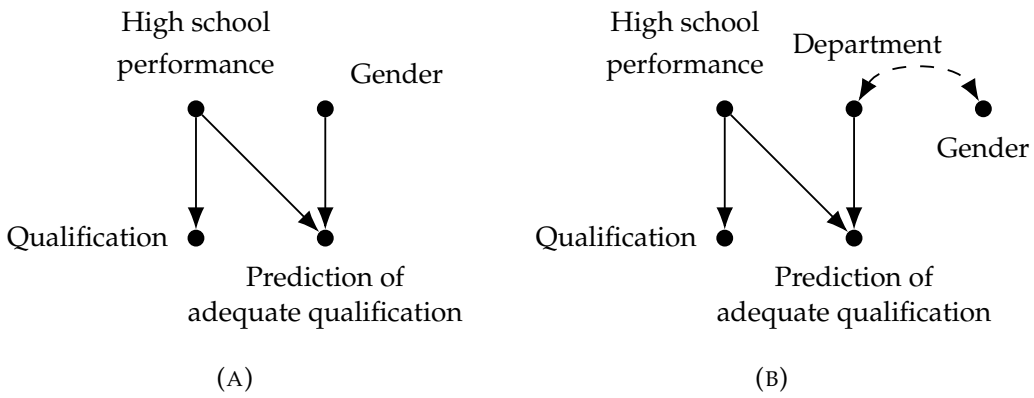


FIGURE 3.3: Two different causal models potentially producing the same joint probability distribution. This time, the target variable ("Qualification") is included in the causal models as well.

$$R(A = female, \hat{Y} = 0, X = x) = 0 \tag{3.9}$$

It then follow that for some predictions, it is the case that

$$I(A = female, \hat{Y} = 0, X = x) > R(A = female, \hat{Y} = 0, X = x), \tag{3.10}$$

and that hence causal relevance fairness is not satisfied. There are cases in which being female influences the prediction, even though being female is irrelevant here.

Things are slightly different in Scenario 2. If we look at the causal graph for this scenario, which is depicted in Figure 3.3b, we see that there is no causal link from gender to the target variable *qualification*. Being female can consequently not influence the prediction in any way, and so

$$I(A = female, \hat{Y} = 0, X = x) = 0 \tag{3.11}$$

As before, there is also no causal link from gender to the target variable, and hence

$$R(A = female, \hat{Y} = 0, X = x) = 0 \tag{3.12}$$

Causal relevance fairness is satisfied because

$$I(A = female, \hat{Y} = 0, X = x) = R(A = female, \hat{Y} = 0, X = x) \tag{3.13}$$

Being female does not influence the prediction at all, and hence it is trivially true that being female does not influence the prediction to a degree that exceeds its relevance. Challenge 1 is therefore met.

Let us now turn to Challenge 2. The causal graphs in Figure 3.2 already include the target variable, which, in this case, is the *actual accident risk*. Scenario 3 is easily analyzed: there is a causal link from gender to predicted accident risk, hence there are cases where the influence of gender on *predicted accident risk* is greater than 0, i.e.

$$I(A = female, \hat{Y} = 1, X = x) > 0 \tag{3.14}$$

Since, however, there is no causal link from gender to *actual accident risk*, gender is irrelevant to accident risk, and hence

$$R(A = female, \hat{Y} = 1, X = x) = 0 \tag{3.15}$$

Consequently,

$$I(A = female, \hat{Y} = 1, X = x) > R(A = female, \hat{Y} = 1, X = x) \tag{3.16}$$

for at least some $x \in D_X$, which means that causal relevance fairness is violated.

Before we begin to evaluate Scenario 4, let us fill in some gaps that we left open in the initial presentation of the example. As stated there, we assumed that the model we are evaluating predicts an individual's accident risk on the basis of their visual ability as well as some other factors. Let us assume we can summarize these other factors in some variable. We will assume that this variable represents all the aspects relevant for driving safely that are independent of the individual's visual ability (e.g. certain cognitive abilities, attention, risk attitudes, etc.). Nothing in the fairness evaluation hinges on

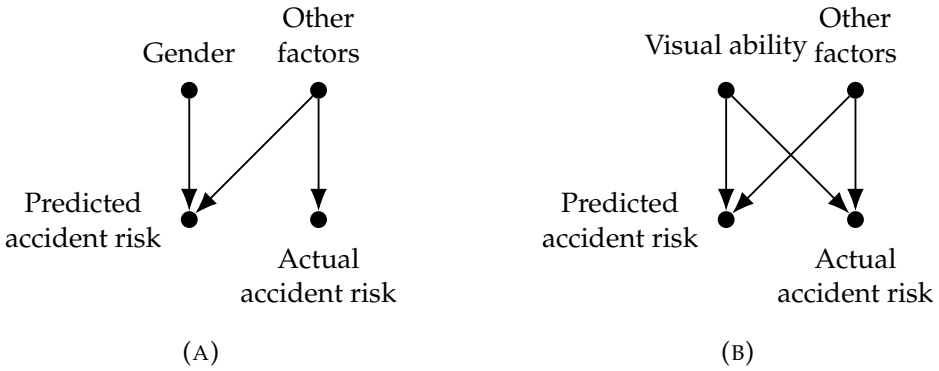


FIGURE 3.4: Two different scenarios in which the protected characteristic (gender in (A), visual impairment in (B)) has a causal influence on the prediction.

the existence of this variable, it just serves to make the example more realistic. The enriched causal model is depicted in Figure 3.4 (for the sake of completeness, an enriched causal model for Scenario 3 is depicted as well).

Analyzing Scenario 4 is a bit more complex. Here, looking at the causal graph alone does not provide enough information to determine whether the predictions the model produces are fair. In addition, we have to look at the structural equations that describe the mechanism of the scenario. In accordance with the data represented in the confusion matrix in Table 3.3, we shall assume that the following structural equation model describes the situation:

$$\begin{aligned} Y &= \neg V \vee \neg F \\ \hat{Y} &= \neg V \vee \neg F \end{aligned} \tag{3.17}$$

In the model, $\hat{Y}$ stands for the predicted accident risk, Y for the actual accident risk, V for good visual ability, and F for the other factors. We will assume that all of the variables are binary, so that we can interpret $\hat{Y} = 1$ as the prediction that the accident risk is above a certain threshold, $Y = 1$ as the risk actually being above this threshold, $V = 1$ as the person having a sufficiently good visual ability for driving, and $F = 1$ as the other factors being present at a level above what is minimally required for driving safely.

Under this interpretation, the structural equations say the following. The accident risk is above the relevant threshold if it is the case that either the

visual ability is not sufficiently good for driving, or the other factors are below the level required for driving safely. The conditions for a *prediction* of accident risk above the relevant threshold are exactly the same as for actual accident risk. For simplicity's sake, the model in this example is deterministic. By adding an unobserved background variable of which we only know the probability distribution, we could turn this example into a probabilistic one. Restricting the example to the deterministic case, however, does not lead to a loss of generality.

Based on the two structural equations, we can now calculate the quantities necessary to determine whether the predictive model in this example satisfies causal relevance fairness — the protected characteristic's influence on and relevance for the prediction. Let us assume an individual is visually impaired $(V = 0)$, but satisfies the minimally required level of other factors $(F = 1)$. Then, this individual will, according to the structural equation model above, have an accident risk above the relevant threshold $(Y = 1)$ and be predicted to be above this threshold $(\hat{Y} = 1)$.

First, we need to calculate the individual's probability of receiving the prediction they actually received, that is, $P(\hat{Y}_{V=0} = 1 \mid F = 1, V = 0)$. To this end, we need to go through the three steps necessary to calculate counterfactual probabilities[11] (Pearl, 2009, pp. 212ff). The first step (*abduction*), in which the probability distribution over the exogenous variables is updated, can be skipped because in this specific case there are no variables whose probabilities would change. Then, in the second step (*action*), we set V to 0 in the structural equation model, and, in the third step (*prediction*) evaluate how this affects the probability of $\hat{Y} = 1$ given the background information available $(F = 1)$. It is easy to see that, due to the fact that our model is deterministic, this probability is 1. Setting the visual ability to below the required level will under these circumstances always lead to a prediction of accident risk above the relevant level, because evaluating the structural equation yields

$$\hat{Y} = \neg V \vee \neg F = \neg 0 \vee \neg 1 = 1 \vee 0 = 1 \tag{3.18}$$

[11] As explained earlier, even though this probability coincides with an actual probability, here it is in fact defined as the subjunctive probability that the individual would have received the prediction, had their visual ability been set to below the threshold — which is what is actually the case. So, in some sense, the intervention which makes it a subjunctive (or counterfactual) probability does not actually change anything. We can think of this as a trivial counterfactual, where nothing disagrees with the actual facts.

Next, we need to evaluate the counterfactual probability that the individual would have received a prediction of being above the relevant risk of accident threshold, had they not been visually impaired. That is, we need to determine $P(\hat{Y}_{V=1} = 1 \mid F = 1, V = 0)$. Applying the same steps as above, we obtain a probability of 0. The hypothetical intervention on the individual's visual ability (so as to make it sufficiently good) leads to a prediction that the risk is below the relevant threshold, since

$$\hat{Y} = \neg V \vee \neg F = \neg 1 \vee \neg 1 = 0 \vee 0 = 0 \tag{3.19}$$

With these two quantities at hand, we can now calculate the influence of being visually impaired on the prediction:

$$\begin{aligned} &I(V = 0, \hat{Y} = 1, F = 1) \\ &\quad = P(\hat{Y}_{V=0} = 0 \mid F = 1, V = 0) - P(\hat{Y}_{V=1} = 1 \mid F = 1, V = 0) \\ &\quad = 1 - 0 \\ &\quad = 1 \end{aligned} \tag{3.20}$$

The influence in this specific case is 1, hence as strong as possible. In order to determine whether this is fair, we need to compare it to the relevance of being visually impaired for the prediction. The calculation here is very similar as above, and yields the following:

$$\begin{aligned} &R(V = 0, \hat{Y} = 1, F = 1) \\ &\quad = P(Y_{V=0} = 0 \mid F = 1, V = 0) - P(Y_{V=1} = 1 \mid F = 1, V = 0) \\ &\quad = 1 - 0 \\ &\quad = 1 \end{aligned} \tag{3.21}$$

As we can see, according to the causal model specified above and given the circumstantial condition that the individual otherwise satisfies the required level of ability $(F = 1)$, the visual impairment is highly relevant to the prediction. With these two quantities at hand, we can now determine whether the prediction is fair using the definition of causal relevance fairness in Equation

3.7. To this end, we have to check whether the following holds:

$$I(V = 0, \hat{Y} = 1, F = 1) = 1 \leq R(V = 0, \hat{Y} = 1, F = 1) = 1 \tag{3.22}$$

This is obviously satisfied. We can conclude that, according to causal relevance fairness, this prediction is considered fair.

To determine whether the predictive model is fair generally, this type of evaluation has to be conducted for all possible predictions $\hat{y} \in D_{\hat{Y}}$ and all possible contexts $f \in D_F$. In other words, for the predictive model to be fair, the influence of the protected characteristic has to be shown to never exceed its relevance, no matter whether the person's other factors are above or below the level minimally required for driving safely (i.e. $F = 1$ or $F = 0$), and whether the prediction is that the individual is above the relevant risk threshold or not (i.e. $\hat{Y} = 1$ or $\hat{Y} = 0$). As it turns out, for Scenario 4, this is the case. We can hence conclude that causal relevance fairness also meets Challenge 2.

3.6 Discussion

I will now turn to the discussion of a number of central points regarding causal relevance fairness. First, I will explain how causal relevance fairness can pick up on indirect causal effects of the protected characteristic on the prediction. Then I will address a methodological question about the sources of the causal knowledge required for determining whether a prediction satisfies causal relevance fairness. Lastly, I will address two potential objections, one conceptual, and one practical.

3.6.1 Detecting indirect causation

It is important to note that causal relevance fairness can pick up on indirect causal effects of the protected characteristic on the prediction. This is a relevant property, as predictive models can be unfair despite being "blinded" with regard to protected characteristics. The widely used practice to ignore protected characteristics when building predictive models is called *fairness through unawareness*.

The idea of fairness through unawareness is, more precisely, to not allow any protected characteristics as input variables of a predictive model, and to thereby prohibit that protected characteristics can influence the prediction. It was convincingly demonstrated, however, that simply turning a blind eye to protected characteristics does not generally guarantee that predictive models are unbiased[12]. The reason for this is that, especially if a predictive model has a large number of input variables, we often find that a protected characteristic is redundantly encoded in (a subset of) those variables, even if the protected characteristic itself is not part of the input variables. In other words, there might be enough information contained in the remaining input variables that allows for inferring information about a person's protected characteristics with relatively high reliability.

An analogy often invoked to illustrate this problem is the phenomenon of *redlining* (see, e.g., Allen, 2019). Redlining was a strategy used by some banks (and other financial institutions) to exclude certain ethnic groups from their services in an indirect way. The idea was to use a person's postal code as a proxy for their ethnicity, while making their decision procedures formally colorblind. Banks would, for instance, not offer mortgage loans to residents of specific neighborhoods that were known to be predominantly inhabited by Black and Hispanic residents. These banks were able to specifically target and discriminate against people on the basis of their ethnicity without explicitly using information on their ethnicity in the decision procedure.

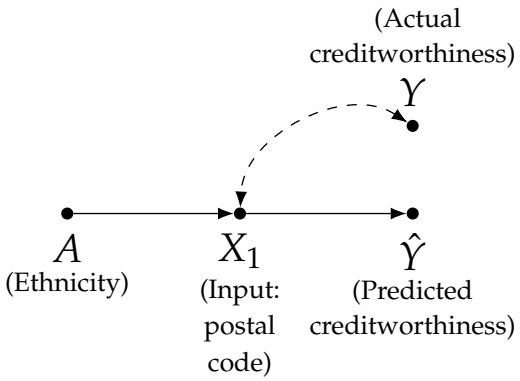


FIGURE 3.5: Indirect causal influence from the protected characteristic *ethnicity* on the prediction of creditworthiness.

A similar situation can, even inadvertently, arise in the context of algorithmic

[12]Empirical studies about the ineffectiveness of such color-blind approaches can be found in Bonilla-Silva (2006) and Apfelbaum et al. (2010). A more detailed explanation why this approach fails in predictive machine learning models can be found in Hardt et al. (2016).

predictions. Suppose a bank uses a predictive model that, among other variables, takes a loan applicant's postal code into account, but does not use any information about the person's ethnicity. The model is intended to predict whether a person is creditworthy or not. Imagine that the (incomplete) causal structure depicted in Figure 3.5 represents the causal mechanisms at work: a person's ethnicity influences where they live (i.e. their postal code), which in turn is an input variable to the predictive model and consequently influences the prediction of their creditworthiness. We can, moreover, see that postal code does not influence a person's *actual* creditworthiness, even though the two variables are correlated (indicated by the dashed bidirectional arrow). Hence, ethnicity indirectly influences the prediction of creditworthiness, but is not causally relevant to the prediction as it does not causally influence a person's actual creditworthiness.

With this predictive model, the following might happen. Imagine there are two individuals, one lives in a neighborhood that is predominantly inhabited by a specific ethnic minority, and the other in some other neighborhood that is predominantly inhabited by the majority ethnicity. Furthermore, imagine that the two individuals are very similar to each other in that they have roughly the same age, a similar education, job and salary, and so on. Now, the latter receives the prediction to be creditworthy while the former receives the prediction not to be creditworthy. Analyzing the predictive model and the causal mechanisms of the context, it might turn out to be the case that had the first person been of a different ethnicity, they would have lived in a different neighborhood, and would consequently have been predicted to also be creditworthy. If this is so, then their ethnicity explains the prediction. Ethnicity is, however, not actually relevant to whether someone is creditworthy or not. Hence, by the lights of the theory of discrimination outlined in Section 3.3, this predictor exhibits discriminatory bias: said person is treated disadvantageously due to their ethnicity, and this constitutes a failure to treat them as an individual, since, as we have stated, their ethnicity is not actually relevant to the prediction.

A fairness criterion should, of course, be able to detect indirect unjustified influences on the prediction as well. Causal relevance fairness is capable of doing this. It is easy to see that in a counterfactual world in which a person's ethnicity was different, their postal code would potentially change as well, which would in turn potentially influence their predicted creditworthiness. Their actual creditworthiness, however, would not change. So, given the

causal model we assumed, a predictor as the one outlined above would not satisfy causal relevance fairness.

3.6.2 Sources of causal assumptions

We can only evaluate whether a predictive model satisfies causal relevance fairness relative to a causal model involving the variables of interest. This includes, at the very least, a model of the causal relationship between the protected characteristic and the target variable that the predictive model is intended to predict, as well as between the protected characteristic and the prediction. The latter can be easily determined by either analyzing how the predictive model is specified, or by conducting (virtual) experiments with it. But how do we obtain the justification to make assumptions about causal relations regarding the former? This is an important methodological question that needs to be answered in order for causal relevance fairness to be useful in practice.

There are two possible approaches to this problem. The first approach is to consult external scientific resources to justify the choice of specific causal assumptions. Imagine we want to evaluate whether a predictive model for health costs is biased against elderly people. In many contexts, age is considered a protected characteristic. This means, unjustified differential treatment on the basis of age is morally impermissible. Imagine the predictive model is sensitive to age, in that, all other things being equal, elderly people are predicted to incur higher health costs than young people. In order to check whether the predictor satisfies causal relevance fairness, we can consult epidemiological and public health research on the effects of age on health and associated costs. This information can then be used to determine the degree to which being elderly explains higher health costs, and hence to which degree it is relevant to the prediction of health costs.

This first approach thus bases the evaluation of causal relevance fairness on causal assumptions which are backed by peer-reviewed social science research. This means the causal assumptions are independent of the data the predictive model is built on and applied to, and obtained in compliance with established scientific standards. This is, in general, desirable. The central

limitation of this approach, however, is that we can only evaluate causal relevance fairness for predictions of variables for which the relevant research exists. Otherwise, assessing a predictive model for fairness would require conducting social science research specifically designed for this purpose. This makes the assessment using this approach complex and costly.

The second approach is a post-hoc analysis of the predictions made. It only requires a sufficiently large set of data on the input variables, the protected characteristic, the predictions, and what the target variable's actual values turned out to be. We can then apply causal inference methods to determine whether in the given dataset the protected characteristic has a stronger causal effect on the prediction than on the target variable itself[13]. If this is the case, it indicates that the predictive model does not satisfy causal relevance fairness. Such a conclusion, however, rests on two assumptions. First, the assumption that the dataset used is representative of the overall population the predictive model is applied to, and secondly, that the dataset contains sufficient observations to allow for statistically reliable conclusions about causal relations (within the predetermined acceptable margin of error) (Wang and Ji, 2020).

To apply this approach to the above example, we would, at some later point in time, need to collect data on the actual health costs of a sufficiently large number of representative individuals that received predictions. Then we would need to run two statistical causal inference analyses to determine the causal effect of age on the predicted health costs, and the causal effect of age on the actual health costs. If the former exceeds the latter, this is an indicator that the predictive model violates causal relevance fairness.

While the second approach is more cost-effective (no new studies have to be conducted) and more universally applicable (the scope of available causal knowledge is not restricted to existing scientific studies), it is less reliable than the externally validated first approach. The reason is that the observations in the dataset might themselves be biased in several ways. It is, for instance, often the case that the target variable cannot directly be observed. In these cases, one can only compare the prediction to some proxy for the target variable, which might not be perfectly aligned with the actual target variable. Another problem is that it can be difficult to check whether the dataset is representative of the overall population. This will often require information on a number of variables which the dataset might not contain. Nonetheless, this

[13] This is the strategy pursued in the empirical case study in Chapter 5.

second approach can often provide a good first approximation of whether and to which degree a predictive model violates causal relevance fairness.

3.6.3 Risk of reinforcing existing biases

One conceptual worry about causal relevance fairness could be that it only guards against new biases, but cannot help to detect and eliminate existing societal biases. We can, for instance, imagine a school class in which most students are implicitly biased against female teachers. A majority of the students incorrectly assumes that male teachers are more competent than female teachers. Imagine that this implicit bias leads the students to be less obedient and less attentive when they are being taught by a female teacher. This, in turn, results in lower average learning outcomes when a class is taught by a female teacher than when it is taught by a male teacher. The existing implicit bias thus introduces a causal connection between gender and learning outcomes — the fact that the teacher is a woman has the effect that, on average, the learning outcomes of the school class are somewhat lower. But this, in turn, means that in this example predicting lower learning outcomes on the basis of a teacher's gender does not violate causal relevance fairness (given that the influence on the prediction does not exceed the actual causal effect on learning outcomes). At first sight, it might seem as if this indicates that causal relevance fairness allows for the reinforcement of existing biases.

This, however, is a faulty line of thinking. Recall the distinction between predictive and allocative fairness introduced in Chapter 1. While predictive fairness ensures that no discriminatory cognitive biases enter algorithmic predictive models, allocative fairness ensures that decisions based on these predictions are fair and equitable. Causal relevance fairness is a predictive fairness criterion — it constrains which evidence can legitimately influence a prediction and to which degree it can do so. Policies aimed at ensuring the fair distribution of goods and opportunities, however, are to be realized in the decision function. They hence fall into the domain of allocative fairness rather than predictive fairness. This includes, for instance, affirmative action policies aimed at correcting injustices that exist due to societal biases.

If we consider the example above in the light of this distinction, it seems fair to acknowledge the existing bias and its consequences when making a prediction — in this case that a female teacher might face resistance and disobedience from students which might result in lower average learning outcomes

— but to make decisions in a way that does not reinforce this existing bias. For instance, if the decision to be made is whom to hire for a teaching job, it might be important to base this decision not only on the predicted learning outcomes, but also on other factors beyond that prediction. We could potentially decide to apply different standards for male and female teachers to ensure an equitable distribution of job opportunities.

One could even go so far as to argue that it is *necessary* to have a predictive fairness criterion that allows for predictions that are sensitive to societal biases. Think, for instance, of a situation in which we have the same predictive model as in the example above, but where predicted learning outcomes are used to decide whether a teacher should receive special support — maybe in the form of a targeted training to deal with disobedient students, or something else along those lines. The predictions could here be used to actively counteract the existing gender bias in teaching. If a fair prediction, however, was not sensitive to causal effects due to societal biases, it would be impossible to create an algorithmic decision system which can ensure that this special support is provided to those people who, in light of the evidence, are likely to need it most. In this example, it would mean that it would not be possible to pick out the female teachers for whom special support would be most useful without violating the hypothetical fairness criterion in question. This seems undesirable.

An overall fair algorithmic decision system might hence require the ability to pick up on such causal effects of the protected characteristic on the prediction. Causal relevance fairness has this ability, while other causal fairness criteria, such as counterfactual fairness, do not.

3.6.4 Causal effects of protected characteristics

Another more practical worry about the fairness criterion presented here could be that it is defined in terms of causal effects of protected characteristics like ethnicity or gender. We need to be able to make sense of counterfactual scenarios in which we imagine that we intervene on or manipulate, say, an individual's ethnicity. But, as some have argued, it is not clear what this means or whether it is possible. If for some protected characteristics, it were not even in principle conceivable to intervene on them, causal relevance fairness would not be well-defined for these protected characteristics. Consequently, it would be unclear what it means to satisfy or violate the criterion.

A number of scientists have argued that some protected characteristics are not manipulable, and that protected characteristics hence cannot be considered or investigated as causes of anything. Holland (1986), for instance, makes this argument for an individual's race and gender, similar to Kaufman and Cooper, who extend the argument to sex, year of birth, and generally all, as they call it, "unalterable" characteristics (Kaufman and Cooper, 1999). An important premise in both arguments is that an individual could not be considered the same individual if we imagined them to have a different such characteristic. In some sense, the protected characteristic is essential to their identity. A counterfactual conditional of the form "x would be the case if the individual's protected characteristic were different" is thus meaningless.

Different lines of argument can be pursued to counter this claim. First, even if we grant that protected characteristics are not manipulable, it is often not the protected characteristic itself that we are interested in. Instead, most predictive modelling applications are typically concerned with relevant proxies of the protected characteristic. These are often specific isolated aspects correlated with or entailed by a protected characteristic, and in most cases it is easy to manipulate them. Say, we want to evaluate whether a predictive model that is used for hiring decisions is biased against certain ethnicities. The proxies for membership in a specific ethnicity that appear in the documents submitted for a job application are a person's name, their address, maybe even the writing style in which their cover letter is written. While it may in practice be difficult to figure out how exactly ethnicity influences or is correlated with all of these, it is at least clear that these variables are in principle easily manipulable.

But one could also argue against the very claim that many protected characteristics are not manipulable. An argument of this form was put forth by Malinsky and Bright (2021). Many protected characteristics, and in particular ethnicity, race, and gender, so the argument goes, can be understood as socially constructed categories. This means that they do not correspond to discrete natural divisions between different groups, but rather to contingent choices as to how to categorize human beings. Mills, for instance, argues that we have to think of racial categories as based on a continuum of phenotypical traits where the lines that demarcate one group from another are drawn in a contingent way that constitutes the outcome of a social decision rather than a biological fact (Mills, 2015, pp. 44ff). We could imagine that in a different social and historical context, the boundaries for racial categories would have

been drawn at different points. On this social constructionist understanding of racial categories, a person's racial category is not essential to their identity but is a contingent social fact. Some people who, in the actual social context fall into one category (e.g. Asian) would, in a counterfactual social context, fall into a different one (e.g. White). Such a counterfactual scenario is clearly conceivable. This, in turn, means that a person's racial category is manipulable — namely by intervening on the social context which sets the boundaries for the racial categories. While it might in individual cases be very difficult to work out what exactly such a counterfactual social context would look like, it is not in principle impossible to do so. A similar argument can be made for ethnicity and gender.

3.7 Conclusion

In this chapter I have introduced and discussed a new causal fairness criterion called causal relevance fairness. The idea underlying causal relevance fairness is that a predictor is fair only if the protected characteristic's effect on the prediction does not exceed its relevance for the prediction. I have shown that the criterion is firmly grounded in ethical theories that interpret wrongful discrimination as differential treatment on the basis of a protected characteristic, where this treatment constitutes a failure to treat the person in question as an individual. Moreover, I have shown that causal relevance fairness, in contrast to other popular fairness criteria, can meet two challenges that any reasonable criterion of algorithmic fairness should meet.

Chapter 4

Reconciling Algorithmic Fairness Criteria

4.1 Introduction

The discourse about algorithmic fairness hit a roadblock early on. Two papers independently proved that the fairness criteria *equalized odds* and *predictive parity* are mutually incompatible under most circumstances (Kleinberg et al., 2016; Chouldechova, 2017). This means, it is in most cases impossible to satisfy both — when one is satisfied, the other must be violated. At the same time, these impossibility results inadvertently provided a justification for companies, governments, and other organizations to use predictive models which violate one of the fairness criteria: they could simply argue that the model cannot but violate the criterion since it satisfies the other. To resolve this issue, it was subsequently discussed whether one of the criteria can be given up (see, e.g., Hellman, 2019; Hedden, 2021), whether it is context dependent which criterion is to be applied (see, e.g., Loi et al., 2021), or whether both criteria should generally be abandoned and supplanted by some other criterion (see, e.g., Kusner et al., 2017; Dwork et al., 2012). Yet, the two criteria equalized odds and predictive parity have some intuitive appeal that makes it hard to accept any of these options. As a consequence, there is still no consensus on how to deal with the impossibility.

In this chapter, I will argue that both criteria can be modified in a way that retains their intuitive appeal and renders them universally compatible. Instead of requiring that error rates must be equal across protected groups, I contend that we should require that the protected characteristic does not cause error rates to be different across groups. By the same token, instead of requiring

that predictive value must be equal across protected groups, I argue that we should require that the protected characteristic does not cause the predictive value to be different for different groups. To formalize these modified versions of equalized odds and predictive parity, I will use a statistical method called *matching*, which is typically used for causal inference in observational studies.

The remainder of this chapter is organized as follows. In Section 4.2, I present the Kleinberg-Chouldechova impossibility theorem, before discussing possible criticisms of the two criteria equalized odds and predictive parity. In Section 4.3, I turn away from fairness for a moment, to introduce the method of matching, which is used for causal inference in statistics. In Section 4.4, I use the matching method to define versions of equalized odds and predictive parity, which, I argue, more adequately capture the ideas underlying equalized odds and predictive parity. As I will show, the two criteria are universally compatible.

4.2 An interpretation of the Kleinberg-Chouldechova impossibility

As in the previous chapters, I will focus on binary predictions (or classifications) and take the predictive model to be a function from a set of input variables to the prediction. I will, as before, denote the variable representing the protected characteristic with A (which will also be assumed to be binary), the prediction with $\hat{Y}$, and the target variable that is to be predicted with Y. Recall the definitions of the two criteria:

Definition 4.2.1 (Equalized odds)**.** A predictive model satisfies equalized odds (relative to protected characteristics $a_1, a_2 \in D_A$) if and only if for all $\hat{y} \in D_{\hat{Y}}$ and $y \in D_Y$, $P(\hat{y} \mid a_1, y) = P(\hat{y} \mid a_2, y)$.

Definition 4.2.2 (Predictive parity)**.** A predictive model satisfies predictive parity (relative to protected characteristics $a_1, a_2 \in D_A$) if and only if for all $\hat{y} \in D_{\hat{Y}}$ and $y \in D_Y$, $P(y \mid a_1, \hat{y}) = P(y \mid a_2, \hat{y})$.

In most contexts, equalized odds and predictive parity cannot be satisfied simultaneously, as was shown by Chouldechova (2017) and Kleinberg et al. (2016). More precisely, whenever the prevalence (i.e.the probability of the target variable taking a specific value) is different for different protected groups,

a predictive model which satisfies equalized odds must violate predictive parity, and vice versa[1]. For example, a predictive model which is intended to predict whether a defendant will re-offend (i.e. commit a future crime) cannot at the same time produce equal error rates and have equal predictive values for different ethnic groups, if the prevalence of reoffence (i.e. the relative frequency of defendants committing another crime in the future) differs across these groups. To state it concisely, the theorem can be formulated as follows:

Theorem 2 (The Kleinberg-Chouldechova impossibility)**.** If the probability distribution of the target variable differs across protected groups, no (imperfect) predictive model can satisfy both equalized odds and predictive parity.

If equalized odds and predictive parity were both universally necessary conditions of predictive fairness, these impossibility results would be bad news. It would mean that it is impossible to build truly fair predictive models. There is, however, another way to interpret the impossibility: we can understand it as showing that we got something wrong in formalizing our intuitions about what makes predictive models fair. The impossibility result can then be seen as an indicator that we have to rethink the definitions of the two fairness criteria and reevaluate whether they actually formalize the intuitive ideas they are supposed to formalize.

The argument I will pursue here is along those lines. I will argue that, despite the fact that both fairness criteria have intuitive appeal, upon closer scrutiny, they turn out to be stronger than would be required in order to avoid certain types of unfairness in predictive models. I will examine both criteria in turn, beginning with equalized odds.

A reasonable interpretation of the aim behind equalized odds is that it is intended as a criterion that prevents systematic cognitive bias. Systematic cognitive bias can here roughly be understood as misjudging how informative a certain trait is in predicting another trait. Assume, for example, that in making predictions about whether someone will get lung cancer, we overestimate how informative it is that the person smokes. More precisely, if someone is a smoker, we predict that they will get lung cancer, and if not, we predict that

[1]Strictly speaking, this is only true for imperfect predictive models, that is, models that are not guaranteed to always predict correctly. Since in real-world situations no predictive model could ever be guaranteed to only make correct predictions, this restriction does not limit the significance of the impossibility result.

they will not get lung cancer. We are clearly biased with regard to smoking in predicting lung cancer: not everyone who smokes gets lung cancer, and some people get it without ever having touched a cigarette. It is easy to see that this will result in different error rates for the group of smokers and the group of non-smokers. The smokers will have a false negative rate of 0 (simply for the fact that no smoker was predicted to not get lung cancer) but a false positive error rate above 0 (some smokers do not get lung cancer). Vice versa, the non-smokers will have a false negative rate above 0 (there are some who get lung cancer, but we never predict a non-smoker to get lung cancer) but a false positive rate of 0 (because no non-smoker is predicted to get lung cancer). One could say that this model is systematically biased with regard to smoking in predicting lung cancer. If, however, instead of using the predictive model just described we used a predictive model which guarantees that the error rates across smokers and non-smokers are equal, then we could be sure that the predictor contains no such bias.

Yet, it is important to note that a violation of equalized odds across protected groups can only plausibly be understood as an *indicator* and not a *definition* of systematic cognitive bias. To see this, note that the statement relating error rates to bias is a conditional: *if* there is bias with regard to trait A, there will be disparities in error rates between those with trait A and those without. By simple logic, this implies that whenever there are no disparities in error rates, there is no bias. Yet, it does not imply that whenever we observe disparities in error rates between those with trait A and those without, we can conclude that the predictor is biased with regard to A. In other words, equalized odds relative to A is a sufficient condition for the absence of bias with regard to A, but not a necessary one. Hence, trying to deduce that a predictor is biased from the observation that error rates among groups differ amounts to committing the well-known fallacy of *affirming the consequent*. At best, observing disparities in error rates allows one to make an inference to the best explanation: when disparities in error rates between two groups are observed, and there is no other plausible explanation, then one is justified in suspecting that this is due to bias with regard to the trait that distinguishes the groups. This may in many cases be a plausible inference, but what matters for our purposes is that it is a fallible one.

To illustrate this with an example, imagine a health insurance company that tries to predict the healthcare costs an individual incurs in a given year. To

FIGURE 4.1: Two different causal models potentially producing the same discrepancy in error rates across religious groups.

simplify things, imagine the company trying to predict only whether an individual's annual costs are above a certain threshold. This allows us to represent the target variable and the prediction as binary variables. Now imagine that in country C, citizens of religion R_1 are, on average, younger than citizens of religion R_2 (we can imagine that this is due to the fact that many people of religion R_1 in C have recently immigrated, and that people generally tend to immigrate when they are somewhat younger). Suppose that, upon examination, the predictions turn out to have a higher false positive rate for people with religion R_1 than for people of religion R_2. Can we conclude that the predictive model the insurance company used is biased against people of religion R_1? The observation of different error rates does not conclusively establish this. Different explanations for this discrepancy are conceivable.

Imagine first a scenario in which the insurance company uses a predictive model which solely takes the individual's age into account, as depicted in Figure 4.1a. Now imagine further that the predictor is biased with regard to age, in that it overestimates how informative young age is of risky behavior (e.g. reckless driving or extreme sports), and hence of increased health costs. This, as I have shown above, will obviously lead to higher false positive rates for predictions of high health costs among young people. Because, on average, people of religion R_1 are younger, and the predictive model is biased with regards to age, it will produce predictions with a higher false positive rate for people of religion R_1.

It is arguable, however, that this predictive model is not biased against people of religion R_1. To see this, consider the following. Imagine that instead of C, the health insurance company operated in a different country D, where citizens of religion R_2 are, on average, younger than citizens of religion R_1 (again because people of religion R_2 are mostly recent immigrants to D). It is easy to see that here, the predictive model (even though it is exactly the same

model as above) would produce predictions with a higher false positive rate for people of religion R_2 (other than in C, where the reverse was the case). If we define bias as disparities in observed error rates, we would come to the somewhat contradictory conclusion that the predictive model is biased against people of religion R_1, but that, had the insurance company applied the exact same predictive model in a different country, the model would be biased against people of religion R_2. We can see that which religion a person has does not, in any sense, influence the predictions (or, for that matter, the error rates). It only happens to be the case that, in the given context, the predictive model works on average less well for one religious group than for another. This, I contend, should not be considered systematic cognitive bias.

Compare this with a second scenario, depicted in Figure 4.1b, in which the insurance's predictor takes a person's religion into account in order to make a healthcare cost prediction. From an observational point of view, the two predictors' performances might be indistinguishable, as they could both produce the same discrepancies in error rates between different religious groups. Yet, on a narrow understanding of what systematic cognitive bias is, only the latter can be said to be biased against people of religion R_1.

This is of course not to say that disparities in error rates among different protected groups are of no moral concern by themselves, but just that there is a conceptual difference between matters of distributive justice and those of systematic cognitive bias. It is problematic that the burdens of predictive errors fall disproportionately on one religious group, as this might lead to an unequal distribution of goods that ought to be distributed equally between groups. Nonetheless, claiming that an algorithmic decision-making system produces an unjust distribution of goods is not equivalent to claiming that the predictions its decisions are based on are biased. This is an important difference, as achieving distributive justice will most likely require different interventions than removing bias.[2]

Let us now turn to predictive parity. Here, a similar observation can be made. Imagine a medical device that tests for a specific disease. Given a person has the disease, there is a 95% probability that the test turns out positive. When applied to a person who is healthy, there is a 5% probability that the test nonetheless turns out positive. This, we can imagine, can be shown to

[2]This was discussed in more detail in Chapter 1.

robustly hold across genders. There is no difference whatsoever in the likelihood of receiving an erroneous result, no matter whether a patient is male or female. Intuitively, it seems, this medical testing device is not gender-biased.

But now imagine that the disease happens to occur more frequently in men. More specifically, we can imagine that one in every ten men has the disease, but only one in every hundred women. Then the positive predictive value, that is, the probability of actually having the disease given that one receives a positive test result is different for men and women. For men, it is roughly 68%, whereas for women it is only about 16%.[3] This means, in this intuitively fair case, predictive parity is not satisfied. But it seems that this is not due to some bias in the testing device, but just to the prevalence of the disease, which differs across genders. In other words, it is not gender that causes the difference in predictive value (since the testing device works, by assumption, equally well for a randomly chosen man as for a randomly chosen woman). So it seems that here, too, we want to distinguish between discrepancies in predictive value which are (causally) explained by gender, and discrepancies in predictive value which are due to external factors, such as differences in the prevalence of a disease.[4]

In light of these criticisms, it seems that the definitions of both, equalized odds and predictive parity, do not adequately explicate the underlying moral intuitions they were designed to capture[5]. This, in turn, could mean that the Kleinberg-Chouldechova impossibility result is not so disastrous after all. If neither equalized odds nor predictive parity are, as they are currently defined, necessary conditions for fairness, the impossibility loses its bite. There is a chance that the impossibility theorem is just an artifact of the way the criteria are defined. In the remainder of this chapter, I will examine this possibility by trying to provide modified definitions of equalized odds and predictive parity which retain all the intuitively plausible aspects of the current definitions but avoid the impossibility.

[3] See calculation in Appendix B.

[4] To my knowledge, there is only one other article that addresses the problem that discrepancies in statistical fairness metrics might, in some cases, not be due to unfair bias but to differences in the prevalence of the target variable. Eva (2022) takes this as the motivation for developing an alternative criterion of predictive algorithmic fairness for risk scoring algorithms, which he calls *base rate tracking*.

[5] *Explication* is a method which seeks to turn an informal concept into an exact, formally defined concept. For an overview of the method of explication, see, e.g., Novaes (2020).

4.3 The matching method

I will use matching — a method for causal inference on the basis of observational data (Stuart, 2010) — to define modified versions of equalized odds and predictive parity. In this section, I will explain the method.

The motivation behind matching stems from the following problem. In many scenarios, it would be useful to be able to infer whether and to what degree a given variable has a causal effect on some other variable. The "gold standard" for estimating causal effects is the so-called *randomized controlled trial* — a specific type of experimental study. Often, however, it is practically impossible or unethical to run experiments, or the only available data is observational data. Think, for instance, about studying the health effects of passive smoking on children. It would be unethical to actively expose a group of children to secondhand smoke. Yet, there might be observational data on the health of children who live in a home where at least one of the parents smokes. Matching aims to, as best as possible, replicate the properties of a randomized controlled trial for observational data, to allow for the estimation of causal effects in cases like the above, where experimental data is unavailable.

How do randomized controlled trials work? The starting point of a typical randomized controlled trial is to split the participants of the trial into two groups via random selection. Random selection of participants ensures that there are no systematic differences between the groups. If slight differences remain, these are due to chance. In other words, randomization ensures that the distribution of observed and unobserved properties is similar for both groups. One group then receives some kind of treatment — accordingly, they are called the treatment group —, while the other doesn't — the latter are called the control group. If it turns out that (potentially at some later point in time) there is a statistically significant difference in some other variable (the effect variable), it can be concluded that this must be caused by the treatment, as due to randomization there are no other systematic differences between the two groups. Randomization and the intervention on one group, but not the other, make randomized controlled trials ideal for inferring causal effects.

A typical example of a randomized controlled trial is drug testing. Say, a new drug for lowering systolic blood pressure was developed and its effectiveness has to be evaluated. A sensible way of doing this is to select a sufficiently

large group of people and randomly assign them to either the treatment or to the control group. The treatment group receives the new drug while the control group receives a placebo. After a few weeks of taking the drug, the participants' blood pressure is taken. If it turns out that the treatment group has a significantly lower average blood pressure than the control group, it can be concluded that the drug does indeed cause a lowering of the systolic blood pressure. If, however, it turns out that the treatment and the control group have similar average levels of blood pressure, it can be concluded that the drug has no effect.

To highlight why for such a conclusion a randomized controlled trial is superior to a (naive) observational study, imagine the following. The drug is given to anyone who wants it, and later the blood pressure of those who decided to take the drug is compared to the blood pressure of those who didn't take the drug. Assume we don't find any significant differences in blood pressure between the two groups — can we conclude that the drug has no effect on blood pressure? The answer is, of course, no. The reason why we observe these results is, most likely, a phenomenon called *confounding*. Confounding means that there is some other variable that influences the effect to be measured and, at the same time, affects who receives the treatment. In this example, such a confounding factor could be overweight. It is likely that people who are overweight struggle with symptoms of high blood pressure and are therefore more interested in taking a blood-lowering drug than people without any such symptoms. At the same time, being overweight often leads to higher blood pressure. Consequently, people in the (non-randomized) treatment group will most likely have a higher baseline blood pressure than people in the control group. Now, even if the drug is highly effective in lowering the high initial blood pressure of overweight individuals to normal levels, we will observe no difference in average post-treatment blood levels between the two groups. Someone who only sees the post-treatment data would be led to falsely conclude that the drug does not make a difference. Randomized controlled trials allow us to control for confounding factors like, in this case, overweight, which could influence who receives the treatment and who doesn't, because we would have roughly equal body weight distributions in the treatment and control group.

When we only have observational data, we can try to mimic randomization via matching. Instead of randomly choosing who to assign to the treatment or control group, however, the effect of randomization is supposed to be

achieved by using the observational data to create a *synthetic* control group that does not systematically differ from the treatment group on any observed or unobserved variables (other than the treatment variable). This is done as follows. Assume that the data consists of information on the treatment variable, the effect variable, and a number of other variables, which, in this context, will be called the *covariates*. For each individual in the treatment group[6], we try to find an individual in the initial control group whose covariate values are identical or as similar as possible to the covariate values of the individual from the treatment group. This individual is added to the synthetic control group. We end up with a treatment and a synthetic control group with identical or very similar distributions over the covariates. Given certain assumptions that we will get to in a moment, this allows us to conclude that any significant difference between the groups with regard to the effect variable is caused by the difference in the treatment variable.

	T	Blood pr.	BMI
Person 1	1	85	28
Person 2	1	102	32
Person 3	1	85	23
Person 4	1	75	27
Person 5	0	82	23
Person 6	0	65	21
Person 7	0	84	19
Person 8	0	118	32
Person 9	0	80	27

(A) Original

	T	Blood pr.	BMI
Person 1	1	85	28
Person 2	1	102	32
Person 3	1	85	23
Person 4	1	75	27
Person 9	0	80	27
Person 8	0	118	32
Person 5	0	82	23
Person 9	0	80	27

(B) Matched

TABLE 4.1: Post-treatment data (Unmatched on the left and matched on BMI on the right).

Let us illustrate this by applying the matching method to the above example. Table 4.2a represents a (fictitious) dataset with data on nine individuals. We have observations on whether individuals took the supposedly blood pressure lowering drug (the treatment, T), their blood pressure (measured in millimeters of mercury, abbreviated as mmHg) sometime after they started taking the drug (*Blood pr.*), and their body mass index (*BMI*). We are interested

[6]Note that the terms *treatment* and *control group* are, in the context of observational data, to be understood figuratively. No individual is actually assigned to a specific group. Rather, it is the case that for some individuals, the supposed causal property is present ($T = 1$) — these we call the treatment group —, and for others, it is not present ($T = 0$) — these we call the initial control group. To highlight the analogy to randomized controlled trials, we will stick to the terms.

in whether the drug works, i.e. whether there is a causal link from taking the drug (T) to blood pressure. If we were to follow the naive approach outlined above, we would check whether the average blood pressure of the group that took the drug (i.e. individuals for which $T = 1$) is significantly lower than the average blood pressure of the group which didn't (i.e. individuals for which $T = 0$). Doing this, we find that the average blood pressure of individuals who took the drug is 86.75 mmHg, while the average blood pressure of individuals who didn't take the drug is 85.8 mmHg.

This naive analysis of the observational data seems to indicate that the drug has the opposite effect of what we would expect — individuals who took the blood pressure lowering drug have, on average, higher blood pressure after taking the drug than individuals who did not take the drug. But, as explained above, this is a false conclusion, as it is likely that there is a confounding factor present. And indeed if we look at the BMI by group, we find that the average BMI for individuals who took the drug is 27.5, while it is only 24.4 for the others. What we are really interested in is whether the group of people that took the drug would, on average, have had a higher blood pressure had they not taken the drug. This question can be addressed using matching. If for each of the individuals who took the drug, we find one individual with similar characteristics among those who didn't take the drug, and then compare the average blood pressure, this comparison will be more meaningful. This is represented in Table 4.2b.

Note that in this context, the treatment variable is T, the effect variable is *Blood pressure*, and the only covariate is *BMI*. In order to obtain a sample that achieves to mimic randomization, we have to find a match for each individual in the treatment group with regard to BMI. We consequently obtain the following matches:

- Person 1 (BMI = 28) is matched to Person 9 (BMI = 27)
- Person 2 (BMI = 32) is matched to Person 8 (BMI = 32)
- Person 3 (BMI = 23) is matched to Person 5 (BMI = 23)
- Person 4 (BMI = 27) is (also) matched to Person 9 (BMI = 27)

Note that not every match is exact, but for each individual in the treatment group, we have taken the individual from the initial control group whose BMI is closest. Note, moreover, that we matched Person 9 to two individuals

from the treatment group. If we now compare the treatment and the synthetic control group, we find that the former has a lower average blood pressure (86.75 mmHg vs. 90 mmHg). This seems to confirm the hypothesis that the drug does in fact help lower a person's blood pressure, or, in other words, that there is a causal link between taking the drug and lower blood pressure.[7]

Let us now address a central methodological question, namely how to choose covariates. In order for matching to be as good a method for estimating causal effects as randomization, it is crucial that the set of covariates contains all the variables that influence both the causal and the effect variable (i.e. all confounding variables). This is important because it entails that there are no unobserved differences between the control and treatment group conditional on the observed covariates (Stuart, 2010, pp. 3f). The assumption that this is the case is typically called *ignorability*. In our example above, ignorability is most likely not satisfied: there could be other variables that influence both treatment and effect variable, like for instance certain pre-existing diseases, smoking, and so on. Even though we did remove some confounding bias by matching on BMI, we could still get a better estimate of the causal effect of the drug on blood pressure if we had moreover matched the data on pre-existing diseases and habits like smoking. Another important norm for choosing covariates is not to include any variables that are causally influenced by the causal variable. This might lead to an underestimation of the causal effect and thereby distort the analysis (Pearl, 2010, pp. 114-118).

Lastly, it is important to mention that there is, in fact, an entire family of matching methods and that we could have chosen a different matching method for the example above. The type of matching we implicitly used can be specified as *Euclidean distance-based 1:1 nearest neighbor matching with replacement*. However, there is a number of choices one can make when deciding between different matching methods. Distance, for instance, can be defined in different ways[8], as well as the strategy for choosing matches.[9] Moreover, there is a choice as to whether individuals from the initial control group can be used

[7]Note that this example serves the purpose of illustration only. Strictly speaking, the difference in average blood pressure in the example is not statistically significant due to the very small size of the sample. The example does, however, allow us to explain the idea behind matching in an easily understandable way.

[8]Besides Euclidean distance, one could use Mahalanobis distance, propensity scores, or a binary exact-match distance.

[9]One could choose a 1:k nearest neighbor matching method for any natural number k, where the match is the data point obtained by averaging over the k individuals, or one could choose a globally optimal matching method, etc.

more than once ("with replacement") or not. The metric we have estimated is the average treatment effect on the treated, which means we have examined whether the drug had a causal effect on the individuals in the treatment group, but not necessarily on the individuals in the control group. As the purpose of the present project is to present a conceptual point, the methodological details do not matter much. This matching method was chosen because it is the simplest method that allows us to illustrate the main idea in the most straightforward way.

4.4 Modifying the criteria

Let us now return to the attempt to modify equalized odds and predictive parity in a way that avoids the Kleinberg-Chouldechova impossibility. Both fairness criteria will be considered in turn.

4.4.1 Matched equalized odds

Let us begin by laying out the intended interpretation of the fairness criterion which is to replace equalized odds, which we will call *matched equalized odds*. Satisfying matched equalized odds means that the relevant protected characteristic has no direct effect on the predictor's error rates. For example, if a recidivism predictor satisfies matched equalized odds, then the fact that a defendant is African American does not increase the probability of receiving a false positive recidivism prediction.

How can we attempt to determine whether a predictor satisfies matched equal odds? We can frame this as a causal inference problem. We are trying to determine the causal effect of the protected characteristic on a predictor's error rates. This problem, as explained in the previous section, can be addressed using matching. We take the protected characteristic to be the treatment variable and then specify an appropriate set of covariates. We next create a matched control group, such that the treatment and control group (i.e. the two protected groups) exhibit no systematic differences other than in their protected characteristics and the predictions they receive. Then we compare the error rates of the two groups. If and only if the error rates are (roughly) equal, matched equalized odds is satisfied.

Some more words are to be said about the choice of covariates. The first point is that we can, of course, only determine whether a predictor satisfies

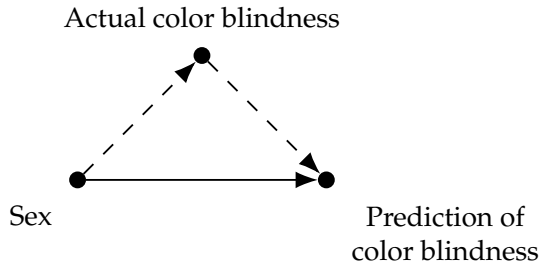


FIGURE 4.2: Direct and indirect (dashed) causal effect from *Sex* to the *Prediction of color blindness*. Matched equalized odds is concerned only with the direct causal effect.

matched equalized odds if the covariates satisfy the ignorability assumption. Any reliable evaluation of a predictor's fairness hence requires data on an appropriate amount of variables so that we can ensure that there are no unobserved differences between the two matched protected groups. The second point is that, in contrast to standard applications of matching, in which we aim to determine the total causal effect of some treatment variable on the effect variable, we are here specifically interested in the *direct effect* (Avin et al., 2005; Pearl, 2014; Weinberger, 2019) of the protected characteristic on the prediction. More precisely, we are interested in the causal effect along those causal paths that are not mediated by the target variable. In many cases, namely when the protected characteristic does not influence the target variable, this just is the total causal effect. But there are some cases where the protected characteristic does in fact influence the target variable. In those cases, we want to exclude the causal effect along this path from our analysis. Technically, this implies that we have to include the target variable in the set of covariates. This violation of the back-door criterion allows us to determine the path-specific effect of the protected characteristic on the prediction.

Why should we exclude this causal path from our analysis? We do this because, from the point of view of moral permissibility, causal effects along paths mediated by the target variable seem unproblematic. When we are trying to predict the value of a variable, the variable that is to be predicted should influence the prediction. Think, for instance, of a medical algorithm to detect whether a person is color blind. Obviously, the fact that a person is color blind *should* influence the algorithm's prediction of whether they are color blind. In the ideal case, the prediction perfectly aligns with whether the person is actually color blind. Color blindness, however, seems to be influenced by sex-specific genetic differences (Abramov et al., 2012). This, in turn, implies that in this example, there is a causal link from a person's sex to

the prediction of whether they are color blind via the target variable (which represents whether they actually are color blind). Figure 4.2 illustrates this. Since we are interested in whether discrepancies in error rates arise from an unjustified *direct* influence of the protected attribute on the prediction (which cannot be attributed to group-specific differences in the target variable distribution), we exclude the mediated causal path from our analysis. This is achieved by adding the target variable to the set of covariates.

Let us now make the above precise by giving a formal definition of matched equalized odds.

Definition 4.4.1 (Matched equalized odds)**.** A predictive model satisfies matched equalized odds (relative to protected characteristics $a_1, a_2 \in D_A$) if and only if for all $\hat{y} \in D_{\hat{Y}}$ and $y \in D_Y$,

$$F_{A=a_1}(\hat{y} \mid a_1, y) = F_{A=a_1}(\hat{y} \mid a_2, y),$$

where $F_{A=a_1}$ is the relative frequency function on a dataset obtained by applying matching to the original dataset such that $A = a_1$ indicates the treatment and $A = a_2$ the control group, and where the covariates include Y.

Let us now consider whether this definition escapes the criticism against equalized odds offered in Section 4.2. Recall that we considered the argument that equalized odds can be interpreted as a criterion against systematic cognitive bias, where this is understood as a misjudgment of how informative a given trait is for predicting another trait. I gave an example in which a predictor of health costs misjudges the informativeness of its only input variable *age*, but, by assumption, not of *religion* (which was not among the input variables). Its predictions nonetheless exhibit a higher false positive rate for people of a specific religion due to the fact that members of this religious group happen to be, on average, younger than members of other religious groups. Despite a discrepancy in observed error rates, we would not want to say that this predictor is unfairly biased against this religious group.

Matched equalized odds gets this right. In order to evaluate the predictor, we would take the data on all the individuals from the relevant religious group as the treatment group to then create a synthetic control group from the data on individuals from other religious groups. We would do this by matching them on a number of covariates, which (minimally) have to include *age* (since

this is a confounder) and *actual health costs* (the target variable). This would, consequently, result in a treatment and a synthetic control group which have equal (or very similar) distributions for the variables *age* and *actual health costs*. Since by assumption the predictor only takes *age* as an input variable, the two groups would not exhibit significant discrepancies in error rates.

We can conclude that matched equalized odds retains the intuitive appeal of equalized odds but can escape certain types of counterexamples, as the health cost prediction example just illustrated. A predictor that satisfies matched equalized odds is guaranteed to not misjudge how informative a protected characteristic is. It does, however, allow for cases in which one protected group happens to have a different observed error rate, provided this is not *because* of the protected characteristic.

4.4.2 Matched predictive parity

We can modify predictive parity in a similar manner in order to define a criterion which I will call *matched predictive parity*. The intended interpretation of matched predictive parity is that the criterion ensures that the protected characteristic does not influence the meaning of the prediction. Meaning is here understood as the confidence conveyed in a given prediction. Another way of stating this is to say that matched predictive parity ensures that the protected characteristic does not influence the predictor in a way that results in discrepancies in (positive or negative) predictive value. While this criterion does not prohibit that for a given protected group the predictions can happen to have a different average predictive value than for some other group, matched predictive parity ensures that such a difference is not due to the protected characteristic (or any of the properties affected by it).

Like before, we can frame the evaluation of whether a predictor satisfies matched predictive parity as a causal inference problem we attempt to solve using matching. As before, it is necessary that the target variable as well as all confounder variables are contained in the set of covariates. In other words, the dataset used for determining matched predictive parity is the same one used for determining matched equalized odds. We can formally define matched predictive parity as follows:

Definition 4.4.2 (Matched predictive parity)**.** A predictive model satisfies matched predictive parity (relative to protected characteristics $a_1, a_2 \in D_A$) if and only if for all $\hat{y} \in D_{\hat{Y}}$ and $y \in D_Y$,

$$F_{A=a_1}(y \mid a_1, \hat{y}) = F_{A=a_1}(y \mid a_2, \hat{y}),$$

where $F_{A=a_1}$ is the relative frequency function on a dataset obtained by applying matching to the original dataset such that $A = a_1$ indicates the treatment and $A = a_2$ the control group, and where the covariates include Y.

Can matched predictive parity handle the counterexample described in Section 4.2? We imagined a medical testing device that, independent of gender, has a 95% probability of correctly predicting that an individual has the disease (i.e. its true positive rate, *TPR*), and a 5% probability of incorrectly predicting that an individual who is actually healthy has the disease (i.e. its false positive rate, *FPR*). Assuming that the disease is much more common among men than among women, the device would produce predictions which, on average, have a lower positive predictive value (*PPV*) for women than for men. Analyzing the example, the conclusion was that the discrepancy in predictive value should not be considered discriminatory bias despite violating predictive parity, since the discrepancy is due only to the different prevalence levels of the disease in women and men. It is easy to see that this device would satisfy matched predictive parity. As a direct consequence of Bayes' theorem, we know that generally:

$$PPV = \frac{TPR * p}{TPR * p + FPR * (1 - p)}$$

where p is the prevalence of the target variable (i.e. $P(Y = 1)$). First, by assumption, TPR and FPR do not differ across gender groups. Secondly, we are looking at the relative frequencies obtained after matching on (among others) the target variable. This entails that in the matched dataset, the prevalence of the target variable is equal (or very similar) for men and for women (i.e. $p_{male} = p_{female}$). Clearly, then, the positive predictive value as well is equal for men and for women. This indicates that matched predictive parity can deal with certain types of counterexamples to predictive parity while retaining those aspects of it which are intuitively plausible. For negative predictive value, the reasoning is analogous.

4.4.3 The Kleinberg-Chouldechova impossibility revisited

Let us now return to our initial question, namely whether we can understand the Kleinberg-Chouldechova impossibility theorem as an indication that the plausible intuitions motivating equalized odds and predictive parity have been formalized in inadequate ways. If this is the right interpretation, then adequately modified versions of equalized odds and predictive parity should be universally compatible and hence escape the impossibility result. I will now analyze whether this is the case for the two modified criteria proposed above.

First, note that, assuming ideal matching conditions, we know that for the relative frequency function $F_{A=a_1}$ on the matched dataset it is always the case that $F_{A=a_1}(y \mid a_1) = F_{A=a_1}(y \mid a_2)$ (for any two groups $a_1, a_2 \in D_A$ and all $y \in D_Y$). This simply follows from the fact that we are required to include the target variable Y in the set of covariates. The matching thus results in a dataset where the distribution of Y is identical for both groups a_1 and a_2. We can now show that matched equalized odds and matched predictive parity are not only universally compatible, but they moreover turn out to be mathematically equivalent.

Proof. To see this, it will first be shown that if a predictive model satisfies matched equalized odds, it also satisfies matched predictive parity. To this end, assume we want to evaluate the predictive model relative to $a_1, a_2 \in D_A$. By the assumption of matched equalized odds, the predictive model satisfies

$$F_{A=a_1}(\hat{y} \mid a_1, y) = F_{A=a_1}(\hat{y} \mid a_2, y) \tag{4.1}$$

for all $y \in D_Y$ and $\hat{y} \in D_{\hat{Y}}$. From this assumption, together with the equality stated above, it follows that:

$$\begin{aligned} &\frac{F_{A=a_1}(\hat{y} \mid a_1, y) * F_{A=a_1}(y \mid a_1)}{F_{A=a_1}(\hat{y} \mid a_1, y) * F_{A=a_1}(y \mid a_1) + F_{A=a_1}(\hat{y} \mid a_1, \neg y) * F_{A=a_1}(\neg y \mid a_1)} = \\ &= \frac{F_{A=a_1}(\hat{y} \mid a_2, y) * F_{A=a_1}(y \mid a_2)}{F_{A=a_1}(\hat{y} \mid a_2, y) * F_{A=a_1}(y \mid a_2) + F_{A=a_1}(\hat{y} \mid a_2, \neg y) * F_{A=a_1}(\neg y \mid a_2)} \end{aligned} \tag{4.2}$$

where y and $\neg y$ abbreviate $Y = 1$ and $Y = 0$, respectively. By Bayes theorem, it follows that for all $y \in D_Y$ and $\hat{y} \in D_{\hat{Y}}$:

$$F_{A=a_1}(y \mid a_1, \hat{y}) = F_{A=a_1}(y \mid a_2, \hat{y}) \tag{4.3}$$

This is the definition of matched predictive parity. We have hence shown that matched equalized odds implies matched predictive parity. Likewise, if we assume matched predictive parity, i.e. that (given fixed a_1 and $a_2 \in D_A$), for all $y \in D_Y$ and $\hat{y} \in D_{\hat{Y}}$ the predictive model satisfies

$$F_{A=a_1}(y \mid a_1, \hat{y}) = F_{A=a_1}(y \mid a_2, \hat{y}), \tag{4.4}$$

we can, together with the assumption that

$$F_{A=a_1}(y \mid a_1) = F_{A=a_1}(y \mid a_2), \tag{4.5}$$

use Bayes theorem to show that this implies matched equalized odds. As both directions of the proof have an identical structure, the proof that matched predictive parity implies matched equalized odds is omitted.

□

The upshot of the above is that the infamous impossibility theorem loses its force when we consider adequately modified versions of equalized odds and predictive parity. From a normative point of view, the modifications can be justified by considering counterexamples where the original criteria seem too demanding (as done in Section 4.2). Modifying the two criteria accordingly not only allows us to rebut the counterexamples but moreover resolves the Kleinberg-Chouldechova impossibility. Surprisingly, under perfect matching conditions, the two criteria even turn out to be equivalent. This can be interpreted as meaning that we have found the unique and robust baseline notion of algorithmic fairness which can be strengthened in at least two (mutually inconsistent) ways.

This conclusion, however, has to be qualified in one respect. We assumed for the proof that the matching conditions are ideal and that hence

$$F_{A=a_1}(y \mid a_1) = F_{A=a_1}(y \mid a_2) \tag{4.6}$$

holds for any two groups $a_1, a_2 \in D_A$ and all $y \in D_Y$. In many realistic situations, however, the conditions for matching are not ideal. Especially if there is a big number of covariates, it is unlikely that for every individual in the treatment group one can find an exact match in the control group. But only exact matching can guarantee the above equality. Typically, the methods of choice are distance-based or propensity score matching. Both can give very good results in that the distributions over individual covariates in the treatment and control group are very similar in the matched dataset. Yet, these two methods cannot guarantee identical distributions over the covariates. The following two theorems, however, show that even though the impossibility persists under imperfect matching[10], the smaller the difference in prevalence, the smaller the trade-off between equalized odds and predictive parity. More precisely, if equalized odds is satisfied, then the smaller the difference in prevalence between the groups, the smaller the difference in positive and negative predictive value. Conversely, if predictive parity is satisfied, then the smaller the difference in prevalence between the groups, the smaller the difference in false positive and false negative error rates. Proofs for the theorems can be found in Appendix B.

Theorem 3. Suppose a predictive model satisfies equalized odds (relative to a_1 and $a_2 \in D_A$). Let Δp denote the *difference in prevalence* between two groups a_1 and a_2, and let $\Delta PPV(\Delta p)$ and $\Delta NPV(\Delta p)$ denote the functions which return the *difference in positive predictive value* and the *difference in negative predictive value* between the groups for a given value of Δp, respectively. Then $\Delta PPV(\Delta p)$ and $\Delta NPV(\Delta p)$ are both monotonically increasing in the interval $(0, 1)$.

Theorem 4. Suppose a predictive model satisfies predictive parity (relative to a_1 and $a_2 \in D_A$). Let Δp denote the *difference in prevalence* between two groups a_1 and a_2, and let $\Delta FPR(\Delta p)$ and $\Delta FNR(\Delta p)$ denote the functions which return the *difference in false positive* and the *difference in false negative error rates* between the groups for a given value of Δp, respectively. Then $\Delta FPR(\Delta p)$ and $\Delta FNR(\Delta p)$ are both monotonically increasing in the interval $[0, 1]$.

[10]Note that this is only a relaxation of the assumption that the distribution of the target variable is identical for both groups. We still assume that the ignorability assumption holds.

These two theorems hence provide evidence that the more closely matched the data is, the more closely we are approximating the underlying criteria of fairness.

4.5 Conclusion

In this chapter, I have argued that, in light of the Kleinberg-Chouldechova impossibility, the two fairness criteria equalized odds and predictive parity seem too demanding. I have further argued that using the causal inference method matching, we can modify both criteria in a way that retains their intuitive appeal but makes them compatible. I was able to show that the modified versions of equalized odds and predictive parity are not only compatible but equivalent. Moreover, I have shown that an approximate version of this result holds when perfect matching is not possible.

Chapter 5

Case Study: COMPAS Recidivism Scores

5.1 Introduction

In 2016, ProPublica, an investigative journalism organization, published an article (Angwin et al., 2016) in which the authors aimed to show that the algorithmic decision support systems used in criminal sentencing in some US states exhibit significant racial bias. In their analysis, they focussed on a tool called COMPAS, developed and distributed by the company Northpointe, which calculates recidivism risks. More specifically, it predicts the probability that a defendant will commit another crime within some time after their trial. The risk score is based on a detailed questionnaire that the defendants have to complete.

The analysis of the COMPAS algorithm found that while the overall accuracy of the predictions was roughly the same for Black (67%) and for White (69%) defendants, the two groups differed significantly in their respective error rates: the false positive rate for Black defendants was 45%, while only 23% for White defendants — indicating that Black defendants were twice as often incorrectly classified as future recidivists; at the same time, the false negative rate for Black defendants was 28%, while 48% for Whites — which means that White defendants who would actually go on to commit further crimes after trial were much more likely to nevertheless receive a low risk score. This was taken to show that COMPAS' predictions were biased against Blacks and that they could potentially result in discriminatory sentencing decisions.

Northpointe contested the claims (Flores et al., 2016), arguing that since the

algorithm can be shown to be *well-calibrated by group* (which entails predictive parity), it cannot be considered unfairly biased. In the COMPAS case, this specifically means that for both groups, Blacks and Whites, of those defendants that were assigned x% risk of recidivating by the algorithm, a proportion of roughly x% did indeed turn out to recidivate. Northpointe argued that if a predictor does not yield results that are calibrated by group, the predicted probability estimate would not have a consistent meaning across different demographic groups. This sparked the debate about the adequacy of different fairness criteria, and in particular about the question of which criterion ought to be applied in order to evaluate the COMPAS algorithm.

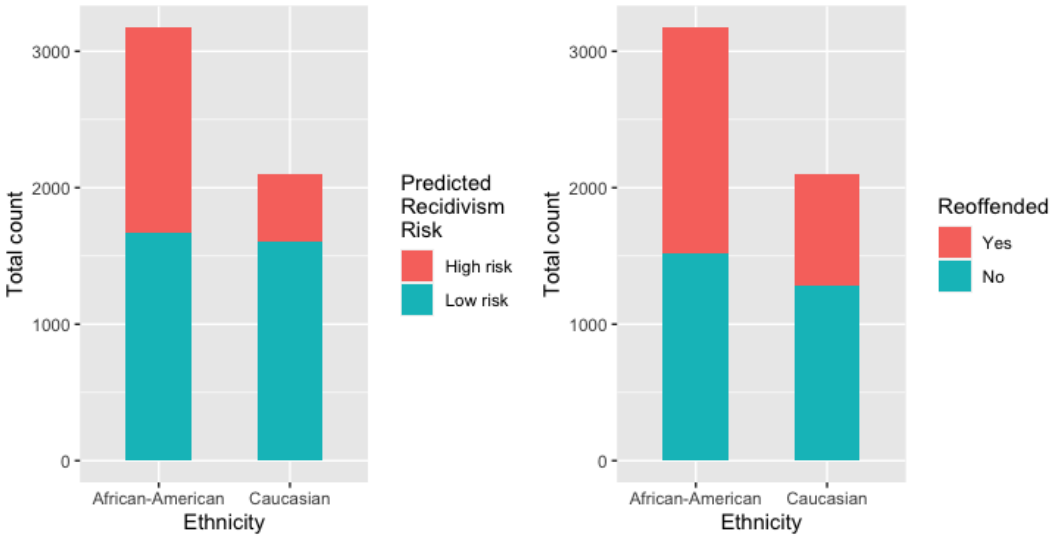
In this chapter, I will show how the previously developed fairness criteria (*causal relevance fairness, matched equalized odds, matched predictive parity*) can be applied to the COMPAS algorithm. For causal relevance fairness, this means testing the hypothesis that the influence of the variable *ethnicity* on the COMPAS recidivism risk predictions exceeds its relevance for those predictions. For matched equalized odds and matched predictive parity it means creating a data set matched on, among other things, the variable representing recidivism, to check whether in this matched data set both groups have equal error rates and predictive values.

The rest of the chapter is structured as follows. In Section 5.2, I will introduce the data set this case study is based on and conduct a brief exploratory analysis of the data. In Section 5.3, I will present the analysis of the COMPAS algorithm with regard to causal relevance fairness. In Section 5.4, I will present the analysis for the two criteria matched equalized odds and matched predictive parity. Section 5.5 discusses the results of both analyses.

5.2 An exploratory analysis of the COMPAS data

The analysis is based on the data set made available by ProPublica[1]. To create the data set, ProPublica merged COMPAS scores obtained from the *Broward County Sheriff's Office* in Florida with public criminal records from the *Broward County Clerk's Office* website. The resulting data set contains 7214 entries, each representing one defendant. The features of interest to our analysis are *age*, the *charge degree* (which takes values "M" for *misdemeanor*, or "F" for *felony*), *ethnicity*, the *number of prior convictions*, the *risk score* assigned by

[1] `https://github.com/propublica/compas-analysis` (Accessed: 4 October 2021).



(A) Predicted recidivism risk (B) Actual recidivism

FIGURE 5.1: Proportion of outcomes by ethnicity

the COMPAS algorithm, and whether defendants *actually recidivated* within two years after their trial.

In line with ProPublica's analysis, I removed a number of rows from the data set. First, those for which the date of the COMPAS evaluation was more than 30 days after the arrest, as this could indicate that the recorded COMPAS score is not for the recorded crime the defendant was arrested for. Second, those for which there was no COMPAS risk score. Third, those for which the charge was an ordinary traffic offense. Since in this analysis the focus is supposed to be on the difference between defendants identified as ethnically Black ("African American" in the data set) and those identified as White ("Caucasian" in the data set), I removed all those data points where the defendants' ethnicity was neither recorded as "Caucasian" nor "African American". This results in a data set containing 5278 entries.

These 5278 defendants can be divided into 3175 African American and 2103 Caucasian defendants, 4247 male, and 1031 female defendants. Of the defendants, 2002 received a high risk score (above five on the ten-point scale); 2647 did in fact recidivate. Like ProPublica, we will interpret a risk score of above five as the prediction that a defendant will recidivate within two years.

Of the African American defendants, 1506 were categorized as high risk (i.e. above 5) by the COMPAS algorithm, and 1661 did actually recidivate. Of the Caucasian defendants, 496 were categorized as high risk, and 822 did

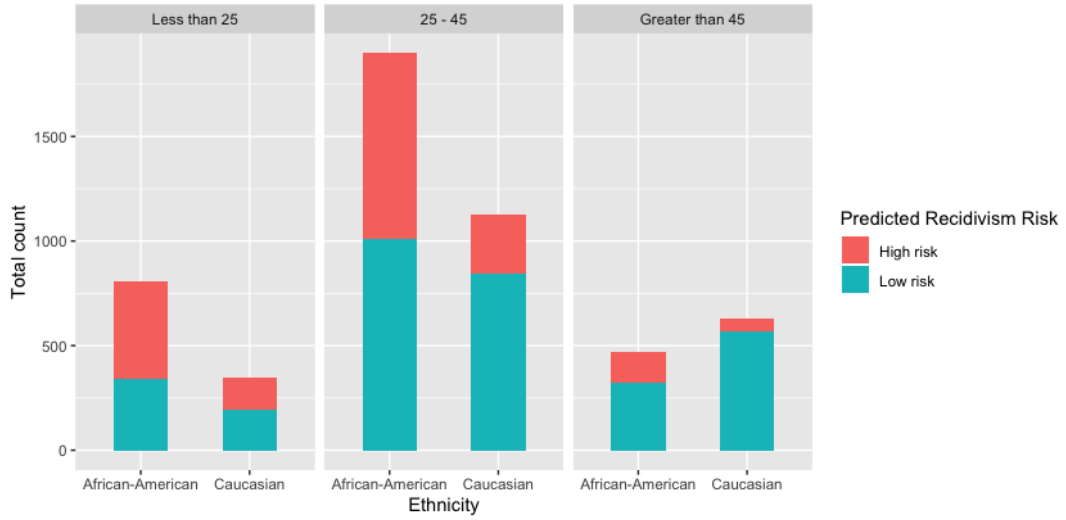


FIGURE 5.2: Proportion of predicted recidivism by ethnicity and age group

recidivate.

If we look at how risk scores are distributed among different age categories, we notice that for those below 25, more than half of the defendants received a high risk score by the COMPAS algorithm, while for those between 25 and 45 it was still a significant portion — about one third. For those above 45 it is only a relatively small fraction. This distribution is similar for actual recidivism.

If we look at risk scores and actual recidivism by ethnicity while controlling for age, we find that young and middle-aged African Americans are the group that had the greatest proportion of high risk predictions (see Figure 5.2). This is also the case for actual recidivism (see Figure 5.3). But, comparing the two box plots, it is striking that especially in the middle-aged group (25-45 years) the proportion of Caucasians that received a high risk score is significantly lower than the proportion that actually recidivated, while for African Americans the proportions of predicted high risk of recidivism and actual recidivism are more closely aligned.

To get a more precise sense of the disparities between African American and Caucasian defendants in terms of risk scores and actual recidivism, we can perform two t-tests in order to assess whether the seeming disparities are due to chance or whether we can assume that there are underlying systematic differences that drive the observed results. The t-test for a high risk evaluation yields a difference in proportion of 0.24. The 95% confidence interval for the

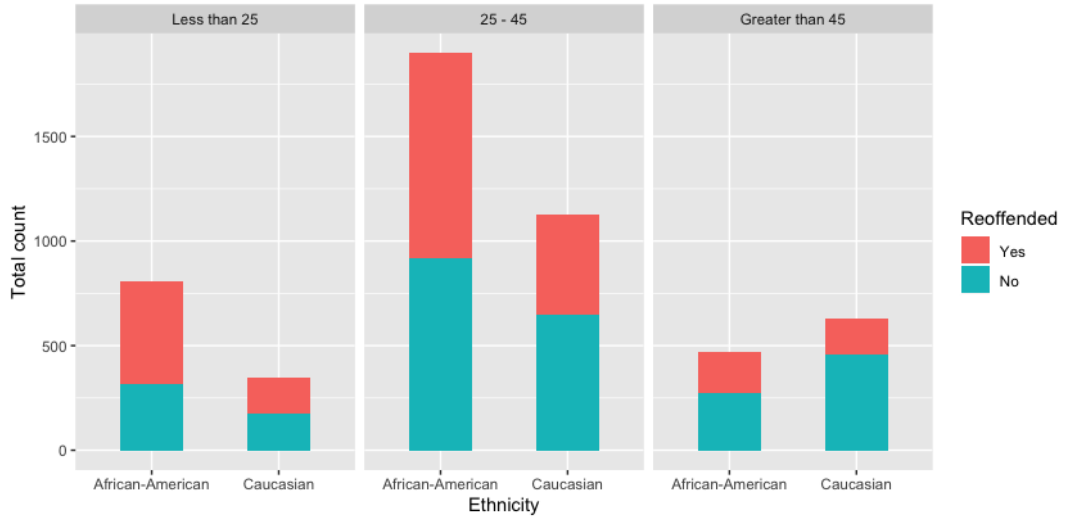


FIGURE 5.3: Proportion of actual recidivism by ethnicity and age group

difference in proportion of high risk evaluations ranges from 0.21 to 0.26. On the other hand, the t-test for actual recidivism yields a difference in proportion of 0.14 (with a 95% confidence interval from 0.12 to 0.17), which is significantly lower than the difference in proportion of a high risk evaluation. In words, this means that the proportion of defendants that are categorized as high risk is 24% higher in the group of African Americans than it is in the group of Caucasian defendants. However, the proportion of African Americans in our sample that actually recidivates is only 14% higher than the proportion of Caucasians.

Note that the above differences result from comparisons in which no other variables were controlled for. Running a multivariate logistic regression for both, high risk evaluation and actual recidivism, while controlling for age, sex, number of priors, and charge degree, yields that for an average African American defendant, the probability of a high risk evaluation is 11% higher than for a Caucasian defendant, but the probability of actual recidivism is only about 1% higher.

5.3 Causal relevance fairness

Let us begin with an examination of whether the COMPAS recidivism risk model is biased using the *causal relevance fairness* criterion developed in Chapter 3.

Recall that causal relevance fairness formalizes the idea that a prediction is fair (relative to a specific protected characteristic, in this case *ethnicity*), only if the influence of the protected characteristic on the prediction does not exceed its relevance. Expressed mathematically, this means we need to consider two causal quantities: the *influence* of the protected characteristic on the prediction, which we define as the causal effect of the protected characteristic on the prediction, and the relevance of the protected characteristic for the prediction, which we define as the causal effect of the protected characteristic on the target variable we aim to predict. In the COMPAS case, we thus need to estimate the causal effect of ethnicity on recidivism risk predictions by COMPAS (the *influence*), and the causal effect of ethnicity on whether defendants actually recidivate (the *relevance*).

Note that, for practical reasons, we here consider a slight generalization of *causal relevance fairness*. While initially, the criterion was defined on the basis of *token-causal effects* — that is, on the basis of whether a specific individual would have received the same risk prediction had they been of a different ethnicity — we here consider whether on the aggregate level those causal effects can be detected. This is because it is easier to estimate aggregate-level causal effects with relative reliability than token-level causal effects. We can take aggregate-level causal effects as indicative of the presence of token-level causal effects.

5.3.1 The matching method

To estimate the two causal quantities, the method of matching is used.[2] We could have chosen to use other causal inference methods because the definition of causal relevance fairness (as opposed to matched equalized odds and matched predictive parity) does not prescribe the use of one particular method. Matching, however, will in this case prove to be a convenient way to estimate the relevant causal quantities.

Recall that matching is a causal inference method that is inspired by the idea of randomization in experimental studies. The central point is to create a synthetic control group that is as similar as possible to the treatment group with regard to some relevant variables (the *covariates*). By ensuring similar distributions over the covariates we can be confident that any observed difference

[2]For a more detailed explanation of the theory underlying the matching method, see Chapter 4.

in outcomes is only due to the causal ("treatment") variable. As matching is usually applied within the *potential outcomes framework* for causal inference in statistics, we will adopt this framework here as well (Rubin, 2005).

In a nutshell, the idea of the potential outcomes framework is the following. Assume we are interested in the effect of treatment T on outcome O, where **X** are the covariates. Moreover, assume our sample consists of n individuals. For every individual i, we can define two quantities: the potential outcome under treatment, expressed as $O_i(T_i = 1)$, and the potential outcome under control, expressed as $O_i(T_i = 0)$. As an individual can only ever be in either the treatment or the control group, we can, of course, only measure either of the two quantities. The other is a counterfactual quantity. The causal effect of T on O (for individual i) can then be expressed as the conditional expectation[3] of the difference between the two quantities, i.e. $E(O_i(T_i = 1) \mid \mathbf{X}) - E(O_i(T_i = 0) \mid \mathbf{X})$ for $i \in \{1, ..., n\}$ (Stuart, 2010, p. 3). To obtain the aggregate-level causal effect, in addition, we need to average over all the individual causal effects.

In the present case, this means we take the group of African Americans and construct a synthetic control group of Caucasian defendants that match the African Americans on a number of relevant covariates, which we will determine in a moment. The frequency distribution over the covariates should be as similar as possible for African American and Caucasian defendants. We then determine (1) the causal effect of ethnicity on COMPAS recidivism predictions, and (2) the causal effect of ethnicity on actual recidivism.

To formalize this, let A denote ethnicity (where $A_i = 1$ is to be interpreted as individual i being identified as African American, and $A_i = 0$ as i being identified as Caucasian, respectively), Y whether a defendant actually reoffends, $\hat{Y}$ whether a defendant is predicted to have a high risk of recidivism, and **X** the set of covariates. We can then define the (average) *influence of ethnicity on the recidivism risk prediction* as follows:

$$\alpha = \frac{\sum_{i \in \{1,...,n\}} (E(\hat{Y}_i(A_i = 1) \mid \mathbf{X}) - E(\hat{Y}_i(A_i = 0) \mid \mathbf{X}))}{n} \tag{5.1}$$

[3]The conditional expectation can be understood as the average value a random variable would take on in the long run. It is defined as

$$E(X \mid Y) = \sum_x xP(X = x \mid Y)$$

.

Analogously, we define the (average) *relevance of ethnicity for recidivism risk predictions* as follows:

$$\beta = \frac{\sum_{i \in \{1,\ldots,n\}} (E(Y_i(A_i = 1) \mid \mathbf{X}) - E(Y_i(A_i = 0) \mid \mathbf{X}))}{n} \tag{5.2}$$

The hypothesis that the COMPAS algorithm does not satisfy causal relevance fairness can hence be expressed as the former being greater than the latter:

$$\alpha > \beta \tag{5.3}$$

5.3.2 Modelling assumptions

In order to determine the covariates, we need to assume a qualitative causal model of the variables in the dataset. The focus here is the potential causal link from *ethnicity* to the predicted recidivism risk $\hat{Y}$ as well as to whether someone actually reoffends (Y). We assume there could potentially be a latent common cause of *ethnicity*, *charge degree,* and *number of prior convictions*. This latent variable could be something like the family an individual is born into, which might determine or at least partially influence genetic factors, socioeconomic level, exposure to violence, and so on. Further, we will draw links from *age* and *sex* to *charge degree* and *number of prior convictions*. This seems reasonable, as there is evidence that both sex (see, e.g., Mawby, 1980) and age (see, e.g., Ulmer and Steffensmeier, 2014) have an effect on criminal behavior. No link is drawn from *ethnicity* to either *number of prior convictions* nor *charge degree*, as studies indicate that if one controls for the right background variables, the correlation between ethnicity and crime rates vanishes (see, e.g., Ulmer et al., 2012).

Based on the COMPAS handbook[4], we know that sex, age, and criminal behavior are taken into account by the COMPAS algorithm, and hence we can assume they potentially have a causal effect on the risk prediction. While ethnicity is not explicitly recorded in the data on which the prediction is based, there may be proxies for ethnicity in the data. This makes it possible that ethnicity causally influences the risk prediction without being explicitly

[4] https://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf (Accessed: 5 April 2022).

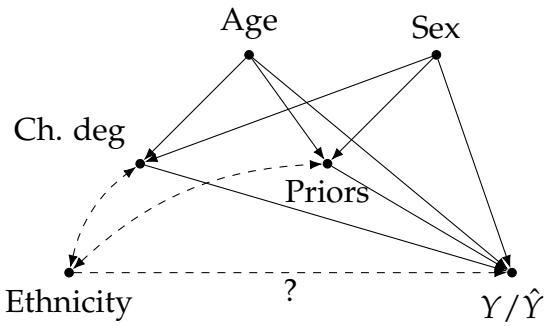


FIGURE 5.4: Causal graph

recorded in the dataset. Figure 5.4 graphically summarizes our causal judgments.

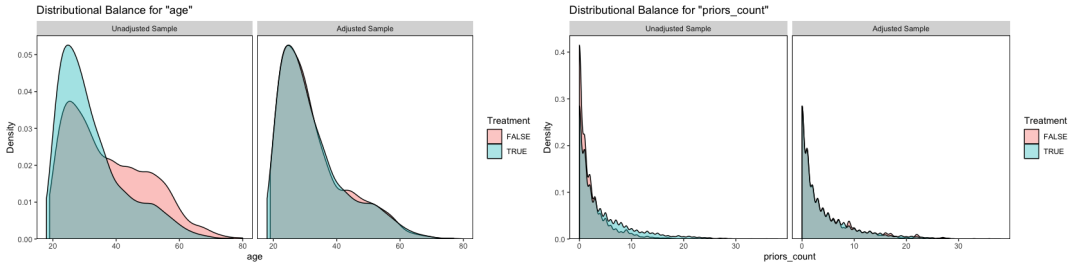
Further potential confounding factors for the risk prediction could include socioeconomic status and previous exposure to crime, which are not included in the dataset. However, assuming that the number of priors and charge degree are sufficiently correlated with these potential confounders and that they can hence be considered proxies, justifies the assumption that we have sufficiently precise observations of all the factors that influence the risk prediction. This assumption, as explained in Section 4.3, is called *ignorability* and is a precondition for reliably inferring causal effect magnitudes via matching.

Given the causal graph above, it seems plausible to include the four variables *age*, *sex*, *number of prior convictions*, and *charge degree* in the set of covariates in order to create a matched control group. Since our set of covariates is not highly dimensional, we can use a distance-based matching method[5]. Specifically, we will use a method based on the *Mahalanobis distance* measure, which is a common choice in comparable studies.

We defined being identified as African American as the treatment and being identified as Caucasian as the control group property. Since the number of observations in the treatment group is significantly greater than the number of observations in the control group, we will use a matching procedure with replacement. This means some of the data points describing Caucasian defendants will be used more than once in estimating the causal effect.

Two matching methods will be tried out: 1:1 and 2:1 nearest-neighbors matching. This means, for each African American defendant we will add the single

[5]In highly dimensional space, distances converge and hence become meaningless. This is known as the *curse of dimensionality* (see, e.g., Beyer et al., 1999).



(A) Age (B) Count of prior convictions

FIGURE 5.5: Balance of covariates *Age* and *Count of prior convictions* after 1:1 matching

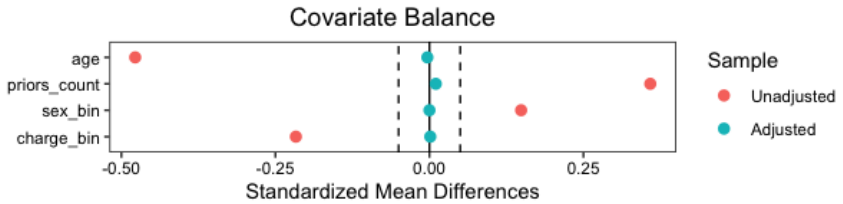


FIGURE 5.6: Summary covariate balance after 1:1 matching

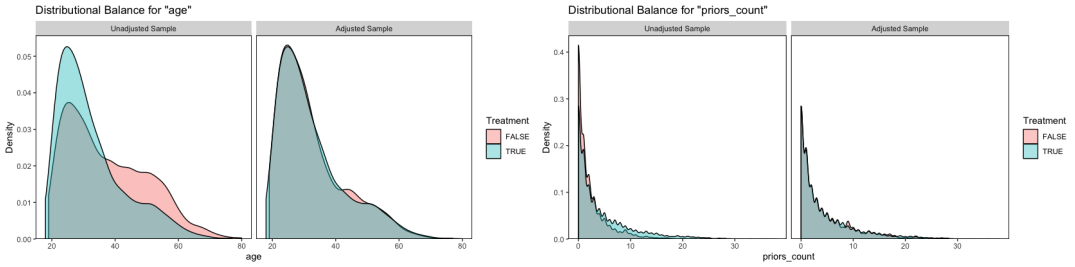
most similar Caucasian defendant to the dataset (1:1). Then we will compare this to matching each African American defendant to the average of the two most similar Caucasian defendants (2:1). Depending on which of the two methods yields the better balance between samples, we will make our choice of matching method for estimating the treatment effect. Let us first analyze the adjusted sample using 1:1 matching.

For *sex* and *charge degree*, the plot indicates that we achieved a (close to) perfect balance, and also for *age* (see Figure 5.5a) and *number of prior convictions* (see Figure 5.5a) it looks like the balance the 1:1 matching achieves is sufficient. If we look at the standardized difference in means[6] for the adjusted sample, we see that with *sex* we indeed have a perfect match, with *charge degree* we have a negligible difference of 0.0006, for *age* 0.0033, and for *priors count* 0.0104. This is by all standards a very close match. The numbers are summarized in Figure 5.6.

Next, we will perform 2:1 matching and see whether we can improve the balance achieved by 1:1 matching. Again, the diagrams indicate that our adjusted sample matches the treatment group relatively closely.

More precisely, there is no difference in means for *sex*, and a negligible one for

[6]That is, the mean difference expressed in units of standard deviation. Standardizing the mean difference helps make differences better interpretable across variables. This is standard practice in matching (Rubin, 2001).



(A) Age (B) Count of prior convictions

FIGURE 5.7: Balance of covariates *Age* and *Count of prior convictions* after 2:1 matching

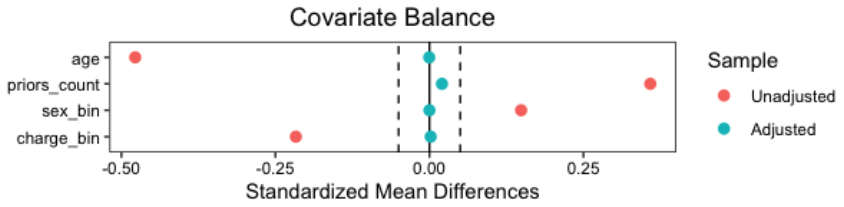


FIGURE 5.8: Summary covariate balance after 2:1 matching

charge degree (0.0009), the standardized difference in means for *age* has gone down by a very small degree (to 0.0003) but the difference in means for the *number of prior convictions* has doubled to 0.0203 (see Figure 5.7 and Figure 5.8). Since a good balance on the number of prior counts seems desirable, we are consequently going to use 1:1 matching for our estimation of the causal effects.

Before we begin the estimation of the causal effect, we have to address the question of whether we should estimate the average treatment effect (ATE) or the average treatment effect for the treated (ATT). The latter would be a good choice in those cases in which we are more interested in the treated population. An example of this sort is smoking: what a study on the effects of smoking intends to assess is what would have been the case if the smoker had not smoked. It is less interesting to ask what would have happened if a given non-smoker would have smoked. In our case, however, we are interested in both counterfactuals. How would the risk evaluation have differed if a given person would have been of Caucasian rather than African American ethnicity? And, equally relevant, how would the risk evaluation have differed if a given person would have been of African American rather than Caucasian ethnicity? Estimating the ATE can be problematic if there is no sufficient overlap between the treatment and the control group. This, however, is not the case in our sample, as the foregoing analysis has shown.

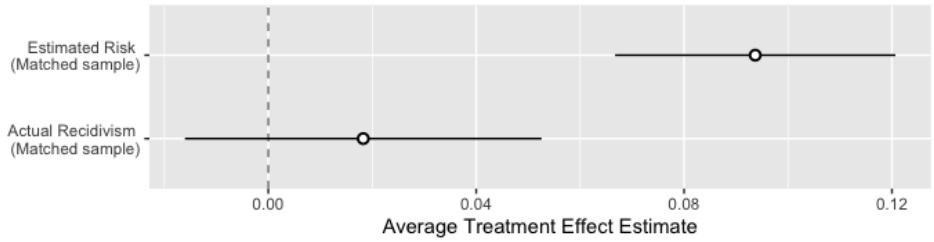


FIGURE 5.9: Comparison of causal estimates. The circles represent the estimated ATE, the lines represent the corresponding 95% confidence intervals.

5.3.3 Results

The estimated difference in proportion for a high risk prediction is 0.094. In other words, being African American makes it almost 10% more likely to be predicted to be at high risk of recidivism by the COMPAS algorithm as compared to being Caucasian. The standard error of our estimate is 0.014. The p-value is well below the 0.01 level of statistical significance. Given that our ignorability assumption holds, we can conclude that the ethnicity of a defendant does indeed have a significant effect on the COMPAS recidivism risk prediction. These results are relatively robust with regard to the influence of unobserved confounders.[7]

To contrast this with the results for *actual recidivism*, we find that the estimated difference in proportion from the adjusted sample is only 0.020. The standard error of this estimate is 0.015, and the p-value is 0.19. This means the result is not significant at any conventional level of statistical significance. We have to conclude that our analysis does not establish a causal link between ethnicity and actual recidivism. In other words, being African American (rather than Caucasian) does not make one more likely to re-offend. This is in line with scientific evidence on racial disparities in crime behavior (see, e.g., Ulmer et al., 2012).

The investigation confirms the hypothesis that $\alpha > \beta$, that is, the influence of

[7] To check how robust our results are with regards to unobserved confounders, we can conduct a sensitivity analysis using Rosenbaum bounds (Rosenbaum, 2002). The upper bound of the p-value remains below the 0.01 level of significance up to a gamma value of 1.7 — that is, we would only change our conclusion if there were an unobserved characteristic that is associated with high risk scores and that is 1.7 times more common among African Americans rather than Caucasian defendants. While this shows that our causal estimate is somewhat sensitive to the presence of unobserved confounders, the result that conclusions are only valid up to such a confounding level is not uncommon in the social sciences (see, e.g., Becker and Caliendo, 2007, p. 80).

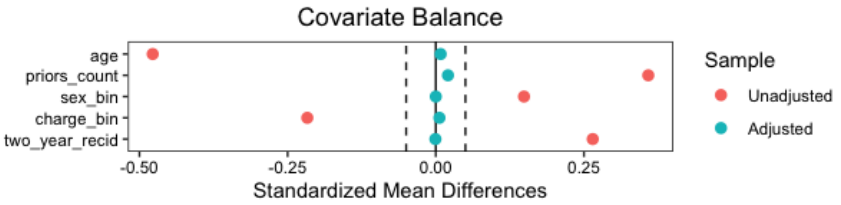


FIGURE 5.10: Summary covariate balance after 1:1 matching

ethnicity on the recidivism risk prediction exceeds its relevance. The COMPAS recidivism risk predictions hence do not satisfy causal relevance fairness, provided our modeling assumptions hold.

5.4 Matched equalized odds and matched predictive parity

Next, we shall examine the COMPAS recidivism risk model according to the two fairness criteria developed in Chapter 4. Recall that in order to evaluate those criteria, we need to consider a matched dataset where the target variable is among the covariates. *Matched equalized odds* is defined as equal false positive and false negative error rates on this dataset, whereas *matched predictive parity* is defined as equal positive and negative predictive values on this dataset. If the two criteria are satisfied, we can interpret this as showing that ethnicity does not have an effect on COMPAS' error rates and predictive values, respectively.

5.4.1 The matched dataset

Other than in the previous section, we here need to include information on whether the defendant *actually recidivated* in the set of covariates on which we match the data points. As before, we will compare 1:1 and 2:1 nearest-neighbors matching. For both, Mahalanobis distance will be used to determine closeness, as well as matching with replacement.

Again, 1:1 matching provides a slightly more closely matched dataset. The covariate balance is represented in 5.10. Note, in particular, the almost exact match on the target variable (whether a defendant recidivated within two years). The standardized mean difference is 0.0002. This is important because it means that the recidivism base rates for both, African Americans as well as Caucasians, are only negligibly different. This, as shown in Section 4.4,

FIGURE 5.11: Comparative false negative and false positive rates, and positive and negative predictive values for African American and Caucasian defendants as well as overall values (in the matched dataset).

resolves the incompatibility between equalized odds and predictive parity. If COMPAS were perfectly fair, we could expect its error rates and predictive values on this dataset to be equal for both groups.

The modelling assumptions for this analysis are, as in the previous section, that the causal relationships of the relevant variables are as depicted in Figure 5.4, and that consequently the set of covariates *age*, *number of prior convictions*, *sex*, *charge degree*, and *whether they recidivated* suffices to satisfy the ignorability assumption.

5.4.2 Results

We can now use the matched dataset to check whether the COMPAS risk predictions satisfy matched equalized odds and matched predictive parity.

First, we find that matched equalized odds is not satisfied. In the matched dataset, the false negative error rate is still significantly lower for African American defendants than for Caucasian defendants (40% vs 53%), while the reverse is true for false positive rates (28% vs 18%). This means that there are discrepancies in error rates that cannot be explained away by differences in base recidivism rates (which are overall higher for the African American defendants in the dataset) and other variables. Recall that in the initial dataset, the discrepancy in false negative rates was 20%, whereas in the matched dataset it is only 13%. Hence, a discrepancy of only 7% can be explained away. We see a similar pattern for false positive rates. Here,

the discrepancy in the initial dataset was 23%, which shrunk to 10% in the matched dataset. Hence, 13% of the discrepancy can be explained away, but the remaining discrepancy of 10% must be assumed to be due to the difference in ethnicity.

Secondly, we find that matched predictive parity is not satisfied either. It has to be acknowledged, though, that the discrepancies in predictive values are much less striking than the discrepancies in error rates. The positive predictive value for African American defendants is 62%, while it is 67% for Caucasian defendants. Hence, a discrepancy of 5% can be attributed to racial bias. The same is true for a 3% discrepancy in negative predictive value (the negative predictive value for African American defendants is 70% and 67% for Caucasian defendants).

5.5 Discussion

The above analyses confirm the hypothesis I initially set out to investigate, namely the claim that the COMPAS algorithm is racially biased. It shows that higher average risk predictions for African American defendants cannot be explained away as mere correlations that come about through differences in other factors (such as for instance, different base recidivism rates, different age distributions, or different average socioeconomic levels). Nor can different error rates or predictive values be explained away by those factors. Being African American makes a systematic difference to COMPAS' risk predictions and to how accurate they are. The evidence, however, does not lend credence to the claim that being African American makes a difference to whether someone actually ends up reoffending — which could potentially have justified these disparities to some degree.

Our investigation hence supports ProPublica's hypothesis that the COMPAS recidivism risk predictions are not fair, yet it does so from a different perspective than ProPublica's own analysis. Applying *causal relevance fairness*, *matched equalized odds*, and *matched predictive parity* as criteria of fairness gives us, at least in principle, a more robust assessment of racial bias than merely comparing different error rates between ethnic groups. The reason for this is that we aim at identifying the cause of the disparity in the outcome, rather than just observing a correlation between ethnicity and higher or lower error rates. Comparing only observed error rates (as done by ProPublica) might

lead to skewed results if there are unobserved factors present that are correlated with ethnicity and which incorrectly drive the predictions in one direction.

Yet, this robustness comes at a price. The assumptions we have to make in order to conduct the analysis of whether an algorithm satisfies these causal criteria of fairness are stronger than the assumptions ProPublica needed to make to check for equal error rates. This is particularly problematic when the full dataset on which predictions were based is not available because then there is no guarantee that the crucial *ignorability* assumption holds. A potential confounder we did not control for is socioeconomic status. Our analysis implicitly relied on the assumption that *charge degree* and *number of prior crimes* are sufficiently correlated with socioeconomic status, such that matching on these variables also yields a balanced distribution of socioeconomic status. Hence the validity of our fairness evaluation depends on the plausibility of this assumption.

5.6 Conclusion

In this chapter, I have attempted to assess the hypothesis that COMPAS — a computational tool used in some US courts for assessing defendants' risk of recidivism — is racially biased against African Americans. In order to do so, I applied three previously developed criteria of fairness. As it turned out, the analysis confirmed the hypothesis: based on the available data, it seemed that none of the three fairness criteria were satisfied by the COMPAS algorithm. I ended with a discussion of the assumptions we had to make to arrive at our conclusions.

Conclusion

The point of departure for this thesis was that in the discussion around algorithmic fairness there is a gap to be filled between philosophy and computer science: philosophers, I argued, ought to engage in more depth with proposed technical solutions to the problem(s) of algorithmic fairness, while computer scientists ought to engage in more depth with the normative theory that motivates them.

How, then, did this thesis contribute to filling this gap? I began by introducing a conceptual distinction between the ethics of algorithmic predictions — *predictive algorithmic fairness* — and the ethics of decisions based on algorithmic predictions — *allocative algorithmic fairness* (Chapter 1). This served as a precondition for an orderly discussion of algorithmic fairness, as it allowed us to put aside questions of distributive justice, which belong to the normative realm of decision-making. Instead, we were able to focus on the normative realm of predictions, which is to ensure the absence of discriminatory bias.

Zeroing in on predictive algorithmic fairness, I then started to investigate the role and use of causal reasoning in algorithmic fairness (Chapter 2). The starting point was a popular causal criterion of predictive algorithmic fairness called counterfactual fairness. I showed that under certain conditions, counterfactual fairness and two other popular and intuitively appealing criteria of fairness, namely equalized odds and predictive parity, are pairwise incompatible. The upshot of this investigation was the following. If we deem the relevant type of situation conceivable, insist that there always has to be a fair way of making predictions, and are unwilling to simultaneously give up equalized odds and predictive parity, then we have to weaken, replace, or fully give up counterfactual fairness. This follows from the logical relations between the three different mathematical fairness criteria alone.

This motivated the investigation into the normative roots of the logical impasse in the next chapter (Chapter 3). I argued that the central wrong involved in biased algorithmic predictions is a failure to treat a person as an individual. If understood as a constraint aimed at prohibiting this type of wrong, counterfactual fairness is too strong a criterion to ensure fair predictions. I consequently developed a weakened version of counterfactual fairness — *causal relevance fairness* -, which is more closely in line with moral theory and which avoids the logical impasse. Causal relevance fairness is firmly grounded in moral theory but is defined as a formal criterion that can be applied in a straightforward and rigorous manner to evaluate the fairness of a given predictive algorithm.

Next, I turned to another logical incompatibility that had beset the discussion around algorithmic fairness, namely the incompatibility between equalized odds and predictive parity (Chapter 4). Examined through a causal lens, it seemed that the incompatibility is due to an overly strong formalization of two intuitively appealing aspects of fairness. As in the previous chapter, I showed how the criteria can be modified in a way that retains their intuitive appeal but resolves the incompatibility. As it turned out, thus modified the criteria did not only become compatible but equivalent: whenever one is satisfied, the other is as well.

The last chapter (Chapter 5) aimed to demonstrate the practical usefulness of the foregoing philosophical investigations. In it, I examined data on the infamous COMPAS tool, a tool that predicts how likely it is that a defendant will recidivate. I used the previously developed criteria of predictive algorithmic fairness to evaluate the fairness of the tool. The study confirmed the hypothesis that COMPAS is unfairly biased against African American defendants.

The upshots of this thesis will hopefully be of interest to philosophers and computer scientists alike. From a philosophical point of view, this thesis aimed to raise and clarify a number of conceptual, logical, and normative issues in the discussion of algorithmic fairness. From a technical point of view, it aimed to show how moral theory can be translated into applicable mathematical fairness constraints for machine learning algorithms.

If this thesis was successful in its aim, it will have shrunk the gap between philosophy and computer science. Yet, much is left to be done. To conclude this thesis, I will provide a brief overview of three possible avenues for future work.

First, an examination of the question of whether there is a tension between rationality constraints and the three fairness criteria developed in this work. For many existing fairness criteria, it was shown that (at least under non-ideal circumstances) enforcing the criteria engenders a decrease in prediction accuracy (see, e.g., Menon & Williamson, 2018; Chen et al., 2018). Ideally, a predictive fairness criterion should be such that being fair is compatible with being maximally accurate. Future work could address whether this is the case for causal relevance fairness, matched equalized odds, and matched predictive parity. If not, it would be valuable to rigorously characterize the trade-off.

Secondly, an examination of the potential risks of fairness gerrymandering. A common worry regarding fairness criteria is that it might be possible for the creators of unfair prediction algorithms to (superficially) manipulate the algorithm or their data in a way such that the algorithm remains unfair, but satisfies the fairness criterion in question. Ideally, a fairness criterion should make this type of gerrymandering as difficult as possible. Future work could engage with the question of how gerrymandering could affect causal relevance fairness, matched equalized odds, and matched predictive parity.

Lastly, from a technical side, it would be interesting to examine how popular machine learning algorithms have to be constrained to be able to minimize their cost function while guaranteeing that the fairness criteria causal relevance fairness, matched equalized odds, and matched predictive parity will be satisfied.

Appendix A

A.1 Decision tree for Scenarios 1 and 2

To better understand the difference between Scenarios 1 and 2, the two scenarios can be represented in the form of a decision tree. In both, Figure A.1 and Figure A.2, we can see that there are, all in all, 200 adequately qualified applicants for university degrees (i.e. $Y = 1$). As can be seen in the confusion matrix, all unqualified applicants are correctly predicted to be unqualified, hence the false positive rate is 0 (for men and women), and so we can leave them aside in the analysis. This will facilitate the presentation.

In both scenarios, there are 100 adequately qualified female and 100 adequately qualified male applicants. In Scenario 1 (Figure A.1), however, all of them apply for a business degree. Of the 100 adequately qualified female applicants, 28 are incorrectly predicted to be unqualified. 72 are correctly predicted to be adequately qualified for the degree. Among the male applicants, on the other hand, only 20 are incorrectly predicted to be unqualified, and 80 are predicted to be adequately qualified. This amounts to a false negative rate of $\frac{28}{100} = 0.28$ for women, and $\frac{20}{100} = 0.20$ for men. Here, the department's predictive model seems to have some sort of gender bias against women.

If, in comparison, we look at Scenario 2 (Figure A.2), things look very different. Here, of the 100 adequately qualified female applicants, 90 apply for a degree in physics. Of the male applicants, only 50 apply for a degree in physics. Since, by assumption, the false negative rate of the physics department is 0.3 (for both men and women), 27 women and 15 men are incorrectly predicted to be unqualified. At the same time, only 10 women but 50 men apply to the business department, which has a lower false negative rate of only 0.1, which results in 1 women and 5 men being incorrectly predicted to be unqualified. This, as above, amounts to a false negative rate of $\frac{27+1}{100} = 0.28$ for women, and $\frac{15+5}{100} = 0.20$ for men. Only in this case, we know that the

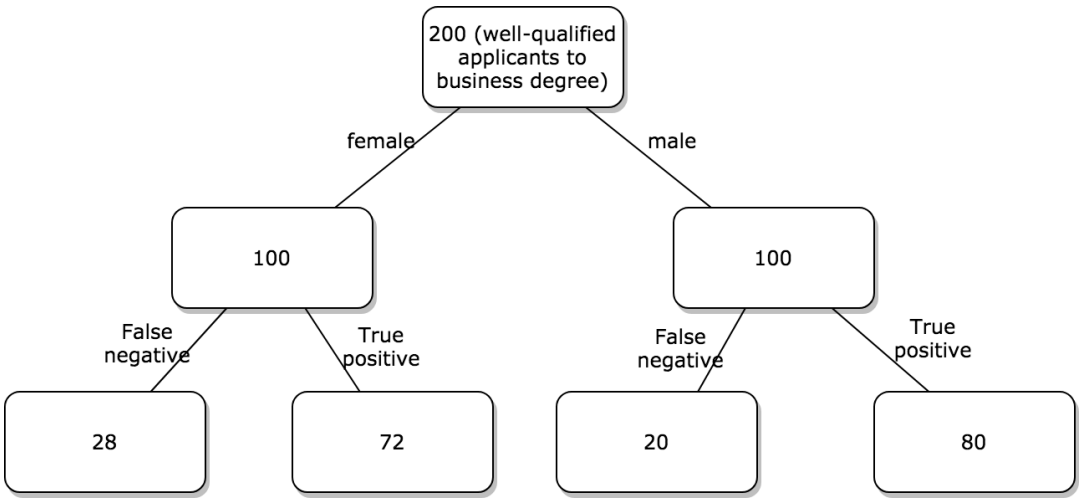


FIGURE A.1: Decision tree for Scenario 1.

different departments' false negative rates are not different for men and for women, and the difference in overall false negative rates between men and women is due to the fact that more women choose to study physics (which has a higher false negative rate) than men. It does not seem that the predictive model is gender biased.

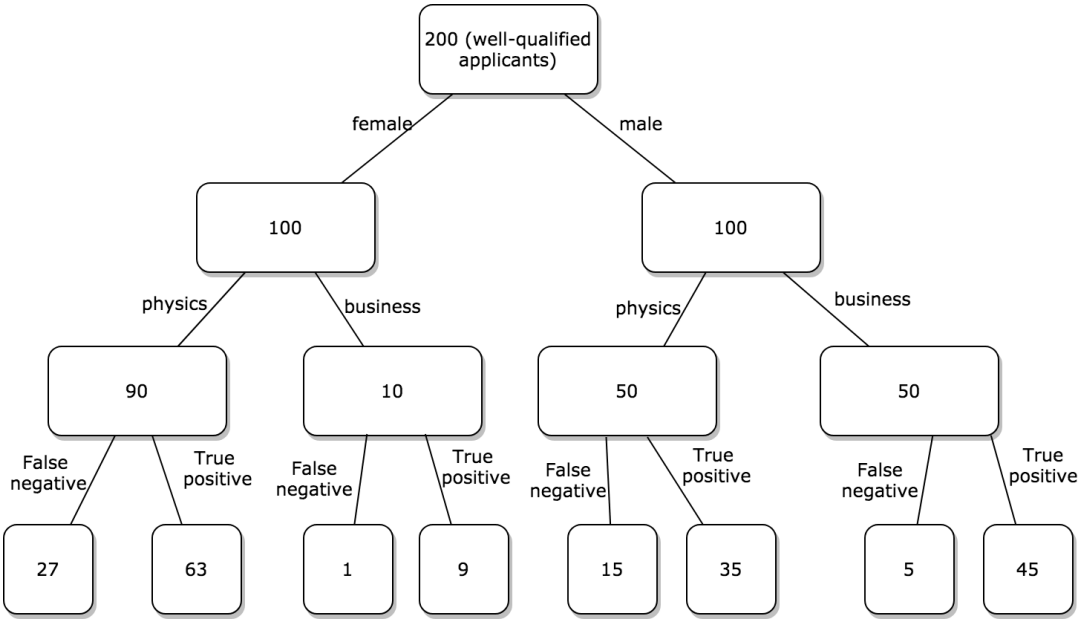


FIGURE A.2: Decision tree for scenario 2.

Appendix B

B.1 Calculation of PPV for men and women

Let the variable T denote the test result of the device (1 indicates a positive and 0 a negative test result), D whether a person has the disease (value 1 indicates the presence of the disease and 0 the absence), and A the gender of the person (male or female, indicated by 1 and 0). To calculate the positive predictive values for men and women, we assume that the device has the same true and false positive rates (0.95 and 0.05, respectively) for men and for women, that is:

$$\begin{aligned} TPR &= P(T = 1 \mid D = 1) \\ &= P(T = 1 \mid D = 1, A = 1) \\ &= P(T = 1 \mid D = 1, A = 0) \\ &= 0.95 \end{aligned} \tag{B.1}$$

$$\begin{aligned} FPR &= P(T = 1 \mid D = 0) \\ &= P(T = 1 \mid D = 0, A = 1) \\ &= P(T = 1 \mid D = 0, A = 0) \\ &= 0.05 \end{aligned} \tag{B.2}$$

We also know that the disease is more common in men than in women:

$$P(D = 1 \mid A = 1) = 0.1 \tag{B.3}$$

$$P(D = 1 \mid A = 0) = 0.01 \tag{B.4}$$

To increase the readability of equations, we will use the following abbreviations:

$$p := P(T = 1 \mid D = 1, A = 1)$$

$$q := P(T = 1 \mid D = 1, A = 0)$$

$$r := P(D = 1 \mid A = 1)$$

$$s := P(D = 1 \mid A = 0)$$

$$t := P(T = 1 \mid D = 0, A = 1)$$

$$u := P(T = 1 \mid D = 0, A = 0)$$

$$v := P(D = 0 \mid A = 1)$$

$$w := P(D = 0 \mid A = 0)$$

Now we can use Bayes theorem to calculate the positive predictive value for men and women.

$$\begin{aligned} PPV_{male} &= \frac{p * r}{p * r + t * v} \\ &= \frac{0.95 * 0.1}{0.95 * 0.1 + 0.05 * 0.9} \approx 0.68 \end{aligned} \tag{B.5}$$

$$\begin{aligned} PPV_{female} &= \frac{q * s}{q * s + u * w} \\ &= \frac{0.95 * 0.01}{0.95 * 0.01 + 0.05 * 0.99} \approx 0.16 \end{aligned} \tag{B.6}$$

B.2 Proof of Theorem 3 and 4

To prove Theorem 3, suppose that the predictive models we consider satisfy equalized odds relative to a_1 and $a_2 \in D_A$, i.e. $P(\hat{y} \mid a_1, y) = P(\hat{y} \mid a_2, y)$ for all $\hat{y} \in D_{\hat{Y}}$ and $y \in D_Y$. We here ignore whether the data is matched or unmatched, as we are only interested in how the difference in positive and negative predictive value between the groups changes when the difference

in prevalence changes.

Without loss of generality, we assume that group a_1 has a higher prevalence than a_2. This allows us to define the difference in prevalence as

$$P(Y = 1 \mid a_1) - P(Y = 1 \mid a_2) = \Delta p \tag{B.7}$$

To make the calculations more readable, we will abbreviate expressions as follows:

$$u := P(\hat{Y} = 1 \mid Y = 1, A = a_1)$$

$$v := P(\hat{Y} = 1 \mid Y = 0, A = a_1)$$

$$w := P(Y = 1 \mid A = a_2)$$

By Bayes theorem, equalized odds, and the definition of prevalence, we can then express the positive predictive value of groups a_1 and a_2 as follows:

$$PPV_{a_1} = \frac{u(w + \Delta p)}{u(w + \Delta p) + v(1 - w - \Delta p)} \tag{B.8}$$

$$PPV_{a_2} = \frac{uw}{uw + v(1 - w)} \tag{B.9}$$

We can now define the difference between the two groups' positive predictive value as a function of their difference in prevalence:

$$\begin{aligned} \Delta PPV(\Delta p) &= PPV_{a_1} - PPV_{a_2} \\ &= \frac{u(w + \Delta p)}{u(w + \Delta p) + v(1 - w - \Delta p)} - \frac{uw}{uw + v(1 - w)} \end{aligned} \tag{B.10}$$

We want to show that the smaller the difference in prevalence Δp, the smaller the difference in positive predictive value, i.e. the smaller $\Delta PPV(\Delta p)$. To do this, we need to show that $\Delta PPV(\Delta p)$ is strictly increasing on the interval $[0, 1]$. To do this, we calculate the derivative of $\Delta PPV(\Delta p)$:

$$\begin{aligned}\Delta PPV'(\Delta p) = &\frac{u\,(\Delta p + w)\,(-u + v)}{\left(u\,(\Delta p + w) + v\,(-\Delta p - w + 1)\right)^2} \\ &+ \frac{u}{u\,(\Delta p + w) + v\,(-\Delta p - w + 1)}\end{aligned} \tag{B.11}$$

This derivative is always positive for $u, w, v, \Delta p \in (0, 1)$. $\Delta PPV(\Delta p)$ is hence strictly increasing on the interval $[0, 1]$.

Analogously, we can show that the difference in negative predictive value $\Delta NPV(\Delta p)$ strictly increases on the interval $[0, 1]$.

Theorem 4 can be proved in a similar fashion, for which reason the proof is omitted.

Bibliography

Abramov, I., Gordon, J., Feldman, O., & Chavarga, A. (2012). Sex and vision II: color appearance of monochromatic lights. *Biology of sex differences, 3(1)*, 1-15.

Alexander, L. (1992). What makes wrongful discrimination wrong? Biases, preferences, stereotypes, and proxies. *University of Pennsylvania Law Review, 141(1)*, 149-219.

Aliprantis, D. (2017). Human capital in the inner city. *Empirical Economics, 53(3)*, 1125-1169.

Allen, J. A. (2019). The color of algorithms: An analysis and proposed research agenda for deterring algorithmic redlining. *Fordham Urb. LJ, 46*, 219.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: there's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica.*

Anstey, K. J., Horswill, M. S., Wood, J. M., & Hatherly, C. (2012). The role of cognitive and visual abilities as predictors in the Multifactorial Model of Driving Safety. *Accident Analysis & Prevention, 45*, 766-774.

Apfelbaum, E. P., Pauker, K., Sommers, S. R., & Ambady, N. (2010). In blind pursuit of racial equality?. *Psychological science, 21(11)*, 1587-1592.

Arnett, J. J. (2002). Developmental sources of crash risk in young drivers. *Injury prevention, 8(suppl 2)*, ii17-ii23.

Avin, C., Shpitser, I., & Pearl, J. (2005). Identifiability of path-specific effects. In *Proceedings of the 19th international joint conference on Artificial intelligence (IJCAI'05)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 357–363.

Becker, S. O., & Caliendo, M. (2007). Sensitivity analysis for average treatment effects. *The Stata Journal, 7(1)*, 71-83.

Beigang, F. (2022). On the Advantages of Distinguishing Between Predictive and Allocative Fairness in Algorithmic Decision-Making. *Minds and Machines*, 1-28.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research, 50(1)*, 3-44.

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is "nearest neighbor" meaningful?. In *International conference on database theory*. Springer, Berlin, Heidelberg, 217-235

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science, 187(4175)*, 398-404.

Binns, R. (2020). On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 514-524.

Bonilla-Silva, E. (2006). Racism without racists: Color-blind racism and the persistence of racial inequality in the United States. Rowman & Littlefield Publishers.

Bradley, R. (2017). *Decision theory with a human face.* Cambridge University Press.

Carey, A. N., & Wu, X. (2022). The statistical fairness field guide: perspectives from social and formal sciences. *AI and Ethics*, 1-23.

Chen, I., Johansson, F. D., & Sontag, D. (2018). Why is my classifier discriminatory?. *Advances in neural information processing systems, 31*.

Chiappa, S. (2019). Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence 33(01)*, 7801-7808.

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

Classen, S., Wang, Y., Crizzle, A. M., Winter, S. M., & Lanford, D. N. (2013). Gender differences among older drivers in a comprehensive driving evaluation. *Accident Analysis & Prevention, 61*, 146-152.

Cleary, T. A. (1966). Test bias: Validity of the Scholastic Aptitude Test for Negro and white students in integrated colleges. *ETS Research Bulletin Series, 1966(2)*, i-23.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 797-806.

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023.

Darlington, R. B. (1971). Another look at "cultural fairness" 1. *Journal of Educational Measurement, 8(2)*, 71-82.

Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). *Mathematics for machine learning*. Cambridge University Press.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214-226.

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

Eva, B. (2022). Algorithmic fairness and base rate tracking. *Philosophy & Public Affairs, 50(2)*, 239-266.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 259-268.

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

Flores, A. W., Bechtel, K., and Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation, 80*, 38.

Fredman, S. (2011). *Discrimination law*. Oxford University Press.

Gelman, A., Fagan, J., & Kiss, A. (2007). An analysis of the New York City police department's "stop-and-frisk" policy in the context of claims of racial bias. *Journal of the American statistical association, 102(479)*, 813-823.

Gölz, P., Kahng, A., & Procaccia, A. D. (2019). Paradoxes in fair machine learning. *NeurIPS'19.*

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1, No. 2). Cambridge: MIT press.

Goodin, R. E., & Spiekermann, K. (2018). *An epistemic theory of democracy.* Oxford University Press.

Hackett, P. M., & Schwarzenbach, J. (2020). Black lives matter: Birdwatching in Central Park and the Murder of George Floyd. In *Handbook of Ethnography in Healthcare Research. Routledge.* 513-521.

Halldenius, L. (2017). Discrimination and irrelevance. In *The Routledge Handbook of the Ethics of Discrimination.* Routledge. 108-118.

Moritz Hardt, Eric Price, and Nathan Srebro. (2016). Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 3323–3331.

Harris, D. A. (2020). Racial profiling: Past, present, and future?. *Criminal Justice, 34*, 10.

Hedden, B. (2021). On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs, 49(2).*

Heidari, H., Loi, M., Gummadi, K. P., & Krause, A. (2019). A moral framework for understanding fair ml through economic models of equality of opportunity. In *Proceedings of the conference on fairness, accountability, and transparency.* 181-190.

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.

Hertweck, C., Heitz, C., & Loi, M. (2021). On the moral justification of statistical parity. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* 747-757.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association, 81(396)*, 945-960.

Holroyd, J. (2017). The Social Psychology of Discrimination. In *The Routledge Handbook of the Ethics of Discrimination.* Routledge. 381–93.

Kaufman, J. S., & Cooper, R. S. (1999). Seeking causal explanations in social epidemiology. *American journal of epidemiology, 150(2)*, 113-120.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & 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.* 616-626.

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

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. In *Aea papers and proceedings Vol. 108.* 22-27.

Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. (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.

Lauritzen, S. L. (1996). *Graphical models* (Vol. 17). Clarendon Press.

Lee Rodgers, J., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician, 42(1),* 59-66.

Lee, M. S. A., & Floridi, L. (2020). Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. *Minds and Machines,* 1-27.

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

List, C., & Goodin, R. E. (2001). Epistemic democracy: Generalizing the Condorcet jury theorem. *Journal of political philosophy, 9(3).*

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

Loi, M., Herlitz, A., & Heidari, H. (2021, July). Fair Equality of Chances for Prediction-based Decisions. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.* 756-756.

Menon, A. K., & Williamson, R. C. (2018). The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency.* 107-118. PMLR.

Mills, C. W. (2015). *Blackness visible: Essays on Philosophy and Race.* Cornell University Press.

Malinsky, D., & Bright, L. K., (2021). On the causal effects of race and mechanisms of racism. Unpublished manuscript.

Mawby, R. (1980). Sex and crime: The results of a self-report study. *The British journal of sociology, 31(4)*, 525-543.

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

Novaes, C. D. (2020). Carnapian explication and ameliorative analysis: a systematic comparison. *Synthese, 197(3)*, 1011-1034.

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

Pearl, J. (2010). The foundations of causal inference. *Sociological Methodology, 40(1),* 75-149.

Pearl, J. (2014). Interpretation and identification of causal mediation. Psychological methods, 19(4), 459.

Pettigrew, R. (2016). *Accuracy and the Laws of Credence.* Oxford University Press.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On fairness and calibration. In *Advances in Neural Information Processing Systems.* 5680-5689.

Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *NYUL Rev. Online, 94,* 15.

Rosenbaum, P. R. (2002). Overt bias in observational studies. In *Observational studies.* 71-104. Springer, New York, NY.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2(3)*, 169-188.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. Journal of the *American Statistical Association, 100(469)*, 322-331.

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.* 6414-6423.

Sen, A. K. (1971). Choice functions and revealed preference. *The Review of Economic Studies, 38(3)*, 307-317.

Simon, F., & Corbett, C. (1996). Road traffic offending, stress, age, and accident history among male and female drivers. *Ergonomics, 39(5)*, 757-780.

Spirtes, P., Glymour, C. N., Scheines, R., & Heckerman, D. (2000). *Causation, prediction, and search*. MIT press.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics, 25(1)*, 1.

Suzumura, K. (2009). *Rational choice, collective decisions, and social welfare.* Cambridge University Press.

Ulmer, J. T., Harris, C. T., and Steffensmeier, D. (2012). Racial and ethnic disparities in structural disadvantage and crime: White, Black, and Hispanic comparisons. *Social Science Quarterly, 93(3)*, 799-819.

Ulmer, J. T., and Steffensmeier, D. J. (2014). The age and crime relationship: Social variation, social explanations. In The nurture versus biosocial debate in criminology: On the origins of criminal behavior and criminality. SAGE Publications Inc. 377-396.

Varian, H. R. (2006). *Intermediate microeconomics with calculus: a modern approach*. Seventh Edition. WW Norton & Company.

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

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware).* IEEE. 1-7.

Walzer, M. (1983). *Spheres of justice: A defense of pluralism and equality.* Basic books.

Wang, X., & Ji, X. (2020). Sample size estimation in clinical research: from randomized controlled trials to observational studies. *Chest, 158(1)*, S12-S20.

Weinberger, N. (2019). Path-specific effects. *The British Journal for the Philosophy of Science, 70(1)*, 53-76.

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford university press.

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

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*. 1171-1180.

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