Chapter 2
Bias and Discrimination in Machine Decision-Making Systems

Jorge Casillas

Abstract There exists a perception, which is occasionally incorrect, that the presence of machines in decision-making processes leads to improved outcomes. The rationale for this belief is that machines are more trustworthy since they are not prone to errors and possess superior knowledge to deduce what is optimal. Nonetheless, machines are crafted by humans and their data is sourced from human-generated information. Consequently, the machine can be influenced by the same issues that afflict humans, whether that is caused by design inadequacies, by deliberately skewed design, or by biased data resulting from human actions. But, with an added problem, any failure of a machine is much more serious than that of a human; mainly due to three factors: they are massive, invisible, and sovereign. When machine decision-making systems are applied to very sensitive problems such as employee hiring, credit risk assessment, granting of subsidies, or medical diagnosis, a failure means thousands of people are disadvantaged. Many of these errors result in unfair treatment of minority groups (such as those defined in terms of ethnicity or gender), thus incurring discrimination. This chapter reviews different forms and definitions of machine discrimination, identifies the causes that lead to it, and discusses different solutions to avoid or, at least, mitigate its harmful effect.

2.1 Introduction

Human decision-making is a complex and nuanced process that involves a multitude of factors and variables. From the most mundane decisions like what to wear in the morning to the most critical ones like choosing a career or partner, our decision-making processes shape our lives and determine our future.

J. Casillas (✉)
Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain
e-mail: casillas@decsai.ugr.es

F. Lara, J. Deckers (eds.), *Ethics of Artificial Intelligence*, The International Library of Ethics, Law and Technology 41,
https://doi.org/10.1007/978-3-031-48135-2_2

Today, many of our decisions are conditioned by the assistance of automatic systems that help digest information to suggest the best decision or, in more and more situations, we delegate directly to these machines to decide for us. This delegation, in turn, is sometimes conscious but, in many other cases, we do it even without awareness.

Automatism helps humans by improving efficiency, thus reducing the time and effort required to complete tasks, reducing errors by avoiding fatigue, distractions, or oversight. Its more interesting aspect is its capability to provide innovation and launch us into doing things that we would never have thought we would be capable of.

Perhaps because of the attractive opportunities it offers, we are falling into a new alienation, trusting more and more in technology with unlimited power, forgetting along the way that machines also fail, and the consequences are a thousand times more serious. Indeed, machine learning algorithms are increasingly being used to make decisions—such as in hiring, lending, and criminal justice—that have significant impacts on people's lives.

When these algorithms are biased, they can perpetuate and even exacerbate existing discrimination and inequality. For example, if a hiring algorithm is biased against a certain demographic group, it can result in fewer members of that group being hired, thus depriving them of work experience and perpetuating the existing discrimination. Similarly, if a lending algorithm is biased against certain groups, it can result in those groups having less access to credit, perpetuating the existing economic inequality. In addition to perpetuating existing discrimination, machine discrimination can also lead to new forms of discrimination. Indeed, algorithms are so powerful these days discovering the unthinkable, that can learn to make predictions based on proxies for protected characteristics, such as ZIP codes or educational attainment, resulting in discriminatory outcomes for certain groups. It has already been observed how difficult it is to tame them (Dastin 2018).

Therefore, it is crucial to ensure that machine learning algorithms are designed and evaluated with fairness in mind, and that they do not perpetuate or create new forms of discrimination. By addressing machine discrimination, we can work towards a more equitable and just society. If we do not soon become aware of how serious the problem is, it will be very difficult to redirect the orientation of these automatisms to the point of being irreversible in certain scenarios. Some may accept risk for reward, but others may think that we are going too far and need to slow down and give ourselves time to reflect in this acceleration of artificial intelligence (AI) (Bengio et al. 2023).

Only a society that is aware and knowledgeable about these issues can build the foundations for a reliable development of AI. To this end, this chapter aims to open a space for reflection on the potential discriminatory danger of machine decision-making systems. I will begin by reflecting on why the failure of a machine is more serious than that of a human; then I will give an informative introduction to the basics of machine learning (without knowing it, it is not possible to gain awareness); I will continue by analyzing what is meant by machine discrimination; then I will identify the main reasons that cause such discrimination; and I will finish with

solutions to avoid or alleviate this discrimination or, in other words, how to ensure fairness.

2.2 Why Machine Failure Is More Serious

In later sections we will see what is meant by discrimination but, first, I want to start the chapter by deliberating on the importance of a machine discriminating. Only when we are able to visualize the magnitude of the problem and the scope that it has, we will be able to assess the seriousness of a machine discriminating and we will look for solutions to alleviate it.

Automation (control, big data, AI...) is used to gain efficiency (resources and time) and effectiveness (better performance). In the end, it all comes down to that, make something faster/cheaper, or better. Who is going to give up those advantages! Indeed, using automation to make decisions is a very tempting and sometimes unavoidable solution in today's society. Companies need it to remain competitive, people need it to do their jobs better, or to spend less time in an increasingly demanding world, or simply because the app we use on our mobile, or the website we visit, offers no alternative.

There is a (sometimes false) belief that when there is a machine involved in decisions, everything works better. We trust machines because they do not make mistakes and are better informed than we are to deduce what is best. However, machines are designed by humans, and machines are fed with data produced by humans. So, whether by design flaws, or by intentionally biased design, or by data collected from biased human behavior, in the end, the machine can be affected by the same problems as humans.

At this point, we might think that well, in the end, although a machine may also have biased behavior, it will be no worse than that of a human. But here we run into another reality, any failure of a machine is much more serious than that of a human. Mainly due to three factors. The automatisms executed by machines are, or can be, massive, invisible, and sovereign.

They are *massive* because they are highly scalable, making millions of decisions in a second. But any failure in its decision is magnified to a scale unthinkable for a human. Suppose a postal officer decides to make blacks wait in line twice as long as whites. First, it would be inconceivable and illegal in our times. But, at the end of the day, the impact of this discriminatory act would reach no more than 100 people. Now, when Amazon decided in 2016 to offer same-day service in select ZIP codes in major U.S. cities based on analysis of data about its customers, it marginalized neighborhoods where primarily blacks live. The impact of that measure affected millions of people every day and caused such a scandal that Amazon had to rectify (Ingold and Soper 2016).

They are *invisible* because the automatisms are often not perceived, there is no awareness that it is a machine that is behind the decision-making process. The case I have just cited of same-day service was very visible, but on other occasions, the

victim of certain decisions does not know that it was all or partly due to a machine. For instance, Deloitte, one of the Big Four accounting firms, explains that it uses powerful machine learning for Credit Risk Management (Phaure and Robin 2020)—they all do it, also the banks. When someone is turned down for a credit application, they simply say that the operation is not viable, they do not say there is an algorithm behind it, let alone how the algorithm arrived at that conclusion (usually because no one really knows how the machine made the decision). Advertising is also a good example of its invisibility; we already know that digital advertising is automated, but we got ads on our screens from Facebook that we do not know have been targeted because of our race, religion, or national origin (Benner et al. 2019).

They are *sovereign* because the machine does not usually "assume" responsibility, there is a lack of accountability. They simply decide, and their decisions are considered final. So it is when you search Google Images for 'unprofessional hairstyles', the photos that come up are mostly of black women. This is not the case if you search for 'professional hairstyles'. Since this became known in 2016 (Alexander 2016), the results have been nuanced, but differences are still observed nowadays (include '-google' in the query to avoid images related to the scandal). Google quickly disassociated itself from this by arguing that it did nothing, it is simply what you see on the internet. The question is whether something that many give more credence to than a god they think so, or do they really expect to find a reflection of society when searching.

Add to this, the belief that thinking that a machine is unbiased masks possible discriminatory effects, and that technology generates dependence on it, it does not turn back, to form a perfect storm. Finally, let us think about scenarios in which the machine already decides for itself, and on particularly sensitive matters such as granting a subsidy, a loan, passing a court sentence, or diagnosing a disease. In addition, of course, to other daily issues where machines condition your way of thinking, such as what news should be interesting for you and what videos you should watch.

Naturally, a well-designed automatic decision-making system would not cause these problems. In this chapter, I will focus on situations where this design is not correct or is subject to an interpretation of what is correct that might be questionable or skewed. Under these premises, a machine can be biased and lead to discriminatory decisions against certain social groups. Therefore, once the reasons for this malfunction have been identified, the chapter will also look at solutions for good design.

2.3 How Machine Learning Works

Before continuing, it is useful to briefly review how to build a model that, based on data, ends up supporting decision making in a given problem, or even directly makes the decision itself. The decision-making system (generally called *model*) most commonly used in AI is based on machine learning from data.

The data contain values of input variables (attributes or features) that determine specific cases, and, in the case of supervised learning (the most common type of machine learning), each data is accompanied by the value of the dependent variable (the one that constitutes the case study or target) that defines what should be the correct response of the system for that combination of input variables. This is why the data is called *example* (or instance), because it serves to teach the algorithm what is the optimal response it should give for a particular case.

When the system is built to predict a certain nominal variable, which takes different categories as possible values, it is called classification. For example, a model, based on different attributes/features such as the income of a family unit, fixed monthly expenses, work stability, outstanding debts, minors and other dependents under their care, can decide whether that family is worthy or not (two classes or categories, yes or no) of receiving a social subsidy. Another example could be using an automated system to recruit applicants. Based on several characteristics such as level of education, affinity of their education to the job position, previous work experience in the sector, outstanding work achievements, etc., the machine can decide to hire or not to hire.

I will return to this last example later. For now, let us move on to a less controversial one. Let us imagine that we want to build a model (decision-making system) that is able to classify traffic related images, something very useful in the driverless era. For instance, it could detect if the image corresponds to a car, a speed limit sign, or a traffic light. The process is depicted in Fig. 2.1.

To do this, we extract different attributes (features) of the images that describe what is there (things such as colors and shapes). Actually, in the era of deep learning (the most successful machine learning tool today), the image is sent raw to the

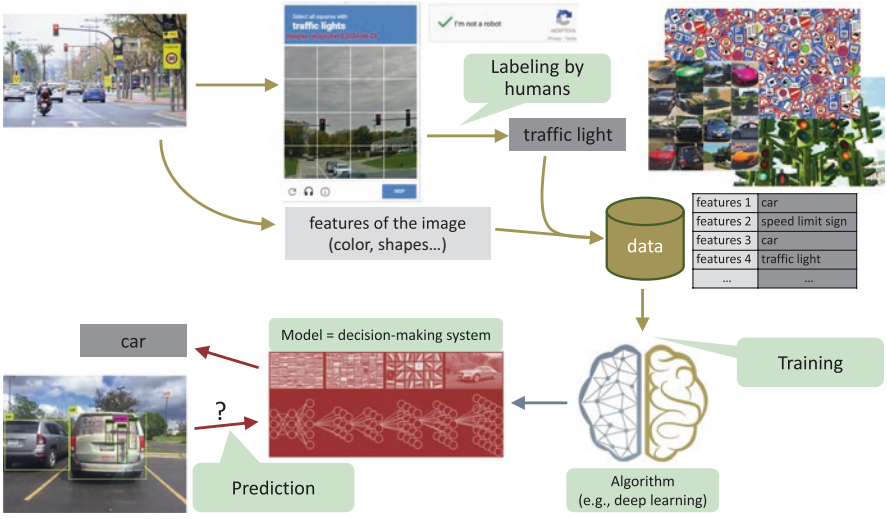

Fig. 2.1 Illustrative example of the machine learning pipeline, from data and labeling to model generation and subsequent prediction

algorithm, but for our purpose of illustrating how machine learning works in general, we will use features. To perform machine learning to build a model capable of automatic classification, we will need to label (or tag) each image. This labeling must be done by humans; this is how we transfer our knowledge to machines. It is still necessary today, despite the overwhelming power of the algorithms currently in use. So, either they pay ridiculous salaries as micro jobs (Hara et al. 2018) for a person to spend hours in front of a screen labelling, or they take advantage of the free labor of millions of people who every day fill out a reCAPTCHA to claim that 'we are not a robot' (Von Ahn et al. 2008).

At the end, we will build a data table where each row represents an example or instance, that is, a particular image. Each column will contain the value of each attribute/feature, plus a special last column that tells if that image is of a car, traffic light, etc. This data set constitutes all the knowledge we have about traffic images, and will serve to illustrate the algorithm, to teach it how to respond in each situation, i.e., it will serve to train it. Thus, in the training phase, the algorithm will build a model whose responses are as close as possible to the real ones in the hope that it will work as well (or even better) as the humans so that we can dispense with their services and use, from now on, the model built by the algorithm to classify traffic images.

Since errors will inevitably be made, we will have different ways of measuring where that error occurs, so there will be multiple possible measures of performance (cost functions), and part of the design of the algorithm will be to decide which measure is best for our interests. For instance, when diagnosing a disease, it is preferable to reduce false negatives (avoid missing someone who does have the disease), while a system that issues traffic tickets is preferable to minimize false positives (avoid issuing tickets to innocent drivers).

However, even if the ticketing machine is conservative, it cannot be so conservative that it does not serve its purpose, and it is not helpful to diagnose everyone for a disease either. Ultimately, the performance measure (cost function) should be a trade-off between hit and miss.

In a binary classification problem where a decision is made between two possible alternatives (usually called the positive and the negative class, the positive being the target of the problem), the outcomes of a classifier can be summarized in a contingency table that collects true positives (cases that are positive and are indeed predicted to be so), true negatives (cases that are negative and are so predicted), false negatives (cases that are positive but are erroneously predicted to be negative), and false positives (cases that are negative but are erroneously predicted to be positive). From these values, a series of measures are derived that assess the performance of the classifier from different points of view. Figure 2.2 shows those of interest to us throughout the chapter to define different fairness criteria.

In summary, we can see that in these machine learning tasks there are some key ingredients that determine the whole process. On the one hand, we have the data, which condense the human knowledge that we want to imitate. It is clear that biased data will lead to a biased algorithm. We also have the features, the variables with which we define each possible case, the lenses we use to see the real world. Poorly

		Predicted condition			
		Positive (PP)	Negative (PN)		
Actual condition	Positive (P)	True positive (TP)	False negative (FN)	True positive rate TPR = TP / P	False negative rate FNR = FN / P
	Negative (N)	False positive (FP)	True negative (TN)	False positive rate FPR = FP / N	True negative rate TNR = TN / N
	Prevalence = P / (P+N)	Positive predictive value PPV = TP / PP	Negative predictive value NPV = TN / PN		
	Accuracy = (TP+TN) / (P+N)				

Fig. 2.2 Contingency table and some performance measures derived therefrom

chosen or faulty features can also lead to undesired algorithm behavior. I have also mentioned the performance measure or cost function, which is the way we decide to quantify what is right or wrong, and consequently the algorithm will generate a model that satisfies that criterion in the best way it is able to find.

Other important design decisions are the structure and size of the model we want to generate. If it is too simple (for example, a small decision tree), it will have very low efficacy and will not be useful. If the system is excessively intricate (for example, a huge artificial neural network), it becomes challenging or even unfeasible to interpret. Consequently, we might not have the capability to elucidate the rationale behind decision-making, a necessity for addressing especially sensitive issues like those exemplified by social subsidy or hiring practices, as introduced earlier in this section.

Understanding the role played by each of these ingredients (data, variables, cost function, model type and size...) is key to identify risks where a bad design can lead to a discriminating machine.

2.4 What Is Meant by Machine Discrimination

Machine discrimination refers to the unfair treatment of individuals or groups based on the results of automated decision-making algorithms. A recurring question here is whether the machine discriminates, whether something inert can indeed discriminate. Or, ultimately, it is just a tool at the service of the human, who is the one who really discriminates.

I believe this is a superfluous question. A smokescreen to draw attention away from algorithms, to offload responsibility onto AI, and to frame the debate exclusively on humans. Although, for the moment, the truth is that there are no laws for algorithms, there are only laws for those humans who design and use those algorithms. Perhaps these words are still premature in 2024, to talk about an algorithm discriminating may seem futuristic, unrealistic, or simply sensationalist. I take that risk in this chapter. The reader will end up drawing his or her own conclusions after following this book and completing the puzzle with other sources.

In my view, the algorithm does discriminate. Something that makes an autonomous decision is responsible for its actions. So is the adult who decides based on the education he/she received from his/her parents. Perhaps in an early version of algorithms, say a decade ago, they were still in their teens and were somewhat irresponsible for their actions, as they were dedicated to supporting human decision making rather than deciding for themselves. But that phase is over, that screen has passed, what we have now well into the twenty-first century is a scenario in which there are algorithms designing other algorithms, machines trained with trillions of data representing trillions of real-world cases that decide, and their word is the law: unquestionable, irrefutable, irreversible.

In these preceding paragraphs I have deviated from the interest of this chapter, but they have been necessary to justify why throughout the text I will speak of machine discrimination. Those who do not recognize themselves with this definition can continue to speak of human discrimination through the machine…; the result, after all, is the same.

Let us look for some formality in this discrimination thing, although it is something that has already been dealt with in such depth and at such length, I refer the reader to other sources such as (Ntoutsi et al. 2020; Hardt et al. 2023) for a deeper understanding of the issue. Here I will limit myself to summarizing some keys that may help to explain the milestones of this chapter. I will try to do it in an informative way that brings this field closer to the general reader, so I will skip some excessively formal and rigorous descriptions.

We have seen how we built a machine to recognize traffic-related images. Suppose we reduce it to just recognizing whether there is a car in the image. This typical case is that of binary classification. Classification, because the decision consists of choosing (predicting) a category within a possible set of alternatives (with no order among them). Binary, because there are only two possible categories: there is a car or there is not. To now bring this problem of binary classification to the field where discrimination is relevant, let us substitute images for persons: ‘there is a car’ would stand for ‘being hired’.

Before continuing, we will call a decision-making system that chooses a response for a given situation a *decider*. This situation is measured through different attributes/features/variables, and among them there will be at least one that we will call *protected attribute*, that is, an attribute that determines a group against which discrimination could be exercised. Examples of protected attributes may be ethnicity, gender, or socioeconomic status.

A decider discriminates with respect to a protected attribute if for cases that only differ by their protected attribute, that decider makes different decisions (choose different classes). For example, if the machine systematically decides to hire a male person and not a female one, even though both are equally qualified, the system is discriminating against women and the protected attribute is ‘gender’.

To discriminate is to make an unfair decision, so discrimination can be measured in terms of fairness. Greater fairness means less discrimination. There is extensive literature around the definition of fairness (Mitchell et al. 2018; Barocas et al. 2017;

Gajane and Pechenizkiy 2017); here I just introduce the best known and used for machine learning.

2.4.1 Fairness Through Unawareness

A decider is said to achieve fairness through unawareness if protected attributes are not explicitly used in the decision process (Chen et al. 2019). In our example, if the decider ignores the gender to hire people, which can be easily done by simply hiding this attribute during training stage. This approach may be naïve because the interdependence with other factors may mean that, even without knowing the protected attribute, one is discriminating. For example, it may reward a type of work experience that, due to past discrimination, has been more accessible to one gender than to the other. Besides, while there may be situations where concealing the protected attribute is sufficient—think on blind auditions for musicians (Goldin and Rouse 2000)—, at other times it proves to be insufficient—e.g., in race-blind approaches (Fryer Jr et al. 2008).

In general, rather than relying on the system to decide fairly if it is not told the protected attribute, it is better to have control over the process to measure and regulate the degree of unfairness, for which it is necessary to know the value of the protected attribute.

2.4.2 Individual Fairness

A different approach is proposed in the seminar work of Dwork et al. (2012), where being aware that failure to control fairness leads to discriminatory systems, a mechanism is proposed to guarantee fairness based on an irrefutable fact: two equal cases (except for the difference that they belong to different groups) should be treated equally. In the hiring example, if a man and a woman are equally qualified (equal education, experience, professional achievements, etc.) when applying for a job, both should be treated equally in terms of being hired or not.

Since this definition considers fairness on an individual basis, the authors call it individual fairness: similar individuals are treated similarly. The question here is to define what is similar, how we assess that two individuals are similar except for the group to which they belong. To address this in machine learning, we must define a metric, i.e. a quantitative measure that assesses the degree to which two individuals are equal or not. On paper, this seems an ideal fairness criterion, provided that the metric is well chosen and made public. However, sometimes it is not an easy task to quantify social similarities. To the extent that this metric is fair, individual fairness will be achieved. In addition, we will get that on many occasions it will not be easy to find similar individuals among the data set; given an individual, we may have

difficulty finding his or her peer in the other group, so the individual fairness can be difficult to prove.

2.4.3 Counterfactual Fairness

Another way of looking at fairness is to analyze what would happen if an individual were to be changed from one group to another. Ideally, the system should have the same results, which would be a sign that it is making a fair decision. In other words, a decision is counterfactually fair toward an individual if it is the same in (a) the real world and (b) a counterfactual world in which the individual belonged to a different demographic group (Kusner et al. 2017).

Finding this counterfactual world is not so simple, it is not enough to flip the protected attribute. In fact, normally that attribute is not taught to the algorithm, it does not know it, so there is nothing to change there.

There really is a causal relationship that makes one attribute influence others. For example, even if race is not considered in hiring someone, that condition may have influenced a whole prior history with respect to the educational opportunities or prior work experience they had. Therefore, it becomes necessary to first define this causal graph (which itself may not find consensus) and then determine the process by which reversing the protected attribute that is not being directly observed triggers changes in other observable attributes that, in turn, propagate other changes according to that causal graph.

2.4.4 Group Fairness

Due to the above reasons that hinder achieving individual or counterfactual fairness, in most cases where machine learning is being applied, other measures are sought. In this way, instead of following an individual definition—that do value the specific discrimination of the individual—, group definitions are chosen—discrimination of an individual is not analyzed, but that of the group as a whole to which that individual belongs. In this type of group fairness criteria we find, in turn, several families of definitions:

- Demographic parity (also known as independence or statistical parity): it refers to a situation where the results of the decider ensure a proportional balance between the groups. For example, in a hiring process, to select a similar ratio of male and female. If ten employees are to be hired, five should be men and five women. It should be noted that this measure does not assess the correctness of the decision; it does not ultimately matter if those selected are well qualified, only the final proportion of the decision is assessed. The fifth best qualified candidate in one group may be much less qualified than the candidate holding a similar position in the other group.

- Equalized odds (also known as separation or positive rate parity): it measures the degree to which a decider provides similar rates of true positive (among all truly positive, how many are chosen as positive by the decider) and false positive (among all truly negative, how many are chosen as positive by the decider) predictions across different groups (Hardt et al. 2016). It can be relaxed to ensure only equal true positive rate (what is known as equal opportunity). In the hiring example, equal opportunity is guaranteed if in both groups (men and women) the same percentage of applicants is selected from all qualified candidates in that group. If, in addition, we have that the percentage of those selected among all the unqualified (those who should not deserve the position) is similar in both groups, we would have equalized odds.
- Predictive rate parity (also known as sufficiency): both the positive predictive ratio (among all cases where the decider chooses the positive class, how many of them are truly positive) and the negative predictive ratio (likewise for the negative class) are equal in the two groups. If a decider is 80% correct in choosing in the man group (8 out of 10 cases hired were actually qualified for the position), a similar percentage of correct choices should be made in the women group to ensure positive predictive ratio. If, in addition, the precision in denying employment is also similar, we would have predictive rate parity.

2.4.5 Impossibility of Fairness

The main challenge is that these three group measures of fairness are mathematically incompatible, they cannot be simultaneously satisfied (except under unrealistic circumstances). Therefore, satisfying two of them results in non-compliance with the third one. Even under certain conditions that are easily encountered in real problems, these three fairness conditions are mutually exclusive. It is known as the impossibility of fairness (Miconi 2017).

To illustrate the complexity of satisfying different criteria of fairness at the same time, I will borrow the clever example given by Zafar et al. (2017) but with some modifications that will better serve the purpose of our exposition. Suppose we want to build a classifier to decide whether to stop a person on suspicion of carrying a prohibited weapon. For this aim, we have a data set based on real cases where it was known whether the subject was carrying a weapon. As features, we know if the person had a visible bulge in his/her clothing and if he/she was in the vicinity of where a crime had been committed. We also know the gender (male or female), which is the protected attribute. It is shown in Table 2.1.

We will analyze the results shown in Table 2.2 of four classifiers (each one identified with ‘C’) that decide whether the subject should be stopped or not, depending on the case. For each of them, the results obtained with the three group fairness criteria are shown (positive difference of the corresponding measurement in the two groups), as well as the degree of accuracy achieved.

Table 2.1 Illustrative example of a data set and a hypothetical response from four classifiers

Gender	Clothing bulge	Prox. crime	Ground truth (has weapon)	C_1	C_2	C_3	C_4
m	Yes	Yes	Yes	Yes	Yes	Yes	Yes
m	Yes	No	Yes	Yes	Yes	Yes	No
m	No	Yes	No	No	Yes	No	No
f	Yes	Yes	Yes	Yes	Yes	Yes	Yes
f	Yes	No	No	No	Yes	Yes	No
f	No	No	No	No	Yes	No	No

Table 2.2 Results obtained by the four classifiers in Table 2.1 with respect to various measures of fairness

	C_1			**C_2**	
Demographic parity	DR diff.	33%	Demographic parity	DR diff.	0%
Equalized odds	TPR diff.	0%	Equalized odds	TPR diff.	0%
	FPR diff.	0%		FPR diff.	0%
Predictive rate parity	PPV diff.	0%	Predictive rate parity	PPV diff.	33%
	NPV diff.	0%		NPV diff.	0%
	Accuracy	100%		Accuracy	50%
	C_3			**C_4**	
Demographic parity	DR diff.	0%	Demographic parity	DR diff.	0%
Equalized odds	TPR diff.	0%	Equalized odds	TPR diff.	50%
	FPR diff.	50%		FPR diff.	0%
Predictive rate parity	PPV diff.	50%	Predictive rate parity	PPV diff.	0%
	NPV diff.	0%		NPV diff.	50%
	Accuracy	83%		Accuracy	83%

- C_1 is the same response as the ground truth knowledge, so the accuracy is obviously 100%. This is an inconceivable situation in a moderately complex real problem. There is no such thing as a perfect classifier, some mistake is always made. However, if this perfect classifier existed, although it would logically satisfy equalized odds (since TPR would be 100% and FPR 0% in both groups) and predictive rate parity (both PPV and NPV would be 100% in both groups), it could not guarantee demographic parity, since DR (demographic rate) would be 66.6% in men (two of the three men are stopped) and 33.3% in women. Nor does it comply with individual fairness, given that when faced with two identical cases that differ only in gender (both with a bulge in their clothing and not in the vicinity of a crime), the classifier decides to stop the man but not the woman.
- C_2 manages the problem with a simple decision: stop everyone. In this way, it gets to treat everyone with individual fairness, it also ensures demographic parity, and even holds equalized odds. However, while the PPV for men is 66.6% (of the three cases it decides to stop, it is right in two of them), for women it is 33.3% (it fails in two of the three cases). In addition, the classifier makes many errors (accuracy of 50%) with an FPR of 100% in both genders, which is totally unacceptable.

- C_3 and C_4 offer two alternative solutions with a good accuracy of 83% and guaranteeing demographic parity. In both cases, individual fairness is also achieved. However, C_3 has an FPR of 0% in males and 50% in females, while PPV is 100% in males and 50% in females. On the other hand, C_4 has a TPR of 50% in men and 100% in women, in addition to an NPV of 50% in men and 100% in women. Thus, none of them can satisfy either equalized odds or predictive rate parity.

Which classifier would one choose between C_3 and C_4? Or, in other words, if we are going to make a mistake, where do we want it to be made? If we prioritize the safety of no one escaping with a weapon, C_3 is better. If we prioritize the individual's right not to be unfairly stopped, C_4 is more appropriate. But both cases commit discrimination by treating men and women differently. Each problem has different social implications, and it is up to experts in the field to decide what orientation should be given to the machine decider, assuming that it will inevitably be biased.

2.5 What We Are Talking About: Example of Machine Discrimination

There are multiple examples where the use of machine learning has generated cases of discrimination (O'Neil 2016; Barocas et al. 2017; Eubanks 2018). There are also many available data sets on fairness (Fabris et al. 2022). Among all of them, if I have to choose one to illustrate the situation, I will inevitably choose the ProPublica case (Angwin et al. 2016), for multiple reasons. It is a situation of special social connotations where an error has a significant impact of discrimination, where the positions between those who use the system and those who suffer from it are drastically opposed, it is still being used and is on the rise, it has been widely studied and, in a way, it marked a before and after in the way of approaching machine learning by clearly revealing the difficulty of solving the problem with fairness awareness.

In the machine learning community, it is known as the ProPublica case after the name of the media wherein the investigation of four journalists was published under the title "Machine Bias" in 2016. It was already known that several U.S. courts use a decision support system that rates the risk that a defendant may reoffend with a score between 1 and 10—see, e.g., State v. Loomis (Liu et al. 2019). Even U.S. Attorney General Eric Holder warned in 2014 that the use of data-driven criminal justice programs could harm minorities (Holder 2014): "By basing sentencing decisions on static factors and immutable characteristics—like the defendant's education level, socioeconomic background, or neighborhood—they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society."

However, until the article in ProPublica, it had not been possible to demonstrate with data how the system that is known as COMPAS works. Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a software used for years (Brennan et al. 2009) developed by Northpointe (Equivant since 2017) as a

decision-making system powered by collected data from a defendant about different constructs such as current charges, criminal history, family criminality, peers, residence, social environment, or education, among others. Journalists were able to access records on thousands of cases collected between 2013 and 2014 in which they knew the risk assessment made by COMPAS and whether or not each person reoffended in the two subsequent years—which is the criterion used by the company to validate COMPAS (Equivant 2019). This data is now public and is being intensively used by the fairness machine learning community (Larson 2023). COMPAS was found to score African Americans at higher risk of recidivism than Caucasians, as shown in Fig. 2.3.

However, Northpointe argued that there really is equal treatment between blacks and whites because the hit rate is similar in both groups. For this, they refer to the PPV (positive predictive value), which measures the percentage of success in predicting the positive class (medium-high risk of recidivism) among all cases that are truly positive (actually recidivated). According to this measure, COMPAS obtains a value of 63% correct for blacks and 59% for whites, i.e. only a 4% difference, which is considered within the reasonably allowable range. As for the NPV (negative predictive value) measure, i.e., the percentage predicting non-recidivism among those

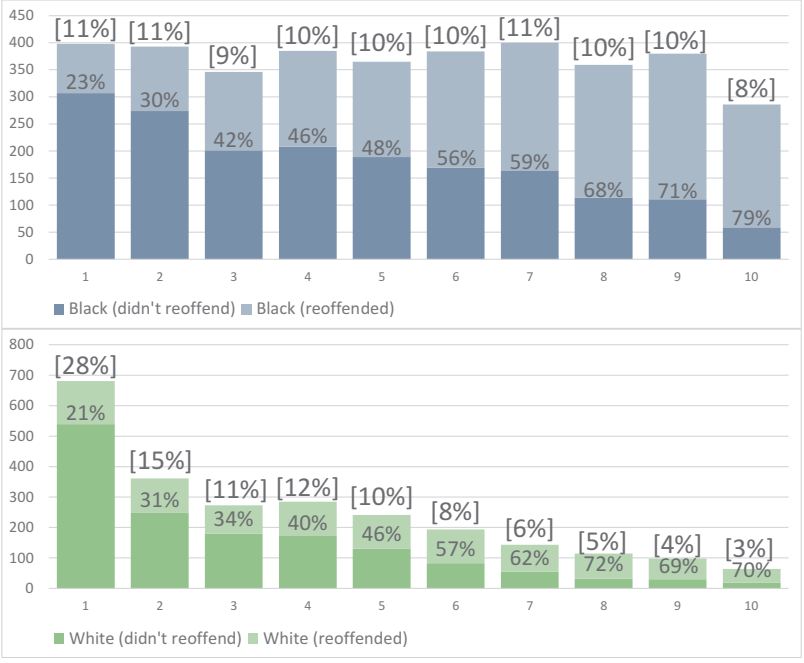

Fig. 2.3 Number of cases according to race and recidivism for each score assigned by COMPAS. Percentage of persons out of the total group assigned to each risk score is shown in square brackets above each bar. Percentage of repeat offenders for each set of scores is shown inside each bar. For example, 10% of blacks are scored with a risk of 9 versus 4% of whites. However, the recidivism rate among those who are rated a 9 is similar in both groups (71% vs. 69%)

who did not reoffend, the results were 65% for blacks and 71% for whites (6% difference). In short, from the point of view of the degree of precision, it could be said that the COMPAS treatment is fair.

But there is another way of looking at things. If we put the focus on the individual and the way in which he or she is severely handicapped by COMPAS in a prediction error, we can analyze FPR (false positive rate, the proportion of cases that do not recidivate despite receiving a high-risk score), where it is observed that there is a significant imbalance between the black and white groups, against the former. The error made in wrongly predicting that someone will reoffend when in fact she/he does not is 22 points higher for blacks (45%) versus whites (23%). Figure 2.4 compares the fairness interpretation of both Northpointe and ProPublica.

The reason for this huge discrepancy lies in a social reality: more blacks than whites recidivate in the data set, 51% versus 39% respectively, as shown in Fig. 2.5. Discussing the causes why there is more black recidivism is beyond the scope of this chapter, but it clearly has its roots in a multitude of historical circumstances. I will only note one reflection: what does recidivism mean? One might think that it refers to "the act of continuing to commit crimes even after having been punished," as defined by Cambridge Dictionary. But, in the eyes of procedural law in any state under the rule of law, recidivism is more than just committing a crime again... as a recidivist must also get caught.

A policing system that is focused more on catching blacks will logically find more crime and make more arrests in that community, reinforcing the perception that there is more crime among blacks. This is a classic chicken-egg example that generates a spiral from which it is difficult to escape. Whilst race is an attribute that is hidden from the machine, among the 137 features (variables) used by COMPAS,

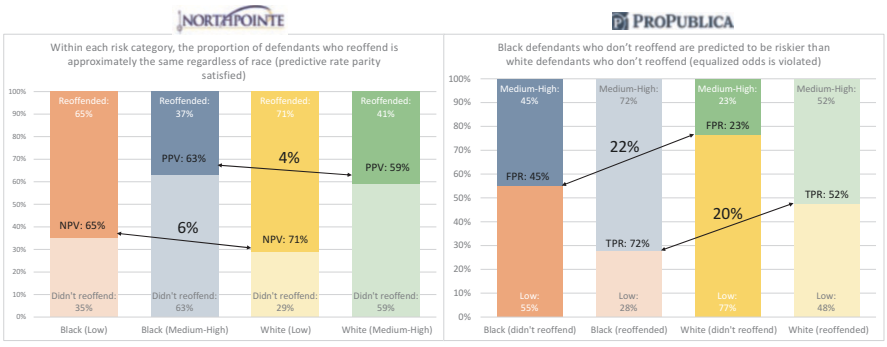

Fig. 2.4 Fairness results obtained in COMPAS as interpreted by Northpointe on the left (predictive rate parity satisfied) and ProPublica on the right (equalized odds violated). From Northpointe's point of view, the system is fair because it maintains a balance in PPV (percentage of cases that do reoffend among all cases that the system predicts will reoffend). However, from ProPublica's point of view, the system is not fair because there is a high disparity in FPR (percentage of cases that the system predicts will reoffend but actually do not) and TPR (percentage of cases that the system predicts will reoffend and actually do). Due to the difference in prevalence (the percentage of recidivists in the black group is higher than in the white group) as shown in Fig. 2.5, it is not possible to achieve both fairness criteria (balance of PPV and FPR at the same time), so it is crucial to understand the nature of the case and choose the most appropriate fairness criterion

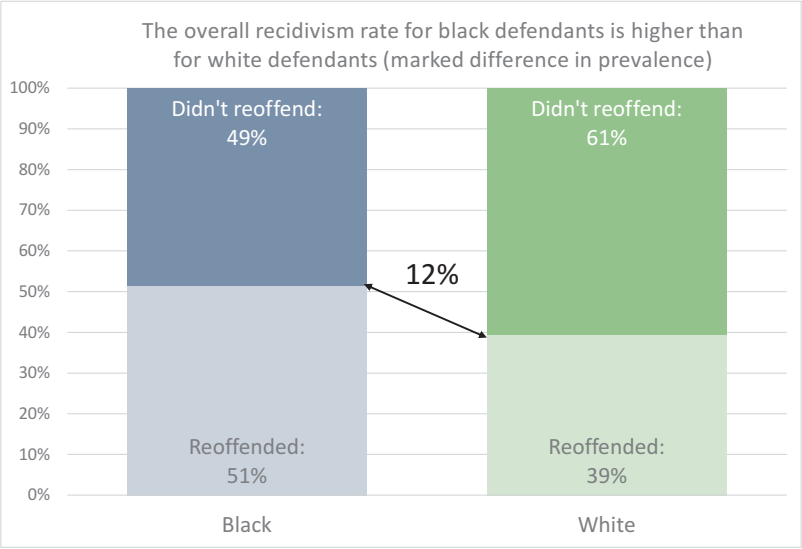

Fig. 2.5 The marked difference in prevalence between blacks and whites makes it impossible to reconcile the two fairness criteria. The reason why there is more recidivism in blacks is complex, but let us keep in mind that recidivism is not only repeating a crime, the subject must also be arrested and prosecuted

questions are asked and records are analyzed that disadvantage blacks (e.g., neighborhood or arrest history questions). So looking the other way does not solve the problem, it just lets it get out of control.

Indeed, in this case, where there is a marked difference of recidivism prevalence between the two groups, forcing the system to have an equal PPV mathematically makes unequal FPR and, therefore, discriminatory from that point of view. In other words, as it is not possible to satisfy different criteria of fairness at the same time (Chouldechova 2017), an in-depth study of the problem is necessary to design the best strategy. The issue here is that the efficiency interest (low false negative) of a company/administration is not aligned with the efficacy interest (low false positive) of the individual. For this reason, the ProPublica case opens an interesting debate that questions the convenience of using machine decision-making systems in a situation where there is no optimal solution and the error is critical, as it seriously harms the individual.

Finally, I show the result of a machine learning algorithm (Valdivia et al. 2021) that, based on a few variables together with the COMPAS prediction, can significantly improve its performance, as a proof of how fair we can go (see Fig. 2.6). Firstly, the COMPAS system is easily improved in terms of accuracy—see (Dressel and Farid 2018)—and, secondly, a well-designed algorithm can offer a range of possible best alternatives with different trade-offs of accuracy and fairness. As shown, the algorithm can generate solutions with an error rate similar to that obtained by COMPAS (35%), but with a difference in FPR between blacks and

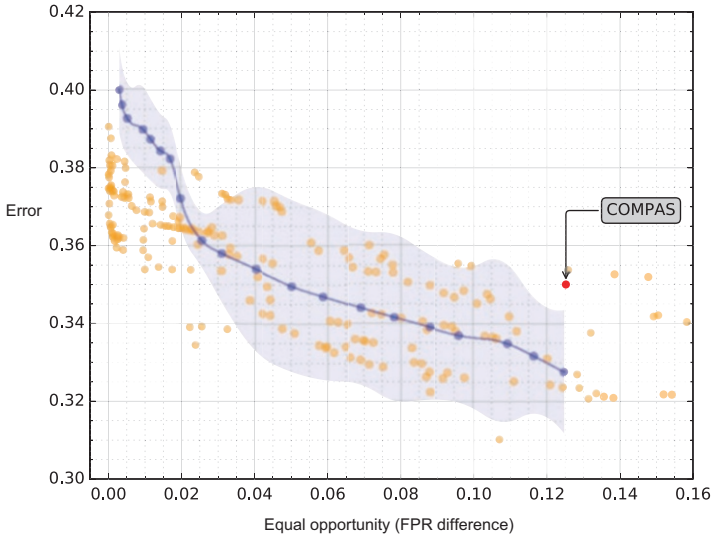

Fig. 2.6 Using a multi-objective optimization technique (Valdivia et al. 2021), multiple alternative classifiers can be generated. Each orange point is the result of an alternative decider, while the blue dots represent the average behavior. It can be seen how COMPAS (red dot) is easily outperformed in accuracy and fairness and that, with the same error rate as COMPAS (about 35%), it is possible to improve equality of opportunity significantly (reducing FPR difference from 12.5% to 4.5%)

whites of lower than 5% instead of the 12% that COMPAS obtained. In other words, if you want to be fairer, you can. It is just a matter of having fairness awareness and the machine learning skills necessary to employ algorithms that do not discriminate.

The ProPublica case exemplifies many of the bad practices that can lead to discrimination, such as the choice of variables, the selection of the set of examples, the use of a certain cost function to guide the algorithm, or the existence of proxies that correlate with race. In the next section I formalize some of the main reasons why a machine can discriminate.

2.6 Why Machine Learning Can Discriminate

Machine learning can end up generating decision-making systems that are unfair or cause discrimination because of different factors (Barocas and Selbst 2016). Sometimes it is due to the use of data that teaches the algorithm discriminatory behaviors previously performed by humans. At other times, a performance measure is used—which, in short, is the reference used by the algorithm to know whether the system it is generating is right or wrong—that leads to biased or unfair behavior.

The problem may also lie in the fact that the system supports decisions on variables that offer an incomplete or distorted view of reality.

Specifically, we can distinguish five causes that lead the algorithm to generate unfair or discriminatory deciders:

1. **The use of contaminated examples**: any machine learning system maintains the existing bias in the old data caused by human bias. Such bias can be aggravated over time as future observations confirm the prediction and there are fewer opportunities to make observations that contradict it.

   An example is the case of Amazon (Dastin 2018), when in 2014 it developed an automated recruitment system based on screening resumes. To do so, it trained its algorithm with the history of recruitment cases carried out by the company in its previous 10 years. The project ended up failing because it was found to have a gender biased behavior, favoring the hiring of men over women with equal qualifications. It was observed that the data used to train the algorithm were biased and, consequently, the algorithm learned to mimic that bias to maximize its performance.
2. **The choice of the wrong performance measure**: the algorithm is guided by a performance measure (cost function) to generate models that maximize (minimize) it; if that measure rewards a certain balance to the detriment of another, from a point of view, the solution could discriminate.

   The ProPublica case is a clear example of this (Angwin et al. 2016). While the company that developed COMPAS, the algorithm that scores an offender's risk of recidivism, claims that its system is fair because it satisfies equal PPV (positive predictive value, the proportion of subjects with a high risk score who actually recidivate) between blacks and whites, in terms of FPR (false positive rate, the proportion of cases that do not recidivate despite receiving a high risk score) there is a high imbalance between the black and white groups, against the former.
3. **The use of non-representative samples**: if the training data coming from the minority group are smaller than those coming from the majority group, the minority group is less likely to be perfectly modeled.

   Let us mention here the cases of AI models used for medical diagnosis based on genomic information. A 2016 meta-analysis analyzing 2511 studies from around the world found that 81% of participants in genome mapping studies were of European ancestry (Popejoy and Fullerton 2016). The data overrepresents people who get sicker. In addition, demographic information on the neighborhood where a hospital is located, how it advertises clinical trials, and who enrolls in them further exacerbates the bias. While another study (Aizer et al. 2014) pointed out that the lack of diversity of research subjects is a key reason why black Americans are significantly more likely to die of cancer than white Americans.

   In short, the survivorship bias (Brown et al. 1992) coined during the World War II when analyzing the impact of projectiles on aircraft—information from heavily damaged aircrafts, the most interesting to be analyzed, was not available as they did not return from battle—is still in force today, especially in the field of health: the algorithm cannot understand what is not shown to it.

4. **The use of limited attributes**: some attributes (variables) may be less informative or less reliably collected for minority groups, making the system less accurate on this group. This fact is aggravated when the variable that is not very representative is the object of study, i.e., the dependent variable or target.

   The scandal in the Netherlands where 26,000 families (mostly immigrants) had been unjustly classified as fraudsters through data analysis between 2013 and 2019, and thus forced to repay thousands of euros worth of benefits, is an example (Peeters and Widlak 2023). This was partly due to the use of a form that was cumbersome and difficult to understand for non-native Dutch speakers, so that erroneous data collected on it were interpreted as fraud for these families. In a nutshell, attributes were used that were not representative for certain social groups, causing them harm.
5. **The effect of proxies**: even if the protected attribute (e.g., ethnicity) is not used to train a system, other features may be proxies for the protected attribute (e.g., neighborhood). If such features are included, bias will occur, and it is sometimes very difficult to determine these dependencies.

   This is a classic case in the field of sociology. There are different causal relationships among variables that generate indirect relationships and back doors. An example is provided by Mitchell et al. (2018), where a complex historical process creates an individual's race and socioeconomic status at birth, both affecting the hiring decision, including through education. Even if we hide an individual's race and socioeconomic status from the algorithm, other variables, such as the quality of their education, will condition their chances of being hired. This variable is not likely to be independent, but likely to be entangled with their race and socioeconomic status.

Unfortunately, in a real-world complex problem, several of these factors combine, making it very difficult to avoid that machine learning does not discriminate, or that, being fair from one point of view, it is unfair from another. However, there are solutions for all these drawbacks. The key is to include validation mechanisms to detect these discriminations and the causes, in order to take action to mitigate discriminatory effects.

2.7 How Machine Discrimination Can Be Overcome

Fortunately, various solutions have been proposed from the machine learning community to address fairness. These techniques can help reduce bias in machine learning models, even if they are not a panacea. It is crucial to continually assess and monitor the performance of the model to ensure that it remains fair and unbiased.

Different approaches have been adopted that can be grouped into three categories, depending on the point in the machine learning pipeline at which the mechanism is incorporated to correct the operation of the process towards better fairness: the pre-processing stage, the in-training stage, or the post-processing stage. These

can be combined, so that it is possible to improve the data first, to apply algorithms designed for fairness subsequently, and finally to polish the results in post-processing. I will describe some existing alternatives in each approach.

2.7.1 Pre-processing for Fairness

Pre-processing approaches can be used to address fairness in machine learning by manipulating the input data before it is used to train a model. They attempt to obtain new representations of the data to satisfy fairness definitions. They are especially useful when the cause of the discrimination is due to biased data. Among the most common pre-processing techniques used for fairness, we find the following:

- Data sampling/weighting: one approach is to sample the data in a way that ensures equal representation of all groups. For example, if a dataset contains unequal representation of different races, we can oversample the underrepresented groups (or undersample the overrepresented one) to create a more balanced dataset (Kamiran and Calders 2010; Gu et al. 2020). Another way to reach a similar effect is to weight data (Krasanakis et al. 2018) in an iterative process, together with the algorithm (so it is a hybrid approach of pre-processing and in-training). Either by repeating (reducing) data from the minority (majority) group, or by giving more weight to data from the minority group, the objective of these techniques is to balance the representativeness of each group in the hope of reducing imbalances that affect fairness.
- Data generation: by using generative adversarial networks it is also possible to generate high quality fairness-aware synthetic data (Xu et al. 2018; Sattigeri et al. 2019). In this case, again, the aim is to create new minority group data, but in this case, it is fictitious data (based on real data), so there is more flexibility to direct the generation to reduce the fairness.
- Feature selection: another approach is to select features that are not biased towards any particular group (Grgić-Hlača et al. 2018). This can be done using statistical methods to identify features that have a low correlation with protected attributes (such as race or gender) or by using domain knowledge to select relevant features that do not discriminate. To the extent that there are attributes correlated with the protected attribute, this selection approach can be effective.
- Data encoding: to achieve fairness, we can frame it as an optimization challenge where we aim to discover an intermediate data representation that optimally captures the information while also hiding certain features that could reveal the protected group membership (Zemel et al. 2013; Calmon et al. 2017). Here we are looking for a data transformation aimed at reducing disparity that causes lack of fairness.
- Preprocessing with fairness constraints: some preprocessing techniques add constraints that promote fairness (Donini et al. 2018). We can also create a new attribute optimized by a kind of adversarial debiasing that trains a model to

minimize the accuracy of a discriminator that tries to predict protected attributes like race or gender (Zhang et al. 2018).

The main advantage of pre-processing is that the modified data can be used for any subsequent task. This helps people who are not very skilled in machine learning or who do not have access to the means to develop new algorithms on an *ad hoc* basis. Besides, there is no need to access protected attributes at prediction time, which is sometimes a limitation in projects where it is not legal or feasible to know that information. However, pre-processing approaches are inferior to in-training approaches in terms of performance of both accuracy and fairness, as well as less flexible compared to post-processing approaches.

2.7.2 In-training for Fairness

The methods in the training phase consist of modifying the classification algorithm by adding fairness criteria or by developing an optimization process considering these fairness measures. Since they work in the training phase, that is, when the algorithm determines how to generate the decider, they have great potential. Some possible in-training techniques are the following:

- Fairness cost-sensitive regularization: one approach is to incorporate fairness constraints into the model's objective function. This can be done by adding a regularization term that penalizes the model (decider) for making predictions that are biased towards certain groups. For example, we can use a fairness metric (usually the group ones) to penalize the model for making biased predictions or incorporate these measures to decide components of the model such as nodes, rules, or weights (Zafar et al. 2017b; Agarwal et al. 2018).
- Adversarial training: another approach is to use adversarial training to make the model more robust to biases in the input data (Kearns et al. 2018; Zhang et al. 2018). This involves training a discriminator model that tries to predict the protected attributes of the input data (such as race or gender) and using the output of the discriminator to update the classifier's weights. In other words, an algorithm oversees the potential discrimination caused by the models generated by another algorithm, so that iteratively the second one manages to improve the solution until it passes the approval of the first one.
- Counterfactual data augmentation: in some cases, it may be possible to generate counterfactual examples that help to mitigate bias in the input data. Counterfactual data augmentation involves generating new training examples that are similar to the original examples but with modified attributes that remove the bias following the given causal graph (Kusner et al. 2017).
- Individual fairness: given the metric to assess similarities among data, it can be used in training stage to force the algorithm to generate equal predictions for similar data (Dwork et al. 2012).

- Multiobjective optimization: finally, a powerful approach is to develop a wrapping scheme (a kind of meta-learning) where the hyperparameters of a standard algorithm are optimized to direct the learner to generate models with specific fairness measures (Valdivia et al. 2021; Villar and Casillas 2021). In this sense, fairness is not conceived as a constraint but as guide to optimize the model. When multiobjective optimization is incorporated in this meta-learning approach, it is possible to generate a wide variety of models with different accuracy-fairness trade-offs.

These in-training approaches achieve the best results in both accuracy and fairness, and they have a greater flexibility to choose the desired balance between them. Any fairness criteria can be incorporated at this stage. However, the adaptation of existing algorithms or the creation of new algorithms is required, which represents a major development effort. In some projects where fairness awareness is incorporated into previous developments, it is not easy to access previous algorithms.

2.7.3 Post-processing for Fairness

Post-processing methods aim to eliminate discriminatory decisions once the model has been trained by manipulating the output of the model according to some criteria. Here we have some common postprocessing techniques:

- Calibrated equality: an approach is adjusting the model's output using statistical methods or by using post-hoc adjustments to calibrate the output for each group (Canetti et al. 2019). This is the case of the equalized odds technique (Hardt et al. 2016) to ensure that the model predicts outcomes with equal false positive and false negative rates for all groups. This is possible in deciders that return an output with a degree of certainty, so that by varying the threshold of that certainty it is possible to alter the final output. For example, output 0 means that the loan is not granted and 1 that it is, so that, initially, a certainty greater than 0.5 concludes that the loan is granted (1) and a lower value that it is not (0). In this case, the 0.5 border can be varied, for example to 0.7, as long as better equity between groups is achieved. As can be seen, neither the data nor the algorithm is modified, what is altered is the output of the decider generated by the algorithm.
- Rejection sampling: in some cases, it may be necessary to reject certain predictions that are likely to be biased. *Rejection sampling* is a technique that involves refusing predictions that are too confident or too uncertain (Kamiran et al. 2018). By doing so, the model can avoid making biased predictions that are likely to be incorrect.
- Model regularization: finally, model regularization can be used to ensure fairness in machine learning. Regularization techniques can be used to penalize/constrain/modify the model to produce outputs that are consistent with a fairness

metric (Pedreschi et al. 2009; Calders and Verwer 2010; Kamiran et al. 2010). Here a transformation of the generated decision plane is pursued in order to improve fairness. It is a similar effect to calibrated equality but through a more complex and potentially more effective process.

These post-processing approaches have a good performance, especially with group fairness measures. As in the case of pre-processing approaches, there is no need to modify the algorithm, it can work with standard machine learning algorithms. However, these approaches are more limited than in-training approaches to get the desired balance between accuracy and fairness. Moreover, the protected attribute must be accessed in the prediction stage, which is sometimes not possible. Whilst the sensitive information can be known during the development of the algorithm for its correct design, it will not always be available (for legal or practical reasons) at the time of using the already trained model.

2.8 Conclusion

Throughout the chapter we have seen many examples where the use of algorithms and massive data analysis generate discrimination based on gender, race, origin, or socioeconomic status in fields as diverse as recruitment, recidivism assessment, advertising, internet searches, credit risk assessment, subsidy payment, health treatment, and online shopping delivery. When discrimination is generated by automatisms, their consequences are much more serious due to its scalability, invisibility, and lack of accountability. Different perspectives on discrimination have been reviewed and different causes have been analyzed. We have also been able to see how there are solutions for most cases, as long as there is a will and the means to solve them.

AI will continue to bring wonderful things to society, but it must also be constrained by the values of that same society. It is advancing at great speed. The best way to deal with this evolution is to move at the same pace in ethics, awareness, education, and regulation.

The legislature should work closely with experts to investigate, prevent, and mitigate malicious uses of AI. AI experts must take the nature of their work seriously, proactively communicating with relevant stakeholders when harmful applications are foreseeable. External audits should be incorporated into potentially discriminatory projects, both in the private and public sectors. We should bring together data scientists and experts in social sciences to infuse a social lens into solutions and examine potential discriminations.

I hope that this chapter contributes to raising awareness about the serious discriminatory potential that machine decision-making systems can have, open the eyes of AI experts to the consequences of their work, and show that, although there are solutions to alleviate it, these are not as simple as fixing a bias in the data.

References

Agarwal, A., A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. 2018. A reductions approach to fair classification. In *International conference on machine learning*, 60–69. PMLR.

Aizer, A.A., T.J. Wilhite, M.H. Chen, P.L. Graham, T.K. Choueiri, K.E. Hoffman, et al. 2014. Lack of reduction in racial disparities in cancer-specific mortality over a 20-year period. *Cancer* 120 (10): 1532–1539.

Alexander, L. 2016. Do Google's 'unprofessional hair' results show it is racist? *The Guardian*. https://www.theguardian.com/technology/2016/apr/08/does-google-unprofessional-hair-results-prove-algorithms-racist-.

Angwin, J., J. Larson, S. Mattu, and L. Kirchner. 2016. *Machine bias*. ProPublica, May 23, 2016.

Barocas, S., and A.D. Selbst. 2016. Big data's disparate impact. *California Law Review*: 671–732.

Barocas, S., M. Hardt, and A. Narayanan. 2017. Fairness in machine learning. *Nips tutorial* 1: 2017.

Bengio, Y., et al. 2023. *Pause giant ai experiments: An open letter*. Future of Live Institute. https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Benner, K., G. Thrush, M. Isaac. 2019. Facebook engages in housing discrimination with its ad practices, U.S. says. *The New York Times*, March 28. https://www.nytimes.com/2019/03/28/us/politics/facebook-housing-discrimination.html.

Brennan, T., W. Dieterich, and B. Ehret. 2009. Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior* 36 (1): 21–40.

Brown, S.J., W. Goetzmann, R.G. Ibbotson, and S.A. Ross. 1992. Survivorship bias in performance studies. *The Review of Financial Studies* 5 (4): 553–580.

Calders, T., and S. Verwer. 2010. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21: 277–292.

Calmon, F., D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K.R. Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Advances in neural information processing systems*, vol. 30, NIPS.

Canetti, R., A. Cohen, N. Dikkala, G. Ramnarayan, S. Scheffler, and A. Smith. 2019. From soft classifiers to hard decisions: How fair can we be? In *Proceedings of the conference on fairness, accountability, and transparency*, 309–318. ACM.

Chen, J., N. Kallus, X. Mao, G. Svacha, and M. Udell. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, 339–348. ACM.

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

Dastin, J. 2018. *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters.

Donini, M., L. Oneto, S. Ben-David, J.S. Shawe-Taylor, and M. Pontil. 2018. Empirical risk minimization under fairness constraints. In *Advances in neural information processing systems*, vol. 31. NIPS.

Dressel, J., and H. Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4 (1): eaao5580.

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

Equivant. 2019. *Practitioner's guide to COMPAS core*. https://www.equivant.com/practitioners-guide-to-compas-core/.

Eubanks, V. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Fabris, A., S. Messina, G. Silvello, and G.A. Susto. 2022. Algorithmic fairness datasets: The story so far. *Data Mining and Knowledge Discovery* 36 (6): 2074–2152.

Fryer, R.G., Jr., G.C. Loury, and T. Yuret. 2008. An economic analysis of color-blind affirmative action. *The Journal of Law, Economics, & Organization* 24 (2): 319–355.

Gajane, P., and M. Pechenizkiy. 2017. On formalizing fairness in prediction with machine learning. arXiv preprint arXiv:1710.03184.

Goldin, C., and C. Rouse. 2000. Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review* 90 (4): 715–741.

Grgić-Hlača, N., M.B. Zafar, K.P. Gummadi, and A. Weller. 2018. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1. AAAI.

Gu, X., P.P. Angelov, and E.A. Soares. 2020. A self-adaptive synthetic over-sampling technique for imbalanced classification. *International Journal of Intelligent Systems* 35 (6): 923–943.

Hara, K., A. Adams, K. Milland, S. Savage, C. Callison-Burch, and J.P. Bigham. 2018. A data-driven analysis of workers' earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, 1–14. ACM.

Hardt, M., E. Price, and N. Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 29. NIPS.

Hardt, M., S. Barocas, and A. Narayanan. 2023. *Fairness and machine learning: Limitations and opportunities*. The MIT Press. (ISBN 9780262048613).

Holder, E. 2014. *Attorney general Eric holder speaks at the national association of criminal defense lawyers 57th annual meeting and 13th state criminal justice network conference*. The United States Department of Justice.

Ingold, D., and S. Soper. 2016. *Amazon doesn't consider the race of its customers*. Should it. Bloomberg, April, 21.

Kamiran, F., and T. Calders. 2010. Classification with no discrimination by preferential sampling. In *Proceedings 19th machine learning Conference Belgium and The Netherlands*, vol. 1, no. 6. Citeseer.

Kamiran, F., T. Calders, and M. Pechenizkiy. 2010. Discrimination aware decision tree learning. In *2010 IEEE international conference on data mining*, 869–874. IEEE.

Kamiran, F., S. Mansha, A. Karim, and X. Zhang. 2018. Exploiting reject option in classification for social discrimination control. *Information Sciences* 425: 18–33.

Kearns, M., S. Neel, A. Roth, and Z.S. Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, 2564–2572. PMLR.

Krasanakis, E., E. Spyromitros-Xioufis, S. Papadopoulos, and Y. Kompatsiaris. 2018. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 world wide web conference*, 853–862. ACM.

Kusner, M.J., J. Loftus, C. Russell, and R. Silva. 2017. Counterfactual fairness. In *Advances in neural information processing systems*, vol. 30. NIPS.

Larson, J. 2023. *COMPAS recidivism risk score data and analysis*. ProPublica, April 2023. https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis.

Liu, H.W., C.F. Lin, and Y.J. Chen. 2019. Beyond state v Loomis: Artificial intelligence, government algorithmization and accountability. *International journal of law and information technology* 27 (2): 122–141.

Miconi, T. 2017. The impossibility of "fairness": A generalized impossibility result for decisions. arXiv preprint arXiv:1707.01195.

Mitchell, S., E. Potash, S. Barocas, A. D'Amour, and K. Lum. 2018. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. arXiv preprint arXiv:1811.07867.

Ntoutsi, E., P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.E. Vidal, et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10 (3): e1356.

O'Neil, C. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown Books. (ISBN 978-0553418811).

Pedreschi, D., S. Ruggieri, and F. Turini. 2009. Measuring discrimination in socially-sensitive decision records. In *Proceedings of the 2009 SIAM international conference on data mining*, 581–592. Society for Industrial and Applied Mathematics.

Peeters, Rik, and Arjan C. Widlak. 2023. Administrative exclusion in the infrastructure-level bureaucracy: The case of the Dutch daycare benefit scandal. *Public Administration Review* 83: 1–15. https://doi.org/10.1111/puar.13615.

Phaure, H., and E. Robin. 2020. *Artificial intelligence for credit risk management*. Deloitte. https://www2.deloitte.com/content/dam/Deloitte/fr/Documents/risk/Publications/deloitte_artificial-intelligence-credit-risk.pdf.

Popejoy, A.B., and S.M. Fullerton. 2016. Genomics is failing on diversity. *Nature* 538 (7624): 161–164.

Sattigeri, P., S.C. Hoffman, V. Chenthamarakshan, and K.R. Varshney. 2019. Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development* 63 (4/5): 3–1.

Valdivia, A., J. Sánchez-Monedero, and J. Casillas. 2021. How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness. *International Journal of Intelligent Systems* 36 (4): 1619–1643.

Villar, D., and J. Casillas. 2021. Facing many objectives for fairness in machine learning. In *Quality of information and communications technology: 14th international conference, QUATIC 2021, Algarve, Portugal, September 8–11, 2021, proceedings*, vol. 1439, 373–386. Springer International Publishing.

Von Ahn, L., B. Maurer, C. McMillen, D. Abraham, and M. Blum. 2008. Recaptcha: Human-based character recognition via web security measures. *Science* 321 (5895): 1465–1468.

Xu, D., S. Yuan, L. Zhang, and X. Wu. 2018. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE international conference on big data (big data)*, 570–575. IEEE.

Zafar, M.B., I. Valera, M. Gomez Rodriguez, and K.P. Gummadi. 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. ACM.

Zafar, M.B., I. Valera, M.G. Rogriguez, and K.P. Gummadi. 2017b. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, 962–970. PMLR.

Zemel, R., Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. 2013. Learning fair representations. In *International conference on machine learning*, 325–333. PMLR.

Zhang, B.H., B. Lemoine, and M. Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society*, 335–340. ACM.