Just Machines

Clinton Castro
Florida International University

Abstract

A number of findings in the field of machine learning have given rise to questions about what it means for automated scoring- or decision-making systems to be fair. One center of gravity in this discussion is whether such systems ought to satisfy *classification parity* (which requires parity in accuracy across groups, defined by protected attributes) or *calibration* (which requires similar predictions to have similar meanings across groups, defined by protected attributes). Central to this discussion are impossibility results, owed to Kleinberg et al. (2016), Chouldechova (2017), and Corbett-Davies et al. (2017), which show that classification parity and calibration are often incompatible. This paper aims to argue that classification parity, calibration, and a newer, interesting measure called counterfactual fairness are unsatisfactory measures of fairness, offer a general diagnosis of the failure of these measures, and sketch an alternative approach to understanding fairness in machine learning.

Whether or not it draws on new scientific research, technology is a branch of moral philosophy, not of science

-Paul Goodman

1 Introduction

Increasingly, judgments formed using patterns found in datasets are being used to project a group or individual's characteristics in order to make decisions about them. The development of these data-driven judgments often

involves the techniques of *machine learning*, the study of programming computers to learn from data. Perhaps the most familiar example of such a judgment is the credit score, a metric for predicting how likely an individual is to repay debt. In recent years, data, data storage, and computing power—the materials needed for developing and assigning these data-driven judgments at mass scale—have become cheap and abundant. As a result, they are now ubiquitous.[1]

Recently, there has been an explosion of interest in the question of how to determine whether a data-driven judgment is "fair", that is, (roughly) not wrongfully biased against members of protected classes.[2,3] Bias against members of protected classes is especially pernicious in this context because data-driven judgement systems have the capacity to reinforce and perpetuate oppressive social practices at scale. The interest in the question of how to determine whether a data-driven judgment is fair has given rise to the study of *fair machine learning*, the study of how to detect and mitigate bias against members of protected classes in judgments developed using the techniques of machine learning.

One center of gravity of the current conversation about fair machine learning is Angwin et al. (2016), a ProPublica investigation of Northpointe Inc.'s COMPAS[4]. COMPAS is software that generates data-driven judgments (which,

[1]If this is not apparent, consider *consumer scores*, scores used to project consumers' characteristics (Dixon and Gellman (2014)). There are consumer scores for just about anything, such as determining an individual's age, ethnicity, gender, frequency of purchasing general apparel, television usage, job security, allegiance to buying name-brand or generic drugs, likelihood of moving to another merchant, likelihood of smoking, and likelihood of being pregnant (Dixon and Gellman (2014)). There are scores for many hundreds, if not thousands, of other characteristics (Dixon and Gellman (2014)). These scores are used to determine which advertisements a consumer sees, which services she is eligible for, and what price she will pay (which might differ from that of someone else buying the same product at the same time from the same merchant (Turow (2017)). Some consumer scores stand in as un- or under-regulated alternatives to credit scores and are used in decisions relating to lending, housing, and employment (Pasquale (2015)). Each of us, whether we know it or not, have many of these scores assigned to us (Dixon and Gellman (2014)).

[2]See, e.g., Kleinberg et al. (2016); Angwin et al. (2016); Northpointe Inc. (2016); Chouldechova (2017); Corbett-Davies et al. (2017); Corbett-Davies and Goel (2018); Huq (2019).

[3]Note that "fair" does not generally have such a narrow meaning. However, in the context of fair machine learning, this typically what is meant by "fair." In what follows, this is what I will mean by "fair," unless I specify otherwise.

[4]COMPAS is an acronym for Correctional Offender Management Profiling for Alter-

in this context, are usually called "risk scores") describing an individual's risk of recidivism. Its scores are used to assist in decisions pertaining to sentencing, bail, parole, and pretrial release. Angwin et al. (2016) purported to find that COMPAS was unfair: they discovered that the software violated a measure known as predictive equality, as it falsely identified black defendants as future criminals at almost twice the rate as white defendants. COMPAS also falsely identified white defendants as low risk more often than black defendants. Northpointe Inc. defended COMPAS, saying that the software was in fact fair. The company demonstrated that COMPAS satisfied another ideal known as calibration, as within each COMPAS-generated "risk group" (which are "low-," "medium-," and "high-risk"), defendants reoffended at similar rates, regardless of race (Northpointe Inc. (2016)).

Paradoxically, both Angwin et al. (2016) and Northpointe Inc. (2016) were right about the underlying facts: COMPAS falsely identifies black defendants as future criminals at higher rates than white defendants; yet, within any given risk group, COMPAS's accuracy rates are uniform across racial groups. As Kleinberg et al. (2016), Chouldechova (2017), and Corbett-Davies et al. (2017) have since shown, in many contexts—including the one COMPAS operates in—parity in overall error rates and parity in accuracy rates within risk groups are mutually exclusive options (this result is described in detail in section 3.2). Whether COMPAS is fair, then, depends on which—if either—of these standards tracks the proper conception of fairness.

In this paper, I survey some of the central concepts of machine learning. I present two measures of fairness that are popular in the machine learning literature, one corresponding to the standard that ProPublica used (i.e., predictive equality) and the other corresponding to the the standard Northpointe Inc. used (i.e., calibration). I show each, as well as a newer measure called counterfactual fairness, to be unsatisfactory. I then argue that these measures ultimately fail for the same reason: they merely track *formal fairness*, the equal and impartial application of rules (Hooker (2005)), which is too narrow a conception of fairness to do the job requested of fairness measures, i.e., to detect wrongfully biased judgements against members of protected groups. This is because, as we will see, the issues that matter in fair machine learning are not merely matters of formal fairness: securing formal fairness is neither necessary nor sufficient for avoiding making wrongfully

native Sanctions

biased judgements against members of protected groups.[5]

2 Machine learning and fairness

To discuss these issues, it will be helpful to have some background on the basics of machine learning. In this section, I provide a very simple example of how one might render a data-driven judgement using machine learning.[6] The example is not meant to give an insight into how all machine learning and data-driven judgment-making works. Instead, it is meant to give those unfamiliar with machine learning enough of a sense of how it works to understand how judgement calls and data collection errors can contribute to the construction of unfair machine learning systems.

2.1 How machines learn: a brief introduction[7]

Let's suppose you have trouble telling Boston Terriers and French Bulldogs (Frenchies) apart and want to develop a system for telling the difference. To get better, you might go into the local dog park with an expert, make some notes on the dogs the expert classifies as Boston Terriers and Frenchies, look for patterns in your notes, and use those patterns as a guide to future attempts at categorizing Boston Terriers and Frenchies. If you did this in a rigorous way—such as the one outlined below—you'd be doing the same thing that a computer does when it uses machine learning to mine data sets to find patterns to render data-driven judgements. By understanding the method, classification, we can achieve an understanding of some of the basics of machine learning.

Let's look at classification, step-by-step.

At step one, you choose what to take notes on. Before you go into the field, you want to know what you plan to take notes on (Length? Weight? Etc.) When you choose what to take notes on, you are engaging in something called *feature selection*. When you choose which labels to use for the thing

[5]Many thanks to Gerard Vong for helping me see this.

[6]A bit more specifically, the example is an example of *supervised machine learning*, machine learning that involves the use of a labeled data set. *Unsupervised machine learning*, in contrast, is machine learning that does not involve the use of any such sets. For helpful introduction to unsupervised learning, see Gerrish and Scott (2018), especially chapter 7.

[7]This presentation was greatly influenced by Green et al. (2017).

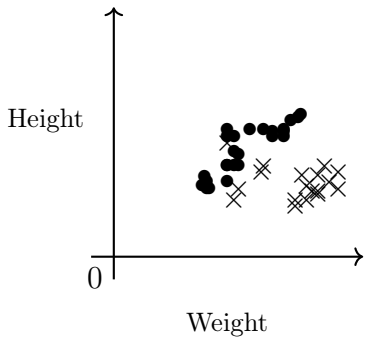
you care about (Boston Terrier, Frenchie, etc.) you are defining your *target variables*, the variables you want to learn about.

At step two, you take notes. When you go into the field with an expert, you're engaging in *data collection*, with the goal of creating a data set that you will try to learn from. You will use this data to train yourself to better tell Boston Terriers from Frenchies; this data set is your *training data*.[8]

Your notes (data) might look like this:

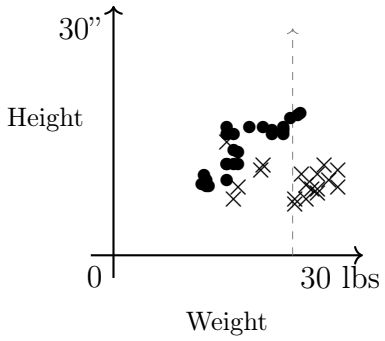
Height	**Weight**	**Breed**
15"	15lbs	Boston Terrier
11"	25lbs	Frenchie
12"	28lbs	Frenchie
17"	16lbs	Boston Terrier
...	...	...

At Step three, you find patterns. Here's one way to organize the training data such that the pattern in it are obvious (where •'s represent Boston Terriers and ×'s represent Frenchies):

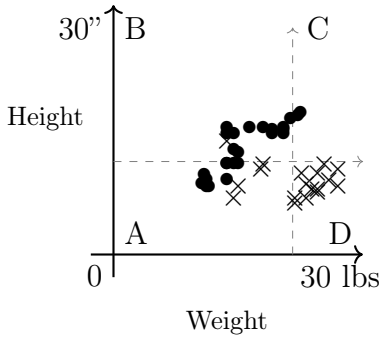


[8]Note that even though feature selection precedes data collection in this example, it isn't the case that feature selection must precede data collection. In fact, these steps are often done in the reverse order. I assumed height and weight will help us discriminate between Boston Terriers and Frenchies, but these might not be (and indeed probably are not) the best features to use for our predictions. In a more sophisticated example, we might start with data that we have already collected (perhaps you own a pet food company and already have a large dataset about about your customer's animals), and then—using the data you already have—make a determination of which features are predictive of the target variable. This complication is worth noting because features selected via certain methods might be inscrutable to us and could track membership to a protected class.

Imagine you want a simple rule for telling Boston Terriers from Frenchies. So, you draw a vertical line that puts most $\times$'s to the right while putting most $\bullet$'s to the left. Let's say the line belongs here:



Let's do the same, with a horizontal line.



Notice that we now have quadrants.

Now we can define a rule: *If you come across a Boston Terrier or Frenchie, take its height and weight.* If the data you collect places you in quadrant D, classify it as a Frenchie. Otherwise, classify it as a Boston Terrier. We'll call the rule for making classifications a *classification rule*; the classification itself is an example of a data-driven judgement.

This will be an accurate way to discriminate Boston Terriers from Frenchies. You will make some mistakes, but much fewer than you did before. The mistakes you will make with respect to the training data are described in this *confusion matrix* (a table that describes the performance of a classification rule with respect to a particular data set):

	Classified as: Boston Terrier	Classified as: Frenchie
Actually: Boston Terrier	23	0
Actually: Frenchie	5	13

If the degree of accuracy supplied by this rule is sufficient for your purposes, you are done. You're ready to use the rule to classify future Boston Terriers and Frenchies. If not, you may want to collect more data, track more than two features, and experiment with drawing angled and maybe curvy lines (as opposed to drawing two perpendicular lines) to track patterns.

As we can see from these suggestions, classification can get difficult and complicated quickly. To improve accuracy, you might track more than two features, which means you're tracking patterns that might be hard for humans to spot. You might also use angled and curved dividers to sort patterns across several dimensions. This is difficult work for the unaided human mind, but common computers can do it with ease. This is why methods like classification are associated with machines. Computers are great at classification in contexts where humans struggle. Further, computing is cheap and using the methods of machine learning is an effective way to increase accuracy across all sorts of contexts, such as lending, predicting future criminality, and advertising. This is why machine learning has become so popular.

2.2 How machines learn to be unfair[9]

There are many ways to inadvertently construct a rule via classification whose application would be unfair. Here, I'll show how unfairness can be introduced at each stage of the process.

Unfairness can be introduced through choices about what to take notes on. Consider a case:

> **Hiring Teachers.**[10] You are hiring teachers and want to know which ones will be effective. You decide to look to your past hires to see which teachers added the most "value" to their students' education, measured using test scores from the year before and

[9]The conceptual organization of this section was influenced by Barocas and Selbst (2016).

[10]This example is loosely based on a real case, described in Quick (2015).

> the year after that student had them as a teacher. You find that teachers that went to certain colleges added more value. So, you only call back teachers that went to those schools.

In our example, you have located a pattern that exists in the data, but this does not mean that using the rule will be fair. To see this, suppose that the following correlations hold in the school districts you run: Minority teachers gravitate towards schools with higher portions of minority students; The higher the portion of minority students, the poorer the funding for that school; Poor funding causes test scores to be stubborn; The colleges that have produced "effective" teachers are not diverse. You've now introduced unfairness into the hiring process that wasn't there before, and you've done this simply by choosing which features to track.

Unfairness can also be introduced via uneven data collection. Consider another case:

> **Pretrial Release.** You are deciding who to release while awaiting trial. You only want to release people who will not reoffend while awaiting trial. You have data about past defendants that suggests that the number of past arrests a defendant has is positively correlated with future arrests. You use this to construct a rule that recommends only releasing those with few previous arrests.

Again, we have located a pattern that exists in the data, but that does not mean that using the resulting classification rule will be fair. We can suppose further that in our case—as in reality, sadly—there are many types of crimes which white people and black people commit at similar rates, but that black people get arrested for much more often (Bunting et al. (2013)). Using arrests as a proxy for reoffense, then, is a way of formalizing and reinforcing an unfair pattern that already exists.

Finally, unfairness can be introduced at the last stage, where patterns are found in the data you've collected. Where to draw your lines reflects your values. How much do you weigh being wrong about taking away the freedom of a person who wouldn't commit a future crime vs. being wrong about letting someone free who would? How much do you weigh missing a qualified teacher vs. hiring an ineffective one? In Hiring Teachers and Pretrial Release, where you draw your lines will not only settle these matters,

it will also determine the proportion of black-to-white non-reoffenders denied pretrial release, black-to-white qualified teachers that do not get a call-back, and so on.

Before moving on, it is worth noting that these examples do not reveal problems that are unique to machine learning. Other methods of data-driven decision making, such as doing statistical calculations by hand, are susceptible to the exact same problems as those discussed in this section. I have focused on machine learning due to the entirely contingent fact that machines are an increasingly popular way to carry out data-driven decision making. That said, it is worth keeping in mind why machine learning has become so popular and how this connects to concerns about fairness. Machines can carry out certain processes—such as developing or updating classification systems—cheaply and at a scale that was recently practically impossible. Further—as we have just seen—the systems they develop can be unfair, and often in ways that can be hard to detect. This is owed to the exact same capacity that makes these systems popular: their ability to leverage hard to detect correlations in large datasets. This, among other things, has driven the search for scalable methods for detecting unfairness in machine learning systems, to which we will now turn.

3 Competing measures of fairness[11]

As the last section demonstrated, we can use classification to inadvertently develop classification rules that are unfair to use. This has lead technologists to ask how they might detect instances where a classification rule's use is unfair. In that discussion, two measures of fairness have emerged as front runners (Corbett-Davies and Goel (2018)). In this section, I'll introduce these measures and demonstrate that satisfying either is neither necessary nor sufficient for a classification rule's use being fair. I will also introduce and discuss counterfactual fairness and subject it to the same treatment. I will then use this discussion to motivate a general explanation of why the measures discussed here fail.

Before we begin, it will be helpful to note that the measures of fairness we will consider focus on aspects of the predictions on which decisions are based, and not the outcomes of the decisions they inform. This raises the question of what these measures have to do with fairness per se, given that

[11]This section was greatly influenced by Corbett-Davies and Goel (2018).

fairness is perhaps more naturally thought of as a feature of decisions or outcomes. I take it that features of predictions are relevant to fairness, given the context: the predictions serve (or at least are intended to serve) as the basis of life-affecting decisions.

It is worth adding that this isn't the full context, either. When we are considering the objects to which fairness measures apply, we are considering them under what we could call "normal operating conditions." These are the conditions under which life-affecting decisions (whether to subject individuals to treatment ϕ) are being based on some prediction (whether the individuals are y); it is at least prima facie reasonable to base decisions as to whether to give someone treatment ϕ on their status whether they are y; and the decision-making system under discussion was developed at least minimally competently (in that its predictions are reliable and incorporated into decisions in an at least prima facie sensible fashion), is functioning properly, is being used as intended, and was not designed maliciously or with any sort of discriminatory intent. I take it that if we can show that a satisfying measure is neither necessary nor sufficient for a system's being fair under these conditions, we have shown the measure to be neither necessary nor sufficient for fairness in the sense that matters for the purposes of this paper.[12]

3.1 Parity and calibration

One popular fairness measure states that

> *Classification parity*: predictive performance ought to be equal across groups, defined by protected attributes (Corbett-Davies and Goel (2018)).

Because there are different ways of measuring classification error—we could, for example, look at false positive rates or false negative rates—there are several fairness measures that pursue classification parity. For the purposes of this paper, I will focus on one of these measures,

[12]This leaves room for the possibility that the measures could capture some other sense of fairness, say some sense of fairness as it applies to predictions per se. My arguments cannot rule this out, but I do not think that it is a problem for my project, as it is concerned with the question of fair data-driven decision making. Whether there is some notion of fairness as it applies to predictions per se will not change the substance of my conclusion, as the arguments will have shown that being fair in this sense is neither necessary nor sufficient for fair data-driven decision making.

> *Predictive equality*: the rate of false positives of a given rule ought to be equal across groups, defined by protected attributes (Verma and Rubin (2018)).

Predictive equality asks us to balance false positives. If we understand "high risk" as the positive, predictive equality tells us that COMPAS is unfair because it is more likely to issue false positives for black defendants.

Another popular conception of fairness states that

> *Calibration*: outcomes (e.g., rearrest) ought to be probabilistically independent of membership to protected a protected group, given one's data-driven judgement (e.g., "high-risk") (Corbett-Davies and Goel (2018)).[13]

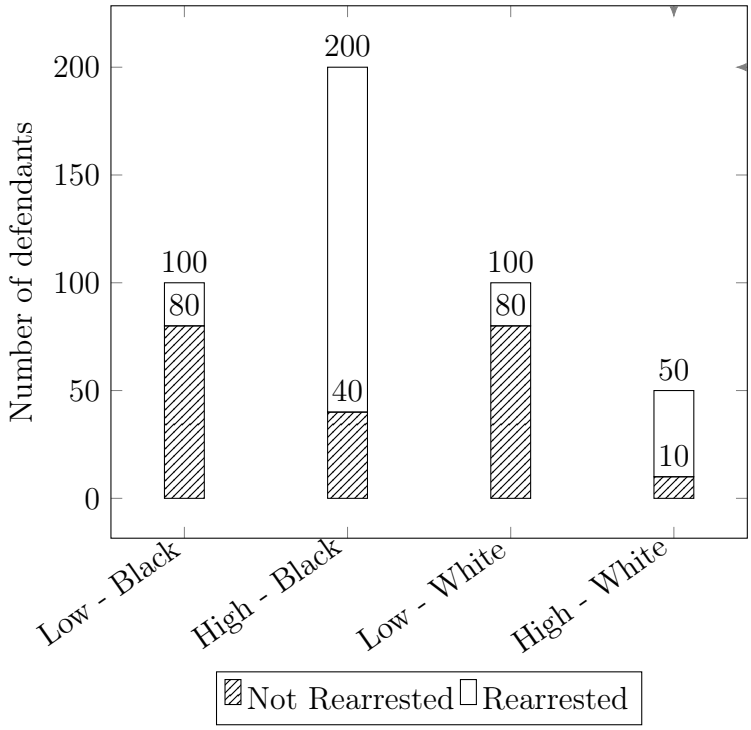
One way to understand calibration is that it requires a rule's judgements to mean the same thing for anyone subject to it. COMPAS meets this standard: the black and white defendants it identifies as "high risk" reoffend at similar rates (Northpointe Inc. 2016).[14]

To see why predictive equality (and other measures that pursue classification parity) and calibration are often at odds with each other, let's consider a

[13]There is a third popular conception, *anti-classification*, which forbids classification rules from taking as inputs data about whether someone belongs to a protected group. Anti-classification is much less plausible than the two we will discuss. In many instances, including Hiring Teachers and Pretrial Release, labels not directly tracking membership to a protected group can serve as a proxy for that membership. For this reason, rules that do not take protected status directly as inputs can discriminate against these groups using proxies. Further, knowing a subjects' protected status might improve accuracy in ways that are required to achieve fair outcomes. Consider, for instance, the fact that in some contexts women reoffend less often than men (Skeem et al. (2016), Binns (2017), DeMichele et al. (2018), Corbett-Davies and Goel (2018)). In these contexts, gender is predictive of reoffense. If we are building a rule to predict reoffense, blinding ourselves to gender will make us less accurate in ways that will systematically disproportionately adversely affect women, because it will lead to our overestimating their likelihood of reoffense (Corbett-Davies and Goel (2018)). As these observations demonstrate, refraining from taking as inputs data about whether someone belongs to a protected class is neither necessary nor sufficient for a classification rule's use being fair: unfair discrimination can occur without using taking group membership as an input, and using group membership might be important for achieving fairness.

[14]This is also true of the low and medium risk groups.

simplified example based on COMPAS.[15] Suppose a hypothetical pretrial risk assessment tool, ASTROLABE.[16] ASTROLABE identifies defendants as "Low" (low risk of committing a crime while out on bail) and "High" (high risk of committing a crime while out on bail). The following chart summarizes ASTROLABE's performance:



Note that the following is true of ASTROLABE:[17]

1. The proportion of low risk defendants who are rearrested—20%—is the same regardless of race, and the proportion of high risk defendants who are rearrested—80%—is the same regardless of race. Put another way: *within* any given groups, rearrest rates are in parity for white and black defendants.

[15] I owe much of this explanation to Corbett-Davies et al. (2016). The example is mine, but the general arch of the explanation is theirs. For a similar explanation of these issues, see Rubel et al. (2021).

[16] I am using a hypothetical tool so that I can stipulate numbers that make the salient aspects of the COMPAS case a bit easier to see.

[17] 1.-4. are also true of COMPAS (Corbett-Davies et al. (2016)).

2. The overall rearrest rate for black defendants—60%—and is not equal to the overall rearrest rate for white defendants—40%.

3. Black defendants are more likely than white defendants to be classified as high risk. (2/3 vs. 1/3.)

4. Black defendants who are not rearrested are twice as likely to be misidentified as high risk than their white counterparts. (2/15 vs. 1/15.)

Note that 1. and 2. guarantee 3. and 4. (cf. Corbett-Davies et al. (2016)). The stacked bar chart can help us see this. We can generalize a bit and say the fact that the system is calibrated and not perfectly accurate (which is entailed by 1.), when combined with the fact that the black rearrest rate is higher than the white rearrest rate (which is entailed by 2.), guarantees that predictive equality does not hold (which is entailed by 4). Therefore, classification parity is not achieved. We are now in a position to see that wide class of cases, calibration and classification parity are incompatible (cf. Kleinberg et al. (2016)).

Now that we have an understanding of predictive equality, calibration, and a sense of why we often can't satisfy both, I will explain why neither is a very good measure of fairness.

Let's begin with predictive equality. To see why abiding by predictive equality is not required for a machine to be fair, consider a case inspired by the fact that in some contexts, men reoffend more often than women:

> **Violent Offense.** You are deciding which defendants to give free anger management counseling to. There are a limited number of counselors, and your task is to increase public safety by giving counseling vouchers to defendants that are at high risk of committing violent offenses while out on bail. Male defendants are much more likely to commit violent crimes while out on release than female defendants, and your data reflect this. You construct an accurate—but not perfectly accurate—and calibrated system for identifying individuals that are at high risk of committing violent offenses while out on bail. You give "high risk" individuals vouchers.

Your system will involve a rule that behaves much like ASTROLABE in that your rule will violate predictive equality. This time, however, it is male defendants (as opposed to black defendants) who are more likely to be misidentified

as high risk, and those defendants will be offered optional anger management counseling as opposed to being denied bail.

Using the rule you construct in Violent Offense seems to be fair, despite its violating predictive equality. If we stipulate that it's highly reliable (maybe it's at least as good as ASTROLABE), it is hard to see how it could be considered unfair. Sure, some men who don't need counseling will be offered it, but they can always decline the invitation. And, of course, more men will be mistakenly offered counseling, but, again, there seems to be no problem here, especially if this is the best we can do with the data and resources we are given.Violent Offense, then, teaches us that abiding by predictive equality is not necessary for a rule's use being fair: the use of the rule in this case is intuitively fair, despite its violating predictive equality (for a similar analysis see Corbett-Davies and Goel (2018); see Hedden (2021) for a different case that shows that abiding by predictive equality is not necessary for a rule's use being fair).

A related case can teach us that achieving predictive equality is not sufficient for a rule's being fair. We can, as the following case illustrates, achieve predictive equality to make up for the unequal false positive rate by systematically overestimating reoffense among women:

> **Violent Offense Two.** You are deciding which defendants to deny bail to. Your task is to increase public safety by identifying defendants that are at high risk of committing violent offenses while out on bail. Male defendants are much more likely to commit violent crimes while out on release than female defendants, and your data reflect this. Valuing predictive equality, You construct a system that has lower standards for identifying women as high risk than it does for men so that the false positive rates among men and women will be in parity.

Despite its achieving predictive equality, this is a deeply unfair system: It involves having different standards form men and women, such that a burden is placed on certain women to make up for the bad behavior of men. As the Violent Offense cases demonstrate, satisfying predictive equality is neither necessary nor sufficient for a classification rule's use being fair.

Let us now turn to calibration. If we return to ASTROLABE, we can quickly see why abiding calibration is neither necessary nor sufficient for a rule's use being fair. Suppose that in ASTROLABE's world—as in the United

States (Bunting et al. (2013))—black people are disproportionately arrested for crimes committed at equal rates by both white and black people. Imagine this has a distorting effect on ASTROLABE, which leads it to overpredict reoffense among black defendants, relative to their white counterparts. This, it would seem, is the paragon of unfairness, despite ASTROLABE's satisfying calibration (as it is commonly measured, which is with rates of rearrest).

If we think through this example one step further, we can see that satisfying calibration isn't necessary for a rule's use being fair either. Suppose we build a second device, SEXTANT, that produces scores adjusted for policing bias. We can imagine that SEXTANT takes as inputs ASTROLABE's judgements as well as other data about disproportionate arrest rates and produces adjusted scores that more closely track reoffense (as opposed to rearrest). SEXTANT will violate calibration: "low risk" black defendants will be rearrested at higher rates than their white, "low risk" counterparts; yet, we'd be loathe to consider SEXTANT unfair in virtue of this.

3.2 Counterfactual fairness

We can round out our consideration of fairness measures by considering a newer measure that is markedly different from the ones we have considered so far, in that—as we will soon see—that it uses the notion of causation to understand unfairness:

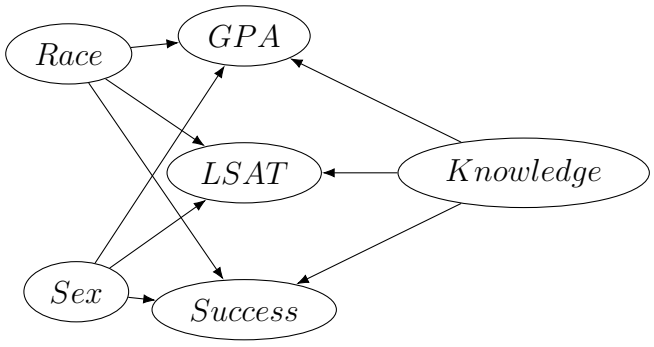
> *counterfactual fairness*: decisions ought to be the same in the actual world and any counterfactual world where the individual belongs to a different demographic group (Kusner et al. (2018)).

Counterfactual fairness may sound complicated, but it is in fact a fairly easy measure to understand. What it amounts to is a prohibition on the use of any variables for the use of prediction that are causally affected by protected attributes. In other words, a prediction-based decision will be fair by the lights of counterfactual fairness iff none of the variables used by the system are affected by protected attributes.

Before moving on, it will be helpful to clarify why this last claim—i.e., that a prediction-based decision will be fair by the lights of counterfactual fairness iff none of the variables used by the system are affected by race—is true.

Counterfactual fairness understands counterparthood by reference to a causal model that maps causal relations among the inputs and outputs of a

given prediction-based decision system, as well as any protected attributes that affect those the inputs and outputs. Take, for example, the following hypothetical causal model, used by Kusner et al. (2018), as the model corresponding to a system used in a case called "Law School Success," where a machine learning system is used to predict success in one's first year of law school (in the model, "Success") on the basis of one's GPA (in the model, "GPA") and LSAT score (in the model, "LSAT"). In the model, "Knowledge" stands for the student's law knowledge, "Race" for their race, and "Sex" for their sex. An arrow from one node—e.g., Race—to another—e.g., LSAT—means that race affects LSAT scores.[18]. We will suppose that the arrows represent (in this example, unspecified) equations, such that if we knew the values of all nodes going into a node, we could calculate the value for that node (e.g., if we know race, gender, and knowledge, we can compute GPA, LSAT, and Success).



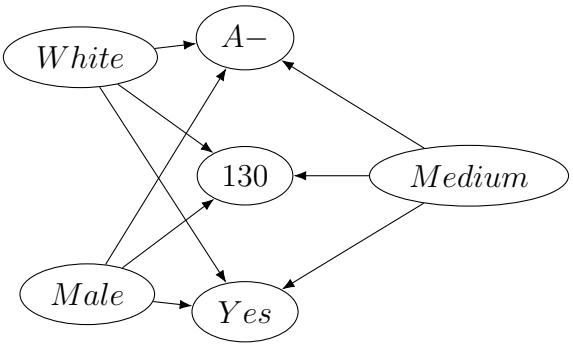
Causal model for Law School Success

Note that this model indicates that race and sex affect GPA, LSAT score, and Success. The idea here is that one's race and sex influence these variables in a variety of ways via structural racism and sexism. Note that law knowledge also affects these variables but that it is—however improbably—not affected by race or gender.

Now suppose that a given system predicts Success on the basis of GPA and LSAT, letting in only those who it predicts will be successful. Is such a system counterfactually fair? To determine this, let's consider a given

[18]For a similar discussion, see Castro et al. (n.d.).

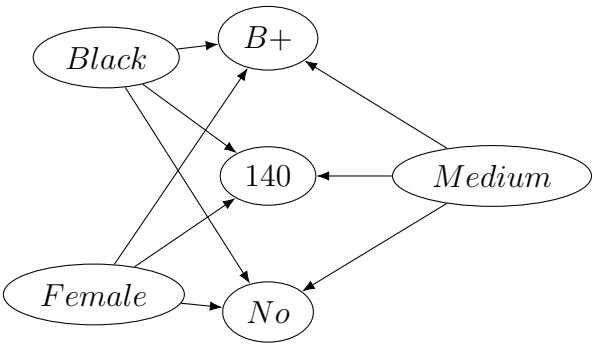
applicant, Hanlon. Here's the causal model with the values for Hanlon filled in:



Causal model for Hanlon

To determine whether this system is fair we need to either find a counterpart of Hanlon who has different protected attributes and does not get in, or show that all his counterparts with different protected attributes would get in. For the purposes of counterfactual fairness, one's counterpart is anyone who has the same assignments for all variables *not* affected by protected attributes. In this case, everyone with a medium level of law knowledge is a counterpart for Hanlon.

Let us now demonstrate that the system is counterfactually unfair. Let us suppose that if we fill "Black" in for race and "Female" in for gender, we get the following:

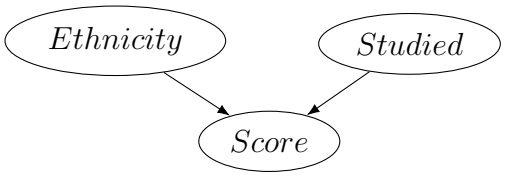


Causal model for Hanlon's counterpart

Given that Hanlon's counterpart gets a different answer—"No"—this system is not counterfactually fair. Note also that we would have a counterfactually fair system if we let applicants in only on the basis of law knowledge, because with such a system we would never have a situation where someone gets in but their counterpart belonging to a different demographic does not.

We can now ask our familiar questions: is being counterfactually fair necessary for a rule's use being fair? Is it sufficient? Characteristically, I will say that it is not in both cases.

Let's start by asking whether being counterfactually fair is necessary. Consider a case, call it Internship, where we are testing for fluency in a language.[19] Presumably, it is legitimate to require fluency in a language for some positions and that fluency could be determined via a test. Suppose, then, that we set up a system where taking such a test is part of applying for an internship. Let us stipulate that the causal model looks like this:



If we specify that Spanish is a legitimate qualification of the position, that all applicants had an opportunity to achieve Spanish fluency by formally studying Spanish, and that applicants pass the test *iff* they are fluent, then it is safe to assume that this system is fair. Yet, counterfactual fairness will judge that it is not. We can imagine a candidate who fails the test because they are non-Hispanic and did not study Spanish. Such a candidate will have a Hispanic counterpart that passes the test even though the counterpart did not formally study Spanish; thus, the system is not counterfactually fair. Yet, the system is fair; so, being counterfactually fair is not necessary for being fair.

Let us now turn to the question of whether being counterfactually fair is sufficient for a rule's use being fair. To see why it is not, consider the following highly stylized case:

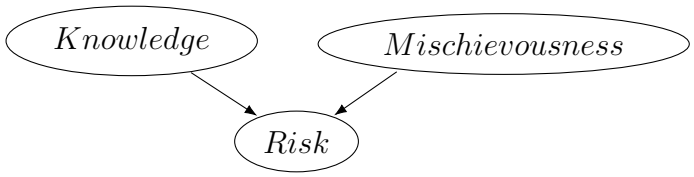
> **Sundial.** You are deciding who to release while awaiting trial. Valuing counterfactual fairness, you have tossed out COMPAS.

[19] This case has many features in common with a case discussed in Castro et al. (n.d.).

> Your new system, SUNDIAL bases its prediction of recitivism off of two features of defendants that are not casually influenced by any protected attributes: law knowledge and mischievousness. The system detects presence of both features, and it deems those who have both to be high risk. Due to policing practices—such as placing police in certain schools but not others—no white people with law knowledge ever face SUNDIAL, meaning no innocent white defendants face SUNDIAL; yet, many innocent black defendants do. As a result, SUNDIAL deems innocent black people high risk more often than their white counterparts.

It might help with what follows to suppose that when SUNDIAL was adopted you very well could have adopted RADAR, which ignores law knowledge and instead detects *kinds* of mischeviousness, one of which is evenly distributed among the white and black defendants that you see and thus does not disproportionately disadvantage black defendants in the way that SUNDIAL does.

In case it is helpful, here is a model for SUNDIAL:



As the causal model makes clear, SUNDIAL is counterfactually fair. Yet, its use is unfair. It serves to only identify black defendants as high risk—it is unfair much in the same way that COMPAS is. For these reasons, counterfactual fairness—like predictive equality and calibration—is neither necessary nor sufficient for a rule's use being fair.

3.3 Why the measures fare poorly

I would now like to attempt a general diagnosis of why the measurements of fairness we have considered fared so poorly.

The problem that predictive equality, calibration, and counterfactual fairness (and anti-classification, which is discussed in footnote 13) share in common is that they are merely measures of formal fairness. To see this, distinguish a classification rule (e.g., "if a suspected Frenchie's height and weight put in it quadrant D, predict Frenchie") from the rule—call it *the ultimate*

rule—that the classification rule attempts to satisfy (e.g., "Label Frenchies as Frenchies"). The measures of fairness that we have considered tell us whether using a classification rule—which attempts to satisfy some ultimate rule by giving instructions about what to do—results in the unequal or partial application of its ultimate rule. What they *do not* do is check to see if, say, a given application of the ultimate rule is applied with sensitivity to the needs of the worse off, imposes unjustifiable burdens arbitrarily, or reinforces and perpetuates an oppressive social practice.

One of the major problems with the focus on formal fairness is that many of our complaints about fairness in machine learning are broader than anything captured by formal fairness. Imagine a case where a community wants to save resources by developing a pretrial risk assessment tool that predicts who will be *rearrested*—as opposed to who will *reoffend*—if released. If black arrest rates in the community are much higher than white arrest rates due to uneven policing, using the pretrial risk assessment tool is unfair regardless of whether it is applying its ultimate rule equally and impartially. This is because applying the ultimate rule "label as high risk those who would be rearrested" is itself unfair. This helps us see why, as a general matter, showing that a pretrial risk assessment tool is or isn't calibrated (or does or doesn't satisfy anti-classification, predictive equality, or counterfactual fairness) does not guarantee that it is (un)fair: many of our concerns with fairness go beyond questions whether a rule—regardless of what it is asking for and regardless of the broader context—is being equally and impartially applied.

We can apply this lesson to COMPAS. It explains why Northpointe Inc.'s response (that COMPAS is fair because it satisfies calibration) to ProPublica's complaint (that COMPAS is unfair because it violates predictive equality) was hardly enough to show that ProPublica hadn't located a problem with the use of COMPAS. Northpointe Inc.'s response leaves open the question of whether COMPAS violates predictive equality because it picks up on and recreates arrest patterns due to uneven policing (as opposed to actual reoffense). If it has, the fact that it perpetuates this injustice while being formally fair is largely besides the point. Our real concern is about the inequality that drives COMPAS to violate predictive equality and whether COMPAS may be playing a role in perpetuating that inequality.

Our discussion also reveals that a moral criticism of COMPAS based on ProPublica's findings is incomplete until it explains why the violation of formal fairness ProPublica discovered is unfair in this context. This is because,

as we saw in 3.2, predictive equality is not required for the application of a rule to be fair. Now, I take it that COMPAS is unfair, and that this has to do with the fact that it places greater burdens on the worse off. This, as I will discuss in the next section, shows that even though predictive equality (as well as various other fairness measures) isn't required for being unfair, it can serve as a useful heuristic in detecting unfairness.[20]

4 Lessons and loose ends

Suppose what I have said so far is correct. What follows?

For one, when it comes to fairness in machine learning, we should worry less about articulating fairness measures that merely track formal fairness. Instead, we should start worrying about better understanding the demands of fairness that got us to worry about machine fairness in the first place. Throughout this paper, we have seen that one perennial concern with machine learning systems is how they treat members of disadvantaged groups. Our investigation has taught us both how automated systems can further disadvantage members of these groups (e.g., by systematically making errors in a way that further disadvantages them) and how satisfying extant fairness measures is not enough to ameliorate these worries.

We should also work towards better understanding how the fairness measures discussed in this paper might serve as reliable heuristics for diagnosing and addressing matters of fairness. In this paper we have caught a glimpse of how this might work. As the ASTROLABE example helps make clear, if a calibrated system violates predictive equality, this is evidence of an underlying inequality in the populations. If members of one population tend to be worse off than another and the system benefits the better off and burdens the worse off, then there will be serious cause for concern from the vantage point of fairness. The upshot here is that predictive equality and calibration can be used in conjunction with one another to detect potentially noxious inequalities that we risk compounding through the use of a data-driven judgment system.

Finally, a worry. Understanding fairness as it applies to data-driven judgments in the way that I have carved our here will be a large and messy undertaking. But, as I hope to have convincingly argued, there is no alternative:

[20] See Hellman (2020) for a defense of a similar claim.

we must be deeply suspicious of any one-size-fits all solutions—such as predictive equality, calibration, and counterfactual fairness—that promise to be both operationalizable and informative enough to algorithmically measure fairness. Instead, we must patiently use a plurality of tools from moral and political philosophy as well as statistics and computer science to answer these questions in piecemeal fashion. And this is as it should be: the question of whether we are permitted to pursue an ultimate rule, and if so, how, just is the question of how and whether to pursue an end (i.e., the central question of ethics).

References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. [Online; posted 23-May-2016].

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, 104(671).

Binns, R. (2017). Fairness in machine learning: Lessons from political philosophy. *arXiv preprint arXiv:1712.03586*.

Bunting, W., Garcia, L., and Edwards, E. (2013). The war on marijuana in black and white.

Castro, C., O'Brien, D., and Schwan, B. (n.d.). [A paper on fair machine learning that is currently under review - email Clinton (clinton.g.m.castro@gmail.com for a draft)].

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

Corbett-Davies, S. and Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *CoRR*, abs/1808.00023.

Corbett-Davies, S., Pierson, E., Feller, A., and Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. [Online; posted 23-May-2016].

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. *CoRR*, abs/1701.08230.

DeMichele, M., Baumgartner, P., Wenger, M., Barrick, K., Comfort, M., and Misra, S. (2018). The public safety assessment: A re-validation and assessment of predictive utility and differential prediction by race and gender in kentucky. *SSRN Electronic Journal*.

Dixon, P. and Gellman, R. (2014). The scoring of america.

Gerrish, S. and Scott, K. (2018). *How Smart Machines Think*. The MIT Press. MIT Press.

Goodman, P. (1969). Can technology be humane?

Green, J., Green, H., and Brungard, B. (2017). Machine learning & artificial intelligence: Crash course computer science 34. [Online; posted 1-November-2017].

Hedden, B. (2021). On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs*, 49(2):209–231.

Hellman, D. (2020). Measuring algorithmic fairness. *Va. L. Rev.*, 106:811.

Hooker, B. (2005). Fairness. *Ethical Theory and Moral Practice*, 8(4):329–352.

Huq, A. Z. (2019). Racial equity in algorithmic criminal justice. *Duke LJ*, 68:1043.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. (2018). Counterfactual fairness.

Northpointe Inc. (2016). *COMPAS Risk Scales: Accuracy Equity and Predictive Parity*. Northpointe.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press, Cambridge, MA, USA.

Quick, K. (2015). The unfair effects of impact on teachers with the toughest jobs.

Rubel, A., Castro, C., and Pham, A. (2021). *Algorithms and Autonomy: The Ethics of Automated Decision Systems*. Cambridge University Press.

Skeem, J. L., Monahan, J., and Lowenkamp, C. T. (2016). Gender, risk assessment, and sanctioning: The cost of treating women like men. *Law and human behavior*, 40 5:580–93.

Turow, J. (2017). *The Aisles Have Eyes: How Retailers Track Your Shopping, Strip Your Privacy, and Define Your Power*. Yale University Press.

Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, page 1–7, New York, NY, USA. Association for Computing Machinery.