2

Algorithmic Fairness

From Parity to Pareto

BIAS BY ANALOGY

If you're sufficiently beyond your high school days, you might remember—with love or loathing—the multiple-choice word analogy problems that were formerly a staple of standardized testing. Such problems ask one to consider relationships such as "*runner* is to *marathon*…" and realize that the correct response is "…as *oarsman* is to *regatta*" and not, for instance, "…as *martyr* is to *massacre*." These linguistic puzzles were removed from the SAT in 2005, in part over concerns that they were biased in favor of certain socioeconomic groups—for example, those who knew what regattas were. Perhaps because the loathers outnumbered the lovers, word analogies and their cultural biases seemed to have been forgotten for over a decade.

But they resurfaced in 2016 in a new context, when a team of computer science researchers subjected Google's publicly available "word

embedding" model to a clever test based on word analogies. The idea behind a word embedding is to take a colossal collection of text and compute statistics about the so-called co-occurrences of words. For instance, the words *quarterback* and *football* are more likely to be found in close proximity to each other in a document, paragraph, or sentence than the words *quarterback* and *quantum.* These pairwise co-occurrence statistics are then fed to an algorithm that attempts to position (or embed) each word in 2-, 3- or higher-dimensional space in such a way that the distances between pairs of words approximately reflect their co-occurrence statistics. In this sense, words that are more "similar" (as reflected entirely by their empirical usage in the documents) are placed "nearer" each other.

Building a large-scale word embedding is an exercise in data collection, statistics, and algorithm design—that is, it is a machine learning project. Google's "word2vec" (shorthand for "word to vector") embedding is one of the best-known open-source models of its kind and has myriad uses in the many language-centric services Google provides—for instance, realizing that *bike* may be a synonym for *bicycle* when in proximity to *mountain*, but for *motorcycle* when in proximity to *Harley-Davidson.*

The starting point behind the 2016 work was an observation dating back to the 1970s: if we have a good embedding of words into, say, 2-dimensional space, then word analogies should roughly correspond to parallelograms. Thus if *man* really is to *king* as *woman* is to *queen*, then the four points corresponding to these words in the embedding should define two sets of parallel lines—namely, the *man-woman* and *king-queen* pair, and the *king-man* and *queen-woman* pair.

We can use this observation to "solve" for the missing word in analogy problems. For if we ask questions of the form "*Runner* is to *marathon* as *oarsman* is to what?," the three specified words—in this case, *runner*, *marathon*, and *oarsman*—define three corners of a

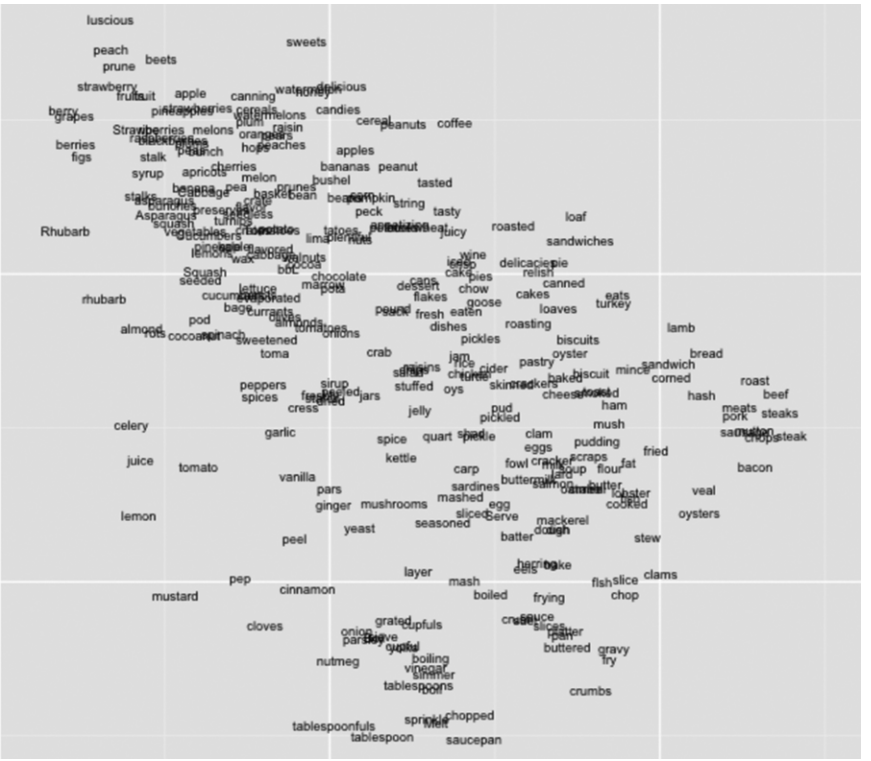


Fig. 6. A small word embedding in two dimensions.

parallelogram in the word embedding, which in turn determines where in space the fourth corner should lie. By looking at the word lying closest to this missing corner, we find the "what"—the word that the embedding "thinks" is the best completion of the analogy. (See Figure 7.)

The team of researchers could have stopped there and used this observation to see how well word2vec would have fared on 1990s SAT tests. But just as with people, sometimes the failings of algorithms and models are far more revealing and interesting than their strengths. So the researchers deliberately restricted the word analogies they investigated to those of the form "*Man* is to *X* as *woman* is to *Y*," where *X* was specified (giving the required three corners) but *Y* (the missing corner) was not—as in "*Man* is to *computer programmer* as

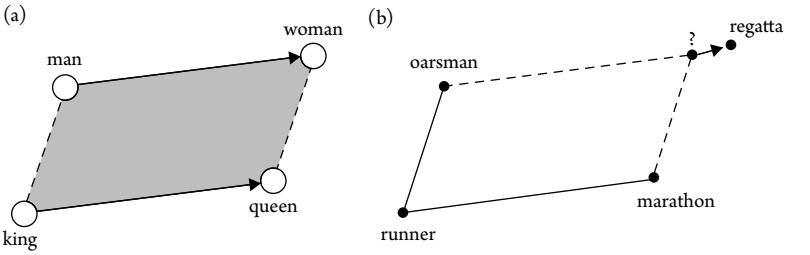


Fig. 7. Word analogies as parallelograms.

woman is to what?" In other words, they specifically tested word2vec on gender analogies.

The findings were dramatic, and demonstrated that word2vec was guilty of rampant gender bias and stereotyping. The example given above was answered by the title of the research article: "Man Is to Computer Programmer as Woman Is to Homemaker?" The paper documented the systematic manner in which word2vec reflected, and perhaps amplified, the biases already present in the raw documents it was trained upon. In the small sample of the embedding reproduced below in Figure 8, one can even visually complete analogies of this form. For instance, it would seem that *ladies* is to *earrings* as *nephew* is to *genius*. While this makes little sense linguistically, it nevertheless feels sexist, with women being associated with ornamentation and men with brilliance. While word analogies may seem a bit esoteric, the important point is that word embeddings are used as basic building blocks for more complicated learning algorithms. And more serious problems can arise when word embeddings and other biased models are used as components in more consequential applications.

In fact, in late 2018, something along these lines was discovered in the machine learning model that Amazon was building to evaluate the resumes of candidates for software engineering jobs. Its algorithm was found to be explicitly penalizing resumes that contained the word *women's*, as in "women's chess club captain," and downgraded candidates who listed the names of two particular all-women colleges. Amazon ultimately disbanded the team working on this particular

project, but not before some damage was done—at least in the form of bad publicity for Amazon, and potentially in the form of implicitly enabling discriminatory hiring.

Neither the authors of the word embedding article nor any of its discussants suggested that the bias of word2vec was the result of sexist programmers at Google, unrepresentative or corrupted data, or coding errors. And there is also no reason to suspect that bias in Amazon's hiring tool was the result of malice either. Any of those explanations might have been more reassuring than the truth, which is that the bias was the natural if unexpected outcome of professional scientists and engineers carefully applying rigorous and principled machine learning methodology to massive and complex datasets.

The problem here is that the training data used in machine learning applications can often contain all kinds of hidden (and not-so-hidden) biases, and the act of building complex models from such data can both amplify these biases and introduce new ones. As we discussed in the introduction, machine learning won't give you things like gender neutrality "for free" that you didn't explicitly ask for. Thus even though probably very few of the documents used to create the word embedding,

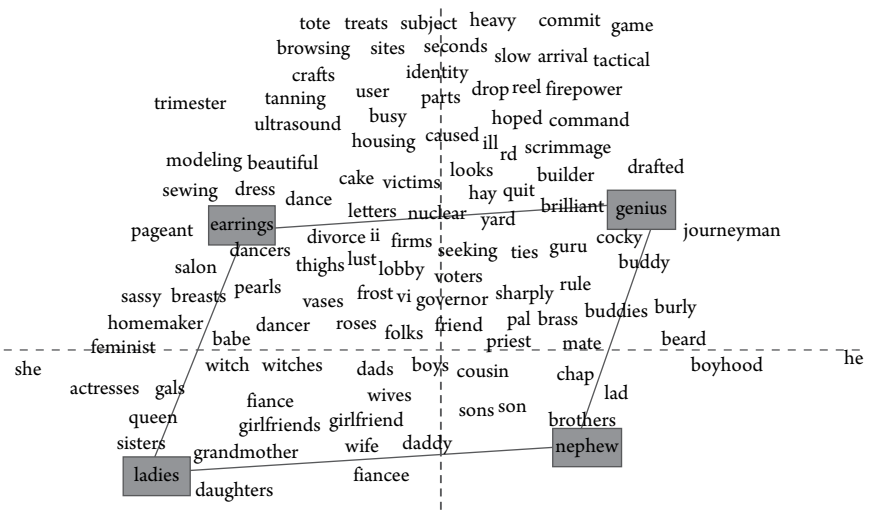


Fig. 8. A word embedding exhibiting gender bias.

if any, exhibited blatant sexism (and certainly none of them actually suggested that homemaker was the best female analogue for male computer programmer), the tiny collective forces of language usage throughout the entire dataset, when compressed into a predictive model for word analogies, resulted in clear gender bias. And when such models then become the basis for widely deployed services such as search engines, targeted advertising, and hiring tools, the bias can be further propagated and even amplified by their reach and scale. This is an example of the complex feedback loops fueled by machine learning that we will examine in Chapter 4.

The embedding bias paper deservedly received a great deal of both academic and media attention. But this is only one example of the growing number of cases in which algorithms—and, most notably, algorithms based on models derived from data via machine learning—exhibit demonstrable bias and discrimination based on gender, race, age, and perhaps many other as yet unknown factors and combinations of factors. And while bias in the search results or ads we are shown may seem (perhaps incorrectly) like a relatively low-stakes affair, the same problems arise in much more obviously consequential domains, such as criminal sentencing, consumer lending, college admissions, and hiring.

As we said in the introduction, many of the problems we identify in this chapter have been discussed well at a high level in other places. Our particular interest will focus on the aspects of these problems that are specific to machine learning, and especially to potential solutions that are themselves algorithmic and on firm scientific footing. Indeed, while the main title of the word embedding paper expressed alarm over the problem uncovered, it also had a more optimistic subtitle: "Debiasing Word Embeddings." The paper reveals a serious concern, but it also suggests a principled algorithm for building models that can avoid or reduce those concerns. This algorithm again uses machine learning, but this time to distinguish between words and

phrases that are inherently gendered (such as *king* and *queen*) and those that are not (such as *computer programmer*). By making this distinction, the algorithm is able to "subtract off" the bias in the data associated with nongendered words, thus reducing analogy completions like the one in the paper's title, while still preserving "correct" analogies like "*Man* is to *king* as *woman* is to *queen*."

These are the themes of this chapter: scientific notions of algorithmic (and human) bias and discrimination, how to detect and measure them, how to design fairer algorithmic solution—and what the costs of fairness might be to predictive accuracy and other important objectives, just as we examined the costs to accuracy of differential privacy. We will eventually show how such costs can be made quantitative in the form of what are known as *Pareto curves* specifying the theoretical and empirical trade-offs between fairness and accuracy.

But ultimately, science can only take us so far, and human judgments and norms will always play the essential role of choosing where on such curves we want society to be, and what notion of fairness we want to enforce in the first place. Good algorithm design can specify a menu of solutions, but people still have to pick one of them.

LEARNING ALL ABOUT YOU

In machine learning, word embeddings are an example of what is known as unsupervised learning, where our aim is not to make decisions or predictions about some particular, prespecified outcome but simply to find and visualize structure in large datasets (in this case, structure about word similarity in documents). A far more common type of machine learning is the supervised variety, where we wish to use data to make specific predictions that can later be verified or refuted by observing the truth—for example, using past meteorological data to predict whether it will rain tomorrow. The "supervision" that guides our learning is the feedback we get tomorrow, when either

it rains or it doesn't. And for much of the history of machine learning and statistical modeling, many applications, like this example, were focused on making predictions about nature or other large systems: predicting tomorrow's weather, predicting whether the stock market will go up or down (and by how much), predicting congestion on roadways during rush hour, and the like. Even when humans were part of the system being modeled, the emphasis was on predicting aggregate, collective behaviors.

The explosive growth of the consumer Internet beginning in the 1990s—and the colossal datasets it generated—also enabled a profound expansion of the ways machine learning could be applied. As users began to leave longer and longer digital trails— via their Google searches, Amazon purchases, Facebook friends and "likes," GPS coordinates, and countless other sources—massive datasets could now be compiled not just for large systems but also for specific people. Machine learning could now move from predictions about the collective to predictions about the individual.

And once predictions could be personalized, so could discrimination. While an abstract model of collective language use such as word2vec can be demonstrably sexist, it's hard to claim any particular woman has been hurt by it more than any other. When machine learning gets personal, however, mistakes of prediction can cause real harms to specific individuals. The arenas where machine learning is widely used to make decisions about particular people range from the seemingly mundane (what ads you are shown on Google, what shows you might enjoy on Netflix) to the highly consequential (whether your mortgage application is approved, whether you get into the colleges you applied to, what criminal sentence you receive). And as we shall see, anywhere machine learning is applied, the potential for discrimination and bias is very real—not in spite of the underlying scientific methodology but often because of it. Addressing this concern will require modifying the science and algorithms, which will come with its own costs.

YOU ARE YOUR VECTOR

To delve deeper into algorithmic notions of fairness, let's consider in more detail the standard framework for supervised learning, but in settings in which the data points correspond to information about specific, individual human beings. What this information contains will depend on the decisions or predictions we want our learned model to make, but it is typical to view that information as being summarized in a list $\boldsymbol{x}$ (technically, a vector) of properties (sometimes called attributes or features) deemed relevant to the task. For instance, if we are trying to predict whether applicants to a college will succeed if admitted (for some concrete, verifiable definition of success, like graduating within five years with at least a 3.0 GPA), the vector $\boldsymbol{x}$ for applicant Kate might include things such as her high school GPA, her SAT or ACT scores, how many extracurricular activities she lists on her application, a score given by an admissions officer to her application essay, and so on. If instead we are trying to decide whether to give Kate a loan for college, we might include all of the information above (since her success in college may impact her ability or willingness to repay the loan) as well as information about her parents' income, credit, and employment history. In either case, the goal of supervised learning is to build a model that makes a prediction y (like whether Kate will succeed in college) on the basis of the vector $\boldsymbol{x}$ (here summarizing Kate's relevant information) using "historical" data of the same $<\boldsymbol{x},y>$ form—in this case, past applicants to the college and whether they succeeded or not. (We'll return later to the important fact that the college really only learns the outcome y for those students it admitted in the past and not for those it rejected, so the college's own decisions are influencing and potentially biasing the data it collects.)

Note that Kate's parents' financial status isn't really information about Kate but still seems relevant to the loan prediction task, especially if Kate's parents will be cosigners on the loan. Looking ahead a

bit, many debates about fairness in machine learning revolve in some way around what information "should" be allowed to be used in a given prediction task. Perhaps the most long-standing and contentious debate is over the inclusion of attributes such as race, gender, and age—properties that people cannot easily change, and that may seem irrelevant to the prediction task at hand, so that it may feel unfair to let them be used to make important decisions.

But what if it turns out that using race information about applicants really does result, on average, in more accurate predictions of collegiate success or loan repayment? Furthermore, what if those more accurate predictions result in some particular racial minority being discriminated against, in the sense that, all else being equal, members of that racial minority are less likely to be admitted to the college than others? Conversely, what if using race allows us to build accurate models that protect a group we wish to protect? Should we allow the use of race in these cases?

These questions do not have easy answers, and human judgments and norms will always need to play a central role in the debate. But the questions can be cast and studied scientifically—and even algorithmically—in a way that is certain to be central to the debate as well.

FORBIDDEN INPUTS

The question of what types of information should be allowed in making various decisions about people has been around for a long time, and is even the basis for significant bodies of law (for instance, in lending decisions and credit scoring, the direct use of race is generally illegal). But it has acquired greater urgency in the Internet era, where so much more data is gathered about people and is available for algorithmic decision-making—whether we know it or not. It is therefore tempting to assert that we can solve fairness problems simply by refusing to

allow models to have access to things such as racial or gender data if we deem these irrelevant to the task at hand. But as we discuss elsewhere, it is difficult to confidently assert that any information about you is "irrelevant" in making almost any decision, because of the very strong correlations between attributes—so removing these features often really will diminish accuracy. Worse, removing explicit reference to features such as race and gender won't be enough to guarantee that the resulting model doesn't exhibit some form of racial or gender bias; in fact, as we will see, sometimes removing racial features can even exacerbate the racial bias of the final learned model.

If someone knows what kind of car you drive, what kind of computer and phone you own, and a few of your favorite apps and websites, they might already to be able to make a pretty accurate prediction of your gender, your race, your income, your political party, and many other more subtle things about you. More simply, in many areas of the United States your zip code is unfortunately already a pretty good indicator of your race. So if there is some property *P* of people that is, on average, relevant or informative in predicting whether they will repay loans, and apparently irrelevant properties *Q*, *R,* and *S* can be combined to accurately predict property *P*, then in fact *Q*, *R,* and *S* are not irrelevant at all. Moreover, removing property *P* from the data won't remove the algorithm's ability to make decisions based on *P*, because it can learn to deduce *P* from *Q*, *R,* and *S*. And given that these combinations may involve many more than just a few properties, and may be so complicated as to be beyond human understanding (yet not beyond algorithmic discovery), defining fairness by forbidding the use of certain information has become an infeasible approach in the machine learning era. No matter what things we prefer (or demand) that algorithms ignore in making their decisions, there will always be ways of skirting those preferences by finding and using proxies for the forbidden information.

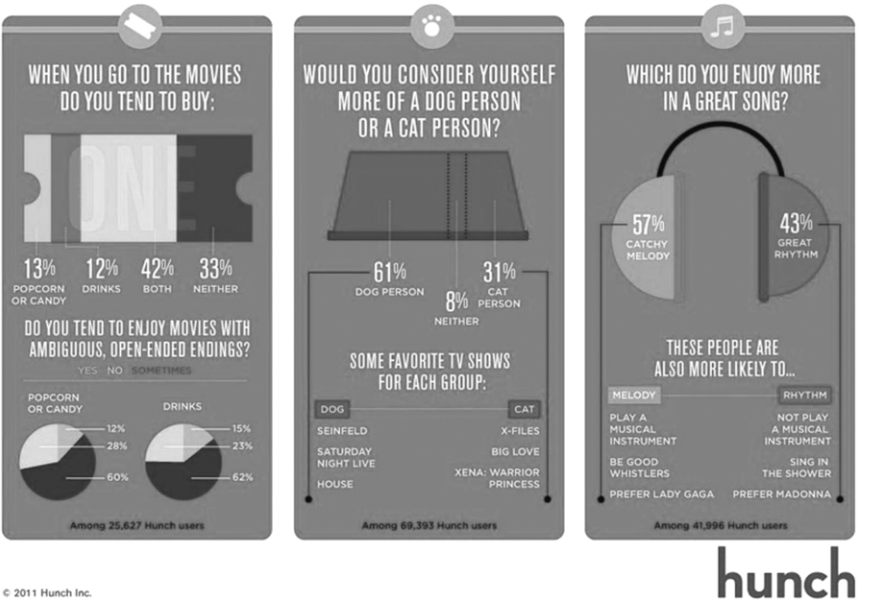


Fig. 9. Illustration of correlations between seemingly unrelated human attributes, such as between preferences for dogs or cats and favorite television shows. From data collected by the former tech startup Hunch (later acquired by eBay).

In other words, it has become virtually impossible to enforce notions of fairness that work by trying to restrict the inputs given to a machine learning or algorithmic decision-making process—there are just too many ways of deliberately or inadvertently gaming such efforts due to the number and complexity of possible inputs. An alternative approach is to instead define fairness relative to the actual decisions or predictions a model makes—in other words, to define the fairness of the model's outputs y rather than its inputs $\boldsymbol{x}$.

While we will see that this approach has been more successful, it also is not without its own drawbacks and complexities. In particular, it turns out that there is more than one reasonable way of defining fairness for the predictions made, and these different ways can be in conflict with each other—so there's no way to have it all. And even if we settle on just one of them, the predictions made by a model obeying a fairness constraint will, as a general rule, always be less accurate

than the predictions made by a model that doesn't have to; the only question is how much less accurate.

In other words, these kinds of more qualitative decisions and judgments—which type of fairness notion to use, when the reduction in accuracy is worth the gain in fairness, and many others—must remain firmly in the domain of human decision-making. The science part can begin only once society makes these difficult choices. Science can shed light on the pros and cons of different definitions, as we'll see, but it can't decide on right and wrong.

DEFINING FAIRNESS

The simplest notion of fairness applied to the predictions or decisions of a model is known as statistical parity. Like many definitions of fairness, defining statistical parity requires that we first identify what group of individuals we wish to protect. For concreteness, let's imagine a planet just like Earth except that there are only two races of people, Circles and Squares. Suppose for some reason we are concerned about discrimination against Squares in the granting of loans by a lender, so we ask that race be a protected attribute. Statistical parity simply asks that the fraction of Square applicants that are granted loans be approximately the same as the fraction of Circle applicants that are granted loans. That's all. The definition doesn't specify how many loans we have to give, or which particular Circle and Square citizens should receive them—it's just a crude constraint saying that the rate of granted loans has to be roughly the same for both races. Note that while our concern might have been discrimination against Squares, the definition is two-sided and thus also demands we not discriminate against Circles (though we could define a one-sided variant if we wanted).

Statistical parity is certainly some form of fairness, but generally a weak and flawed one. First, let's recall the framework of supervised

learning, where loan applicants like Kate have specific individual properties summarized in their vector $\boldsymbol{x}$, and there is some "true" outcome y indicating whether they will repay a loan. Statistical parity makes no mention of $\boldsymbol{x}$ at all—we could satisfy statistical parity by ignoring $\boldsymbol{x}$ entirely and picking an entirely random 25 percent of Circles and Squares to give loans to! This seems like a very poor algorithm for lending decisions, because we are entirely blind to the properties of individuals.

However, if we think about it a bit more carefully, this objection isn't that serious once we realize that statistical parity doesn't specify the goal of the predictions our model makes but is simply a constraint on those predictions. So the fact that there is a bad algorithm—random lending—that perfectly satisfies statistical parity does not mean there are not also good algorithms, ones that give loans to the "right" Circles and Squares. The goal of the algorithm might still be to minimize its prediction error, or maximize its profits; it's just that now it has to work toward this goal while being constrained to give out loans at equal rates. And in some sense it's reassuring that random lending obeys statistical parity, because it makes it immediately clear that the definition can be achieved somehow—which is not true of all fairness definitions.

Moreover, sometimes random lending might actually be a *good* idea, since it lets us obey statistical parity while we gather data. If we are a new lender and know nothing about the relationship between applicant attributes and loan repayment (i.e., between the $\boldsymbol{x}$'s and the y's), we can give out random loans for a while until we have enough $<\boldsymbol{x},y>$ pairs to make more informed decisions, while still being fair (according to statistical parity) in the meantime. In machine learning, this would be called *exploration*—a period in which we are focused not on making optimal decisions but on collecting data. There might also be settings where the deliberate blindness of random decisions is desirable for its own sake—for instance, in distributing a limited number

of free tickets to a public concert, where we don't view some candidates as more deserving or qualified than others.

A second and more serious objection is that statistical parity also makes no mention of the *y*'s, which here represent the ultimate creditworthiness of each applicant. In particular, suppose it is the case that for some reason (and there might be many) Squares in the aggregate really are worse lending risks overall than Circles—for example, suppose that 30 percent of Circle applicants will repay loans, but only 15 percent of Square applicants will. (See Figure 10.) If we managed to find a perfect predictive model—that is, one that takes the $\boldsymbol{x}$ of any applicant, Circle or Square, and always correctly predicts whether that individual will repay the loan or not—then statistical parity forces us to make some difficult choices, because the actual repayment rates of the two races are different, but fairness requires us to grant loans at the same rates.

For instance, we could obey statistical parity by granting loans to exactly the 15 percent of Square applicants who will repay and to half of the 30 percent of Circle applicants who will repay. But this might also feel unfair—especially to the other 15 percent of creditworthy Circle applicants to whom we unjustly deny loans. And if we make money by giving loans to people who will repay them and lose money by giving loans to those who won't, then the lender is also making less money than it could. On the other hand, we could also obey statistical parity by giving loans to all 30 percent of repaying Circle applicants, but then we'd have to also give loans not just to the 15 percent of repaying Square applicants but also to another 15 percent of defaulting Square applicants in order to equalize the loan rates for the two populations. And now we'd lose money on these.

In other words (and again in the language of machine learning), while statistical parity is not at odds with exploration, it is at odds with *exploitation*—that is, with making optimal decisions—any time the optimal thing to do from an accuracy perspective differs between the two populations. In such cases, we can't simply optimize our

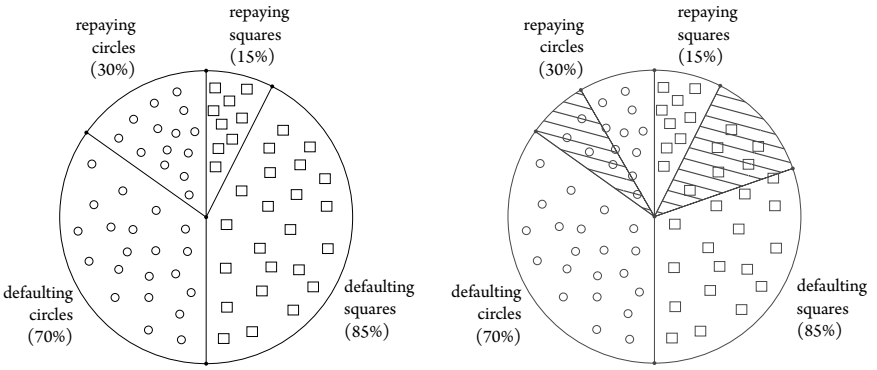


Fig. 10. Illustration of the tension between statistical parity and optimal decision-making. To obey statistical parity the lender must either deny loans to some repaying Circle applicants (shaded) or give loans to some defaulting Square applicants (shaded).

accuracy; we can only try to maximize it subject to the constraint of statistical parity. This is what the two solutions above do, in different ways—one by denying loans to creditworthy Circle applicants and the other by granting loans to Square applicants we know (or at least predict) will default. And while we shall next see that there are various kinds of improvements we can make to this coarse fairness constraint, the tension between fairness and accuracy will always remain, but can be made quantitative. In the era of data and machine learning, society will have to accept, and make decisions about, trade-offs between how fair models are and how accurate they are.

In fact, such trade-offs have always been implicitly present in human decision-making; the data-centric, algorithmic era has just brought them to the fore and encouraged us to reason about them more precisely.

ACCOUNTING FOR "MERIT"

The problem with statistical parity—that it can be violated even when making "perfect" decisions, if the Circles and Squares differ in their

creditworthiness—can be remedied by requiring that we evenly distribute the *mistakes* we make, rather than evenly distributing the loans we give.

In particular, we could ask that the rate of false rejections—decisions by the model to deny a loan to an applicant who would have repaid—be roughly the same for the Circle and Square populations. Why should this be considered a notion of fairness? If we view the creditworthy individuals who are denied loans as being the ones harmed, this constraint requires that a random creditworthy Circle applicant and a random creditworthy Square applicant have the same probability of being harmed. In other words, all else being equal, your race doesn't affect the probability with which our algorithm will harm you. And if we could somehow never make any mistakes at all—achieving perfect accuracy—that would be deemed fair by this notion, since then the rates of false rejections would be zero for both populations, and thus equal.

But now it's also "fair" if our model mistakenly rejects, say, 20 percent of the population of Square applicants who would repay their loans . . . as long as it also mistakenly rejects 20 percent of the population of Circle applicants who would repay their loans. So here we are evenly distributing not the loans themselves (as in statistical parity), but rather the mistakes we make in the form of false rejections. This opens the door to building predictive models that are imperfect (an inevitability in machine learning, as discussed shortly) while still being fair according to this new definition, which is naturally called equality of false negatives. (We can just as easily define equality of false positives for settings in which such mistakes represent the greater harm.)

Of course, if you are one of the creditworthy Square applicants who was rejected for a loan, this might still seem unfair, and it might not comfort you to know that your own unjust treatment is being

balanced by similar injustices to creditworthy Circles.[1] That's because both statistical parity and equality of false negatives are providing protections for groups (in this case the two races), but not for specific individuals in those groups, a topic we shall return to a bit later.

Since perfect decision-making is now deemed fair, we might be tempted to think that equality of false negatives eliminates the tension between fairness and accuracy. Unfortunately, while it does so in theory—namely, if from applications $\boldsymbol{x}$ we really could perfectly predict repayments y—in practice it does not, due to the cold realities of machine learning.

Those realities include the fact that the real world is messy and complicated, and even ignoring fairness entirely, it's rare to find a machine learning problem where there is sufficient data and computational power to find a model that makes perfect predictions—if such a model is even possible in principle, given that we can't hope to measure absolutely every salient property of loan applicants. And once we admit that our models will inevitably be imperfect, it's easy to find both cartoon and real examples of the tensions between accuracy and equality of false negatives.

FAIRNESS FIGHTING ACCURACY

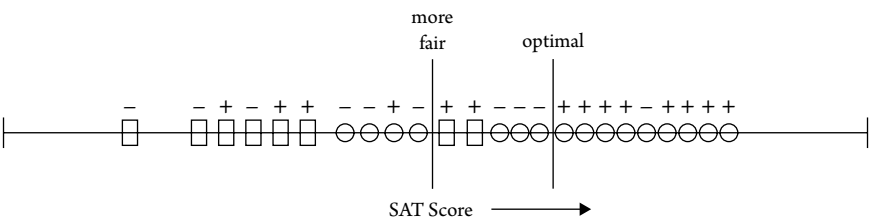
As an example, let's consider a simplistic problem in which we have to decide whom to admit to the fictional St. Fairness College based solely on SAT scores. Again, members of the majority Circle and Square populations are applying, and it turns out that the majority of applicants are Circles. Furthermore, Circle applicants tend to be wealthier and thus can afford SAT preparation courses and pay for multiple retakes

[1] Sydney Morgenbesser, who was a law professor at Columbia University, was reputed to have said about the aftermath of campus protests in 1968, that the police assaulted him unjustly, but not unfairly. Asked to explain, he said that "They beat me up unjustly, but since they did the same thing to everyone else, it was not unfair."

of the exam. The Square applicants are less wealthy and generally take the exam once, with less practice and preparation. Not surprisingly, the SAT scores of the Circles are on average higher than for the Squares, but for superficial reasons. It turns out both populations are equally well prepared for collegiate success. In particular, suppose that the percentage of Circles who would succeed if admitted to St. Fairness is equal to the percentage of Squares who would succeed; it's just that the Circles have inflated SAT scores.

If our model is a simple threshold rule—we admit any applicant whose SAT score is above some cutoff—then we simply may not be able to make perfect predictions, and furthermore, even choosing the most accurate model may badly violate fairness. The fundamental problem is that since the Circles are the majority class, the most accurate model—that is, the one that minimizes the number of mistakes it makes in the aggregate—will set the threshold largely based on the SAT scores and collegiate success of the Circles, since by definition the rate of mistakes on the majority group counts more toward aggregate error than the rate of mistakes on the minority group. This comes at the expense of discrimination (a higher false rejection rate) against the Squares.

To illustrate this, suppose our historical applicant dataset looks like the following:



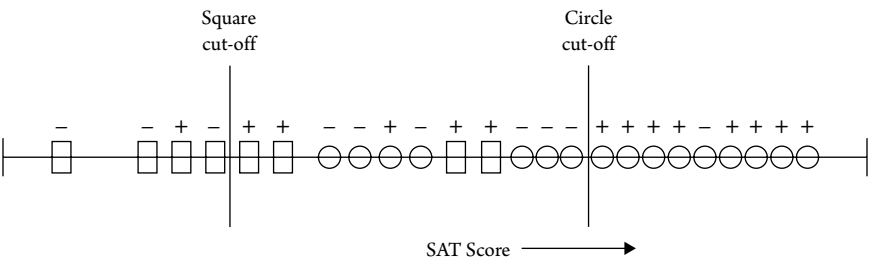
In this figure, circles represent Circle applicants and squares represent Square applicants. The position along the line indicates SAT scores, with higher scores to the right. A "+" symbol above an applicant indicates this person succeeded at St. Fairness, while a "–" indicates he or she did not.

On this data, the best model from a pure accuracy perspective is the cutoff labeled as "optimal." If we admit only applicants with this SAT score or higher, we make exactly seven mistakes: the one Circle – above the cutoff, the one Circle + below the cutoff, and the five Square +'s below the cutoff. But this means we would have falsely rejected these five successful Square applicants, whereas we would have falsely rejected only one Circle applicant. This violates the equality of false negatives notion of fairness. And if we use this cutoff to make decisions about future applicants, we should generally expect the disparity to be at least as bad as it was on the historical training data.

Of course, other models are possible. Moving the cutoff lower—for instance, to the line labeled "more fair"—improves fairness, according to equality of false negatives metric, by accepting two additional successful Squares, but it worsens accuracy, as we now make a total of eight mistakes instead of seven. The reader can confirm that even on this simple dataset, improving fairness will degrade accuracy, and vice versa.

Let's examine a couple of objections to this example. The first is that it is indeed a cartoon—no college would base admissions solely on SAT scores, but would instead build a more complex model incorporating many other factors. But in science generally, and in algorithm design specifically, if bad things can already happen in simple examples, they will also tend to happen in more complex ones—perhaps to an even greater extent. And the recent empirical machine learning literature is rife with examples of real-world problems in which building the optimal model for predictive accuracy results in demonstrable unfairness to some subpopulation. So increased complexity is not coming to our rescue on this issue.

A second objection is that the problem arises from our model not accounting for the fact that Circle and Square SAT scores differ for superficial reasons unrelated to collegiate success. If we know or detect statistically that the distributions of Circle and Square scores are different, why not build separate models for the two populations? For instance, in Figure 12 we show the same data with separate



cutoffs for Circles and Squares. This hybrid model would make only three mistakes (two for the Circle cutoff and 1 for the Square), better than the seven mistakes for the aggregate single cutoff previously discussed, and it is also more fair, since it falsely rejects the same number (one) of Circle and Square applicants. So we've permitted a more complex model but have improved on both criteria.

This might indeed be a good thing to do, increasing both fairness and accuracy at the same time. In fact, it's not so different from how certain affirmative action policies are implemented. But note that this model (which really involves first picking which submodel to use based on the applicant's race) now explicitly uses race as an input, which, as we discussed above, is something some notions of fairness (and many laws) forbid, since race can equally well be used to increase rather than reduce discrimination. If we removed race as an input, it wouldn't be possible to implement this hybrid model. Even if we can use a model like this, it doesn't necessarily solve the problem—what if SAT scores are more predictive of college success for one population than another? For example, suppose the optimal predictive model for the Circle population alone has a false rejection rate of 17 percent and the optimal predictive model for the Square population alone has a false rejection rate of 26 percent. We're still discriminating against the Square population according to equality of false negatives, even if it might be less than with a single, common, race-blind model.

Note that the gender bias observed in the word embeddings model we discussed at the start of the chapter can be blamed on latent human

bias that was present in the data. The algorithm simply picked up on the ways in which human beings used language—on reflection, how could we have expected it to do otherwise? But in the lending and admissions prediction problems we have just discussed, we can't as easily blame human bias in the data. We have assumed that the labels in our data are correct—anyone labeled in our dataset as able to succeed in college really is, and vice versa (of course, things only get worse if the labels might be wrong as well). The disparity in false rejections that emerges from our final admissions algorithm is the natural result of an algorithm trying to optimize for predictive accuracy—an emergent phenomenon that can't be blamed only on the class of models, the objective function, or any part of the data. It's just that when maximizing accuracy across multiple different populations, an algorithm will naturally optimize better for the majority population, at the expense of the minority population—since by definition there are more people from the majority group, and hence they contribute more to the overall accuracy of the model.

So there is simply no escaping that predictive accuracy and notions of fairness (and privacy, transparency, and many other social objectives) are simply different criteria, and that optimizing for one of them may force us to do worse than we could have on the other. This is a fact of life in machine learning. The only sensible response to this fact—from a scientific, regulatory, legal, or moral perspective—is to acknowledge it and to try to directly measure and manage the trade-offs between accuracy and fairness.

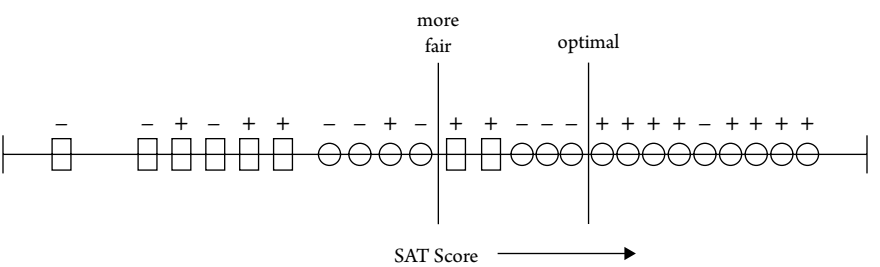
NO SUCH THING AS A FAIR LUNCH

How might we go about exploring this trade-off in a quantitative and systematic fashion—in other words, algorithmically? From the introduction—where we described the process of gradually adjusting a line or curve separating positive points from negative points, as well as the fancier but similar backpropagation algorithm for neural networks—we

already have a sense of how machine learning goes about maximizing predictive accuracy on a dataset in the absence of any fairness constraint. On our St. Fairness College dataset, this process would entail searching through the possible values of the SAT cutoff for the one that minimized the total number of mistakes made (successful students rejected and unsuccessful students accepted, ignoring race). So even though the algorithmic details can be complicated for rich model classes, the basic idea is just a search for the model with the lowest overall error.

But we could equally well search for the model that instead minimized the overall unfairness. After all, for any proposed SAT cutoff, we can easily compute its "unfairness score" by just taking the magnitude of the difference between the number of falsely rejected Circle students and the number of falsely rejected Square students. Using the same principles as for standard error-minimizing machine learning, we could instead design algorithms for unfairness-minimizing machine learning.

Better yet, we could consider both criteria simultaneously. With each model we now associate two numbers: the number of mistakes it makes on the data and its unfairness score on the data. If we had an algorithm that could enumerate these numbers for all models under consideration, we could try to pick the one yielding the "best" tradeoff. But what do we mean by best? Consider our dataset again, and the two race-blind cutoffs we first examined:

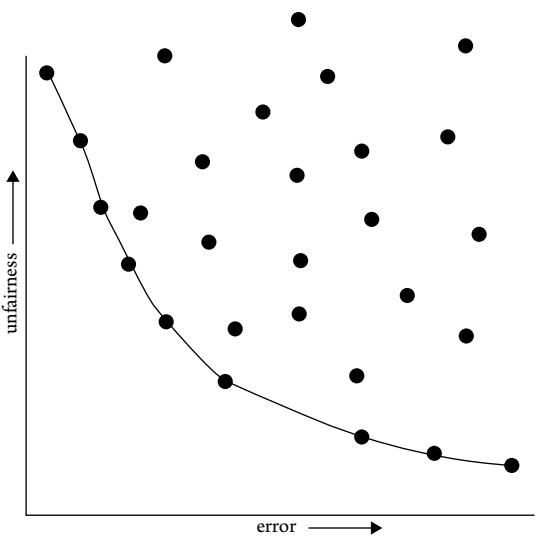


Which is better—the "optimal" cutoff, which makes seven mistakes and has an unfairness score of 4, or the "more fair" cutoff, which makes eight mistakes and has an unfairness score of 2? There is no universally

right answer, because each of these models is better on one criterion and worse on the other. They are thus incomparable, and we should consider both to be reasonable candidates.

But sometimes there are models that really are worse. Consider the cutoff obtained by moving the error-optimal cutoff line three circles to its left, thus now accepting those three unsuccessful students. This model now makes ten mistakes and has the same unfairness score (4) as the optimal cutoff. So it ties the optimal cutoff on unfairness but has strictly higher error. There's no plausible circumstance in which we'd prefer this new model to the optimal one, because we can improve one of our objectives without paying for it in the other—in the language of machine learning, the new model is *dominated* by the error-optimal one. In contrast, neither the "optimal" or "more fair" models dominates the other.

We can generalize this idea across our entire model space. Suppose we enumerated all the possible numerical pairs <*error, unfairness*> achieved by the models we are considering (e.g., SAT cutoffs). Schematically, these pairs would give us a cloud of points that might look something like the following:



So each point corresponds to a different model; the x-coordinate of the point is the model's error, and the y-coordinate is its unfairness score. (For instance, the optimal cutoff would be a point at $x = 7$ and $y = 4$.) Here we have drawn a curve connecting the set of undominated models, which form the southwest (down and to the left) boundary of the set of points. The key thing to realize is that any model that is *not* on this boundary is a "bad" model that we should eliminate from consideration, because we can always improve on either its fairness score or its accuracy (or both) without hurting the other measure by moving to a point on this boundary.

The technical name for this boundary is the Pareto frontier or Pareto curve, and it constitutes the set of "reasonable" choices for the trade-off between accuracy and fairness. Pareto frontiers, which are named after the 19th-century Italian economist Vilfredo Pareto, are actually more general than just accuracy-fairness trade-offs, and can be used to quantify the "good" solutions to any optimization problem in which there are multiple competing criteria. One of the most common examples is the "efficient frontier" in portfolio management, which quantifies the trade-off between returns and risk (or volatility) in stock investing.

The Pareto frontier of accuracy and fairness is necessarily silent about which point we should choose along the frontier, because that is a matter of judgment about the relative importance of accuracy and fairness. The Pareto frontier makes our problem as quantitative as possible, but no more so.

The good news is that generally speaking, whenever we have practical algorithms for "standard," accuracy-only machine learning for a class of models, we also have practical algorithms for tracing out this Pareto frontier. These algorithms will be a little bit more complicated—after all, they must identify a collection of models rather than just a single one—but not by much. For instance, one algorithmic approach is to invent a new, single numerical objective that takes a weighted combination of error and the unfairness score. Thus we might ascribe a

"penalty" to a model that looks like 1/2 times its error plus 1/2 times its unfairness, so the error-optimal cutoff would evaluate to (1/2)7 + (1/2)4 = 5 1/2. We then find the model that minimizes this new weighted penalty, which equally weights error and unfairness. It turns out that the model minimizing this weighted penalty must be one of the points on the Pareto frontier. If we then change the weightings—say, to 1/4 times error plus 3/4 times the unfairness score—we will find another point on the Pareto frontier. So by exploring different combinations of our two objectives, we "reduce" our problem to the single-objective case and can trace out the entire frontier.

While the idea of considering cold, quantitative trade-offs between accuracy and fairness might make you uncomfortable, the point is that there is simply no escaping the Pareto frontier. Machine learning engineers and policymakers alike can be ignorant of it or refuse to look at it. But once we pick a decision-making model (which might in fact be a human decision-maker), there are only two possibilities. Either that model is not on the Pareto frontier, in which case it's a "bad" model (since it could be improved in at least one measure without harm in the other), or it is on the frontier, in which case it implicitly commits to a numerical weighting of the relative importance of error and unfairness. Thinking about fairness in less quantitative ways does nothing to change these realities—it only obscures them.

Making the trade-off between accuracy and fairness quantitative does *not* remove the importance of human judgment, policy, and ethics—it simply focuses them where they are most crucial and useful, which is in deciding exactly which model on the Pareto frontier is best (in addition to choosing the notion of fairness in the first place, and which group or groups merit protection under it, both of which we discuss shortly). Such decisions should be informed by many factors that cannot be made quantitative, including what the societal goal of protecting a particular group is and what is at stake. Most of us would agree that while both racial bias in the ads users are shown online and racial bias in lending decisions are undesirable,

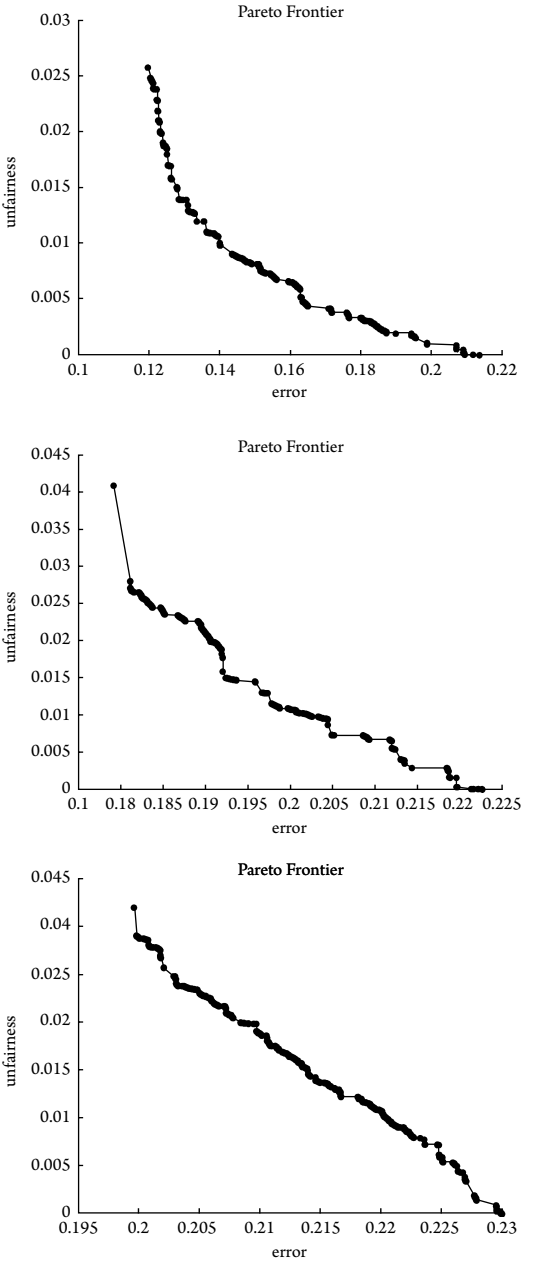


Fig. 15. Examples of Pareto frontiers of error (x axis) and an unfairness measure (y axis) for three different real datasets. The curves differ in their shapes and the actual numeric values on the error and fairness axes, thus presenting different trade-offs.

the potential harms to individuals in the latter far exceed those in the former. So in choosing a point on the Pareto frontier for a lending algorithm, we might prefer to err strongly on the side of fairness—for example, insisting that the false rejection rate across different racial groups be very nearly equal, even at the cost of reducing bank profits. We'll make more mistakes this way—both false rejections of creditworthy applicants and loans granted to parties who will default—but those mistakes will not be disproportionately concentrated in any one racial group. This is the bargain we must accept for strong fairness guarantees.

FAIRNESS FIGHTING FAIRNESS

Even before we arrive at the role of human judgment in the choice of a model on the Pareto frontier, there is the question of which fairness notion we want to use in the first place. As we've already seen, there is more than one reasonable choice. Statistical parity might be appropriate in settings where we simply want to distribute some opportunity, like free tickets to a concert, equally across groups, and there is no notion of merit (like creditworthiness) that is relevant. Approximate equality of false negatives (rejections) across groups might be appropriate in lending decisions. In picking whose tax returns to audit, equality of false positives (audits that discover nothing illegal) across groups might be the goal, since here it is the false positives—law-abiding citizens who are nevertheless subjected to a costly audit—who are harmed. And there are other reasonable fairness definitions where these came from.

In the same way that it was natural to hope for models that are as accurate and fair as possible, we might also hope to have it all when it comes to definitions of fairness. There may always be a trade-off between accuracy and fairness, but why not at least have our fairness notion be as strong as possible? For example, why not define fairness to mean satisfying statistical parity *and* equality of false negatives *and* equality of false positives *and* whatever else we can think of?

Alas, as with the Pareto frontier, we again encounter some stark barriers to all-encompassing definitions of fairness. It turns out there are certain combinations of fairness criteria that—although they are each individually reasonable—simply cannot be achieved simultaneously, even if we ignore accuracy considerations. There are mathematical theorems demonstrating this impossibility. One example is the combination of equality of both false positive and false negative rates across groups, along with another fairness notion called equality of positive predictive value. This simply asks that (for example) among those people the algorithm recommends be granted a loan, the repayment rates across racial groups are roughly the same. This is a desirable property for a predictive algorithm to have, because if the algorithm doesn't have it—that is, if the Circle applicants it recommends for loans end up making the bank less money than the Square applicants—the human decision-makers ultimately responsible for making loans will have a strong incentive to let race influence their decisions when they decide whether or not to follow the recommendations of their model. And they will be rational to do so, since if the model doesn't have equal positive predictive value for both races, its predictions for Circle applicants really will mean something different than its predictions for Square applicants.

So here we have a situation in which there are three different fairness definitions, each of which is entirely reasonable and desirable, and each of which can be achieved in isolation (albeit at some cost to accuracy)—but which together are simply impossible to achieve. Thus in addition to trade-offs with accuracy for any single fairness definition, there are even trade-offs we must make between fairness notions.[2]

[2] This is an interesting contrast to what we saw in the privacy chapter, where there does seem to be a single framework (differential privacy) that captures much of what we could (reasonably) want in a privacy definition and yet still permitted powerful uses of data such as machine learning. In other words, we already know that the study of algorithmic fairness will necessarily be "messier" than the study of algorithmic privacy and that we will have to entertain multiple, incompatible

These stark mathematical constraints on fairness are somewhat depressing, but they also identify and reinforce the central role that people and society will always have to play in fair decision-making, regardless of the extent to which algorithms and machine learning are adopted. They reveal that while algorithms can excel at computing the Pareto frontier once we commit to a definition of fairness, they simply cannot tell us which definition of fairness to use, or which model on the frontier to choose. These are subjective, normative decisions that cannot (fruitfully) be made scientific, and in many ways they are the most important decisions in the overall process we have been describing.

PREVENTING "FAIRNESS GERRYMANDERING"

There's one more crucial and subjective decision in this process that we've ignored so far, and that's the choice of which groups of individuals we want to protect in the first place. We've given examples of concerns over gender bias in word embeddings and racial discrimination in lending and college admissions. But race and gender are just two of many attributes we might want to consider. Extensive debate has also taken place over discrimination based on age, disability status, wealth, nationality, sexual orientation, and many other factors. The US Equal Employment Opportunity Commission even has regulations forbidding discrimination based on any type of "genetic information," an extremely broad category that clearly includes race and gender but also much more, including genetic factors yet to be discovered. As with the choices of fairness definition or which model on the resulting Pareto frontier we want to use, there's no "right answer," and no sensible role for algorithms or machine learning, in the choice

definitions of fairness. We might wish things were otherwise, but we must nevertheless proceed. But perhaps it does increase our appreciation of the strengths of differential privacy.

of which attributes or groups of people we want to protect—this is a decision for society to make.

One theme running throughout this book is that algorithms generally, and especially machine learning algorithms, are good at optimizing what you ask them to optimize, but they cannot be counted on to do things you'd like them to do but didn't ask for, nor to avoid doing things you don't want but didn't tell them not to do. Thus if we ask for accuracy but don't mention fairness, we won't get fairness. If we ask for one kind of fairness, we'll get that kind but not others. As we have seen, sometimes these tensions and trade-offs are mathematically inevitable, and sometimes they arise just because we didn't specify everything we did and didn't want.

This same theme holds true for the choice of which groups we protect. In particular, one recently discovered phenomenon is what we might call "fairness gerrymandering," in which multiple overlapping groups are protected, but at the expense of discrimination against some intersection of them. For example, imagine we want to distribute free tickets to see the Pope and want to protect both gender and race—so the same fraction of men and women should receive tickets, and also the same fraction of Circle and Square people (continuing our use of a fictitious two-race population). Suppose we have enough tickets for 20 percent of the population overall to see the Pope, and there are equal numbers of all four attribute combinations (Circle men, Circle women, Square men, Square women)—say twenty of each, for a total population of eighty, and thus we must distribute sixteen tickets total. (See Figure 16.)

We can probably agree that the "most fair" solution would be to give 4 tickets each to Circle men, Circle women, Square men, and Square women. But that's not what we ask for when we specify that we want to be fair with respect to race and gender *separately*. From that perspective, an equally fair solution is the "gerrymandered" one that gives eight tickets to Circle men, eight tickets to Square women, and no tickets to the other two groups. We still have eight tickets going to men, eight

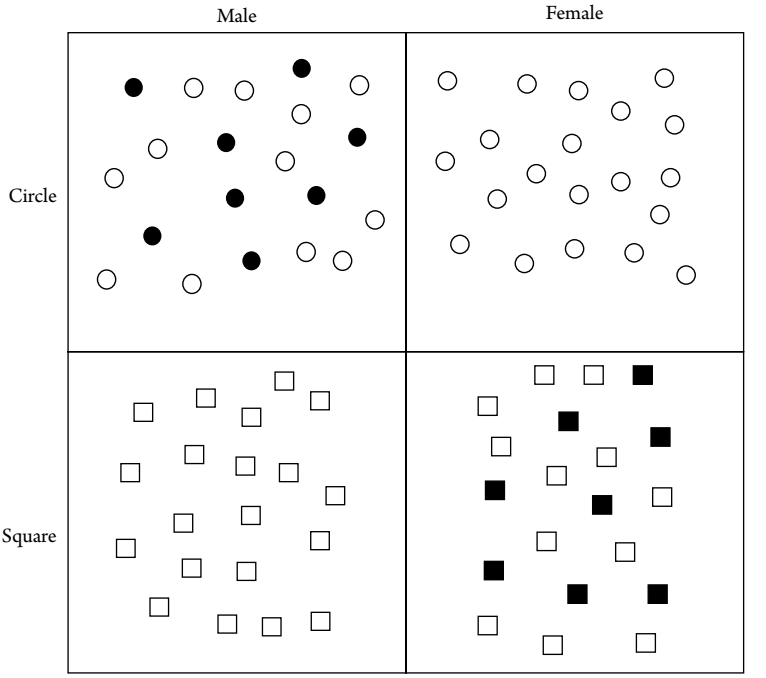


Fig. 16. Illustration of fairness gerrymandering, in which the winners of tickets (shaded circles and squares) are disproportionately concentrated in small subgroups despite being both gender- and race-fair separately.

to women, eight to Circles, and eight to Squares. It's just that we've concentrated those tickets in certain more refined subgroups at the expense of others (namely, Circle women and Square men).

One might ask who would come up with such a complex and discriminatory solution. The answer is that machine learning would, if it improves the accuracy of some prediction task—even by a tiny amount. We didn't ask that these more refined subgroups also be protected, only that the top-level attributes of race and gender be protected. If we also wanted such subgroup protections, we should have said so. And once we see this problem, there seems to be the potential for a kind of infinite regress, in which despite avoiding discrimination by race, gender, age, income, disability, and sexual orientation in

isolation, we find ourselves with a model that, for example, unfairly treats disabled gay Hispanic women over age fifty-five making less than $50,000 annually.

While it is still early days for such topics, recent research has suggested improved algorithms for coping with fairness gerrymandering. These algorithms have the somewhat intuitive and appealing form of a two-player game between a Learner and a Regulator. The Learner is always trying to maximize predictive accuracy but is constrained to be fair to possibly complex subgroups identified by the Regulator (such as disabled gay Hispanic women over age fifty-five making less than $50,000 annually). A back-and-forth ensues, in which the Regulator finds new subgroups suffering discrimination under the Learner's model so far, and the Learner attempts to correct this discrimination while still preserving as much accuracy as possible. This process is guaranteed to quickly result in a model that is fair to all subgroups the Regulator is interested in protecting—even if the Regulator is potentially concerned with a great many groups, of the sort that result when we (for example) consider all possible ways of taking intersections of different categories of people. Experimental work suggests that this much stronger notion of subgroup fairness can be met while still providing usefully accurate models. In fact, Figure 15 earlier in the chapter, showing Pareto curves between error and unfairness on real datasets, is using this gerrymander-free fairness measure.

Of course, once we contemplate fairness for more refined subgroups in a population, it's hard not to take things to their logical conclusion, which would be promises of protections for individuals rather than just groups. After all, if we take a traditional statistical fairness notion such as equality of false rejection rates in lending, if you are one of the creditworthy Square applicants who has been denied a loan, how comforting is it to be told that there was also a creditworthy Circle applicant who was rejected to "compensate" for your mistreatment?

But if we go too far down the path toward individual fairness, other difficulties arise. In particular, if our model makes even a single mistake, then it can potentially be accused of unfairness toward that one individual, assuming it makes any loans at all. And anywhere we apply machine learning and statistical models to historical data, there are bound to be mistakes except in the most idealized settings. So we can ask for this sort of individual level of fairness, but if we do so naively, its applicability will be greatly constrained and its costs to accuracy are likely to be unpalatable; we're simply asking for too much. Finding reasonable ways to give meaningful alternative fairness guarantees to individuals is one of the most exciting areas of ongoing research.

BEFORE AND BEYOND ALGORITHMS

Our focus in this chapter has been on machine learning algorithms, the predictive and decision-making models they output, and the tensions between accuracy and fairness. But there are other places in the typical workflow of machine learning in which fairness concerns can arise, both before we get to algorithms and models and after we've deployed them.

Let's begin with the "before"—namely, with the data collected and fed to a machine learning algorithm in the first place. Throughout much of the chapter we have implicitly assumed that this data is itself correct and not already corrupted by human bias. Our main goal was to design algorithms that do not introduce discrimination as a side effect of the optimization of predictive accuracy. But what if the data itself has been gathered as part of an already discriminatory process? For example, maybe we wish to predict crime risk, but we don't have data on who commits crimes—we only have data on who was arrested. If police officers already exhibit racial bias in their arrest patterns, this will be reflected in the data.

For another example, returning to our hypothetical admissions scenario at St. Fairness College, suppose human admissions officers

historically had a much better understanding of the Circle applicants who form the majority population than they did of the minority Square applicants—being more familiar with the high schools attended by Circles, understanding Circle essays and extracurricular activities better, and generally being more informed about the Circle population. Maybe the admissions officers weren't consciously biased against the Square applicants; they just know the Circle applicants better. It would not be at all surprising if these admissions officers were much better at accurately picking Circle applicants who will succeed at St. Fairness than they are at picking Square applicants who will succeed. Then, remembering that the college only learns about the success or failure of applicants it chooses to admit and not the ones it rejects, the obvious outcome is that the $\langle \boldsymbol{x}, y \rangle$ data generated by past admissions decisions will make Circle applicants look much better than the Square applicants in aggregate—not because they are better in reality but only because of the relative expertise of the admissions officers and the skewed sample of data that they produce.

And if it is the case that in our historical data Circles look better than Squares, there is absolutely no reason to hope that a machine learning algorithm—even one carefully applying all of the anti-discrimination methodology we have been discussing in this chapter—won't learn a model favoring Circle applicants over Square applicants. The problem is that even if we equalize the false rejection rates between Circle applicants and Square applicants on the training data, we will not do so on general populations of Circles and Squares, because the training data is not representative of those general populations. Here the problem is not the algorithm but the mismatch between the input to the algorithm and the real world, caused by the bias already embedded in the data. And this is something we simply cannot expect the algorithm itself to discover and correct. It's just the machine learning version of the computer science adage "Garbage in, garbage out." We might call this version "Bias in, bias out."

The problems can become even more insidious. Sometimes decisions made using biased data or algorithms are the basis for further data collection, forming a pernicious feedback loop that can amplify discrimination over time. An example of this phenomenon comes from the domain of "predictive policing," in which large metropolitan police departments use statistical models to forecast neighborhoods with higher crime rates, and then send larger forces of police officers there. The most popularly used algorithms are proprietary and secret, so there is debate about how these algorithms estimate crime rates, and concern that some police departments might be in part using arrest data. Of course, even if neighborhoods A and B have the same underlying rates of crime, if we send more police to A than to B, we naturally will discover more crime in A as well. If this data is then fed back into the next update of our model, we'll "confirm" that sending more police to A than to B was the right thing to do, and send even more next time. In this way, even small random fluctuations in observed crime rates can lead to self-fulfilling prophecies of enforcement that have no underlying basis in reality. And if the initial training data in this process has not just innocent random fluctuations but is in fact the result of historical crime rates measured during a biased period of police deployment (e.g., one that disproportionately policed minority neighborhoods), the amplification is all but guaranteed. It's another example of a mismatch between the data fed to an algorithm and the world that data is meant to represent, but now accelerated with a feedback loop.

As we have seen, the design of fair machine learning algorithms can be made scientific and is (at least in principle) easy to implement in practice. It would be necessary for companies, organizations, and engineers to be aware of this science, to take it seriously, and to want to incorporate it into their code. Even complex computer programs and systems are often built by relatively small teams, so the number of

people who need to be educated and trained is manageable. But what about the problems arising from biased data collection?

Unfortunately, in many cases these problems are as much social as algorithmic, and are accordingly more difficult. When data collection for college admissions or predictive policing is done by large, distributed, and heterogeneous groups of people, each with their own unknown strengths, weaknesses, and biases, the challenge of imposing clean and principled practices can feel daunting. And in many cases, the scientific solutions suggested by machine learning—such as only training on data gathered during "exploration" phases in which (say) random applicants to St. Fairness are granted admission without anyone looking at their applications—are simply impractical to implement for policy, legal, or social reasons. So while there is now quite a bit of solid science around fairness, there's much more to do to understand how to better connect the narrow purview of an algorithm in isolation with the broader context in which it is embedded. There is much we don't know—but that also makes it an exciting time to be working in this area.