It’s not a hot take to say algorithms are the future. They give us the chance to make sense of whole swaths of data we could never sift through manually, which makes them useful scientific tools, and they are becoming increasingly attractive to policymakers, because they allow computers to bear the brunt of decision making that would ordinarily be tainted with human bias, in areas like insurance, employment, and criminal sentencing. I’m not one of those tech alarmists who believe that robots will rise up to become our overlords, I genuinely think this is a good thing. And it might seem counterintuitive that I think this, because my capstone project is about how much algorithms suck, but here’s a little history lesson for you.

The mid-twentieth century conceptualized prison sentencing as mostly rehabilitative, which gave judges a wide berth within sentencing guidelines to pick the sentence that they thought would serve as the best remedy, given each defendant’s unique circumstances. With almost no objective guidelines and almost no oversight, this quickly created racially biased sentencing trends. Between the latter half of the twentieth century and now, sentencing guidelines became stricter, then looser. Judges have the same berth of discretion, but are now much more cautious - they believe an algorithmic approach could seek out an objective sentence divorced from any prejudices, especially with algorithms like COMPAS, that exclude race as a variable from consideration at all. Most states, from Virginia to Wisconsin, to Colorado, which requires the use of algorithms, to Texas, gravitate towards algorithmic sentencing - even Dallas County uses algorithms to predict the risk of recidivism to assign bail. This could usher in a new era of equalized sentencing, a complete break from our unfortunate past.

Except, the data going into these algorithms bears all the hallmarks of that unfortunate past. To understand the largest problem with sentencing algorithms, we need to talk about what algorithms are. The biggest misconception is that they have to be AI. Algorithms, at their core, are functions. You put data in, the algorithm performs a mathematical operation on that number, and you get an output. The trick is figuring out what to do to them - if humans are the ones controlling the weights, this could contribute human bias to the end result, so more often than not, it’s left up to the algorithm to come up with appropriate weights after being trained on the correspondences between old factors and the outcomes in previous sentencing data. You might have seen the humorous example of this in facial recognition algorithms, that skew towards white data, but that previous sentencing data is the problem. First, overcriminalization of minorities feed skewed outcomes in the training data, which the computer uses to create its new weights. Second, even if the computer doesn’t know the race of the defendant to be racist, it can find other correlations. Other factors included in the algorithm, like zip code, primary language, and prior contact with the police, are highly correlated with race, which simulates the effect of race in these algorithms - this is called induced bias, and it’s the most dangerous because we don’t realize it’s there.

My capstone sought to quantify this bias from a large sample size of published COMPAS data from ProPublica, which was this database of a lot of the common inputs used and the resulting risk score. Obviously, I didn’t have the algorithm, so I had to reverse-engineer the weights that these algorithms most likely used by training my own classifier on this data, teaching the model to match up certain inputs with certain outputs. I used two different classifier approaches so that I could more rigorously verify my results - first, a random forest classifier, which acts like a bunch of trees branching out at several points - this measurement ended up being more precise with its high-risk results. Second, a logistic regression model that fits the data to a kind of s-shaped curve, which, while less precise, gives me the absolute significance of all the factors, rather than their importance relative to each other. I ended up being able to predict a test set of COMPAS data with an accuracy in the high 80s to the low 90s, which is about B+/A- territory. That’s fine, as far as algorithms go, but it’s when the algorithm makes those mistakes that we find the juicy information.

The three standard measurements for calculating bias are proportional parity, specificity parity, and predictive parity. I won’t spend too much time on these because it requires an explanation about precision and false positives and true negatives and that stuff, but the important thing to understand here is that if you fail one of these tests, there is a disparate impact in the data, and there are two pretty clear disparities here. The proportional parity graphs basically signal that the algorithm predicted a lot more positive results for high risk scores for African American defendants, and specificity parity shows that the algorithm predicted a lot more accurate negative results in the high risk scores for Caucasian defendants.

Here’s a measurement that isn’t an official measure of bias but that I think y’all will find very illuminating. I took stock of every time the model incorrectly labeled a person’s risk as higher than it should be, and compared the number of times that happens across races, and here’s the result from our good friend, the random forest classifier. To analyze these results properly, we’re going to need to back up a little bit. When algorithms determine a sentence, they’re measuring the defendant’s risk – specifically, the risk of recidivism, or committing a second crime once you’re out of jail. So, when the computer makes a positive prediction, it’s saying “yes, this defendant is at a high risk for committing another crime.” The algorithm can either get these high-risk predictions right, a true positive, or incorrectly assign a higher risk when the defendant should be lower risk, a false positive. I measured the ratio of all the positive predictions to the true positives, such that the true positive will be 1 when we correctly predict ALL and no more of recidivism, so we want this graph to be as close to the center as possible. The resulting number is basically how many people the algorithm needs to label as high risk to catch one actually high-risk defendant. I’ll let that sink in. For one correct attribution of risk to a high defendant, the computer incriminates 1.83 black defendants – that’s almost one additional person as collateral damage. On the other hand, the odds ratio is 1.22 for Caucasian defendants, which is a much lower probability of incorrectly being labeled as high risk.

Here’s the kicker - the real world is super complex, and the amalgamation of subjective factors that determine a sentence can’t always be predicted accurately from a set of training data. So, it’s important to look at what a computer does when it doesn’t know what to do. When a score is wrong, it is FAR more likely to be a harsher score for black defendants and a more lenient score for white defendants, which is a scarily racialized statistic for an algorithm that doesn’t include race.

So how do we fix this? The obvious solution is an intervention at the level of data, and this is where things get complicated. I want all of us to pretend that we’re designing some sort of classification system, and we’ve just been told that we can either create a perfectly fair model with about 50% accuracy, or a completely accurate model that’s pretty biased. Raise your hand if you’d accept the loss of accuracy for fairness. Raise your hand if you wouldn’t.

This is tricky. An unfair model replicates dangerous problems of the status quo. An inaccurate model defeats the purpose of using computers to begin with. So I researched three broad approaches based on their ability to reconcile this tradeoff.

First is regularization. Regularization is a process that basically takes the weights of a lot of the non-race variables, looks for if any are suspiciously similar to race, and tries to minimize the similarity, which seems promising because it is a pretty objective way to memorize a numerical carrier of bias. A lot of literature describes the problem with trying to minimize the redundant information from this protected variable, which is that it’s not exactly redundant. Arbitrarily removing information from a proxy variable because we think it’s associated with race obviously removes a lot of the power it has in predicting the target.

Next is adversarial debiasing. We have two neural networks. One of them attempts to predict the target risk scores with the information we have. The other is trying to guess what race the defendant is. The fairer the algorithm, the harder it should be to guess the race. This one tries to resolve the problem by having two algorithms that are trying to maximize fairness and accuracy, respectively. The difficult part is getting these two models to converge to a final set of results that maximizes both.

Last is reweighting – we identify the data that causes the bias and re-weight those problem variables, and other similar variables, to explicitly counter the bias. This requires a lot of human experimentation and tinkering with the data, but it’s the one I ended up using. So, after localizing the correlation to a few problem variables and re-weighting to counteract those variables, here’s that graph again, with our new and improved weights. By the way, keep in mind that race is not an explicit variable that is used, so the algorithm has no idea who it’s sentencing. We have a marked improvement – odds go down for black defendants from 1.84 to 1.22, and odds go down for white defendants from 1.22 to 1.11. This is definitely not mission accomplished, and we’ve still got a ways to go, but we’re moving in the right direction towards that gold standard of 1.

Okay, sure, but I just spent the past two minutes telling you that something has to give. Our fairness standard measures precision, the idea that we don’t accidentally label low risk people as high risk, which seems to say good things about accuracy. What gives is recall, the idea that the model actually catches everybody that is high risk.

This is the tradeoff. Fairness versus accuracy. Precision versus recall. Human empathy versus algorithmic efficiency.

I put it to you that the way our justice system works, precision is more important than recall, because we start with the presumption of innocence. “Innocent until proven guilty” means that guilty should be a high bar. Scrutinizing the source of this bias is important, because it is way too easy to naturalize differences across race and say “well, x group is just *like* that”, and that is honestly an abdication of responsibility to change things. This is really what worries me. Our reweighted model (and other systems like regularization and adversarial debiasing) try to address this. There’s obviously more work to be done, more data to be collected, and more solutions to be tested to see if reweighting and other human intervention is viable, and data poses a promising, objective solution. However, the move to data, and the removal of human discretion that comes with it, is also a way to shield human blame and human responsibility for affirmative responses to correct this bias. So the big question I had earlier, do I choose fairness or accuracy, is less important than the recognition that either fairness or accuracy is a choice. The advent of sentencing algorithms has simply illuminated how important that choice is, for millions of Americans. I hope we as a society make the right one.