PETER SZOLOVITS: OK, so a little over a year ago, I got a call from this committee. NASEM is the National Academy of Science, Engineering, and Medicine. So this is an august body of old people with lots of gray hair who have done something important enough to get elected to these academies.

And their research arm is called the National Research Council and has a bunch of different committees. One of them is this Committee on Science, Technology, and the Law. It's a very interesting committee. It's chaired by David Baltimore, who used to be an MIT professor until he went and became president of Caltech. And he also happens to have a Nobel Prize in his pocket and he's a pretty famous guy.

And Judge David Tatel is a member of the US Court of Appeals for the District of Columbia circuit, so this is probably the most important circuit court. It's one level below the Supreme Court. And he happens to sit in the seat that Ruth Bader Ginsburg occupied before she was elevated to the Supreme Court from that Court of Appeals, so this is a pretty big deal. So these are heavy hitters.

And they convened a meeting to talk about the set of topics that I've listed here. So blockchain and distributed trust, artificial intelligence and decision making, which is obviously the part that I got invited to talk about, privacy and informed consent in an era of big data, science curricula for law schools, emerging issues, and science, technology, and law. The issue of using litigation to target scientists who have opinions that you don't like. And the more general issue of how do you communicate advances in life sciences to a skeptical public. So this is dealing with the sort of anti-science tenor of the times.

So the group of us that talked about AI and decision making, I was a little bit surprised by the focus because Hank really is a law school professor at Stanford who's done a lot of work on fairness and prejudice in health care. Cherise Burdee is at something called the Pretrial Justice Institute, and her issue is a legal one which is that there are now a lot of companies that have software that predict, if you get bail while you're awaiting trial, are you likely to skip bail or not? And so this is influential in the decision that judges make about how much bail to impose and whether to let you out on bail at all or to keep you in prison, awaiting your trial.

Matt Lundgren is a radiology professor at Stanford and has done some of the really cool work

on building convolutional neural network models to detect pulmonary emboli and various other things in imaging data. You know the next guy, and Suresh Venkatasubramanian is a professor. He was originally a theorist at the University of Utah but has also gotten into thinking a lot about privacy and fairness.

And so that that was our panel, and we each gave a brief talk and then had a very interesting discussion. One of the things that I was very surprised by is somebody raised the question of shouldn't Tatel as a judge on the Circuit Court of Appeals hire people like you guys to be clerks in his court? So people like you guys who also happen to go to law school, of which there are a number now of people who are trained in computational methods and machine learning but also have the legal background.

And he said something very interesting to me. He said, no, he wouldn't want people like that, which kind of shocked me. And so we quizzed him a little bit on why, and he said, well, because he views the role of the judge not to be an expert but to be a judge. To be a balancer of arguments on both sides of an issue. And he was afraid that if he had a clerk who had a strong technical background, that person would have strong technical opinions which would bias his decision one way or another.

So this reminded me-- my wife was a lawyer, and I remember, when she was in law school, she would tell me about the classes that she was taking. And it became obvious that studying law was learning how to win, not learning how to find the truth. And there's this philosophical notion in the law that says that the truth will come out from spirited argument on two sides of a question, but your duty as a lawyer is to argue as hard as you can for your side of the argument.

And in fact, in law school, they teach them, like in debate, that you should be able to take either side of any case and be able to make a cogent argument for it. And so Tatel sort of reinforced that notion in what he said, which I thought was interesting.

Well, just to talk a little bit about the justice area because this is the one that has gotten the most public attention, governments use decision automation for determining eligibility for various kinds of services, evaluating where to deploy health inspectors and law enforcement personnel, defining boundaries along voting districts. So all of the gerrymandering discussion that you hear about is all about using computers and actually machine learning techniques to try to figure out how to-- your objective function is to get Republicans or Democrats elected,

depending on who's in charge of the redistricting. And then you tailor these gerrymandered districts in order to maximize the probability that you're going to have the majority in whatever congressional races or state legislative races.

So in the law, people are in favor of these ideas to the extent that they inject clarity and precision into bail, parole, and sentencing decisions. Algorithmic technologies may minimize harms that are the products of human judgment. So we know that people are in fact prejudiced, and so there are prejudices by judges and by juries that play into the decisions made in the legal system. So by formalizing it, you might win.

However, conversely, the use of technology to determine whose liberty is deprived and on what terms raises significant concerns about transparency and interpretability. So next week, we're going to talk some about transparency and interpretability, but today's is really about fairness.

So here is an article from October of last year-- no, September of last year, saying that as of October of this year, if you get arrested in California, the decision of whether you get bail or not is going to be made by a computer algorithm, not by a human being, OK? So it's not 100%. There is some discretion on the part of this county official who will make a recommendation, and the judge ultimately decides, but I suspect that until there are some egregious outcomes from doing this, it will probably be quite commonly used.

Now, the critique of these bail algorithms is based on a number of different factors. One is that the algorithms reflect a severe racial bias. So for example, if you are two identical people but one of you happens to be white and one of you happens to be black, the chances of you getting bail are much lower if you're black than if you're white. Now, you say, well, how could that be given that we're learning this algorithmically? Well, it's a complicated feedback loop because the algorithm is learning from historical data, and if historically, judges have been less likely to grant bail to an African-American than to a Caucasian-American, then the algorithm will learn that that's the right thing to do and will nicely incorporate exactly that prejudice.

And then the second problem, which I consider to be really horrendous, is that in this particular field, the algorithms are developed privately by private companies which will not tell you what their algorithm is. You can just pay them and they will tell you the answer, but they won't tell you how they compute it. They won't tell you what data they used to train the algorithm.

And so it's really a black box. And so you have no idea what's going on in that box other than

by looking at its decisions. And so the data collection system is flawed in the same way as the judicial system itself.

So not only are there algorithms that decide whether you get bail or not, which is after all a relatively temporary question until your trial comes up, although that may be a long time, but there are also algorithms that advise on things like sentencing. So they say, how likely is this patient to be a recidivist? Somebody who, when they get out of jail, they're going to offend again. And therefore, they deserve a longer jail sentence because you want to keep them off the streets.

Well, so this is a particular story about a particular person in Wisconsin, and shockingly, the state Supreme Court ruled against this guy, saying that knowledge of the algorithm's output was a sufficient level of transparency in order to not violate his rights, which I think many people consider to be kind of an outrageous decision. I'm sure it'll be appealed and maybe overturned. Conversely-- I keep doing on the one hand and on the other-- algorithms could help keep people out of jail. So there's a Wired article not long ago that says we can use algorithms to analyze people's cases and say, oh, this person looks like they're really in need of psychiatric help rather than in need of jail time, and so perhaps we can divert him from the penal system into psychiatric care and keep him out of prison and get him help and so on. So that's the positive side of being able to use these kinds of algorithms.

Now, it's not only in criminality. There is also a long discussion-- you can find this all over the web-- of, for example, can an algorithm hire better than a human being. So if you're a big company and you have a lot of people that you're trying to hire for various jobs, it's very tempting to say, hey, I've made lots and lots of hiring decisions and we have some outcome data. I know which people have turned out to be good employees and which people have turned out to be bad employees, and therefore, we can base a first-cut screening method on learning such an algorithm and using it on people who apply for jobs and say, OK, these are the ones that we're going to interview and maybe hire because they look like they're a better bet.

Now, I have to tell you a personal story. When I was an undergraduate at Caltech, the Caltech faculty decided that they wanted to include student members of all the faculty committees. And so I was lucky enough that I served for three years as a member of the Undergraduate Admissions Committee at Caltech.

And in those days, Caltech only took about 220, 230 students a year. It's a very small school. And we would actually fly around the country and interview about the top half of all the applicants in the applicant pool. So we would talk not only to the students but also to their teachers and their counselors and see what the environment was like, and I think we got a very good sense of how good a student was likely to be based on that.

So one day, after the admissions decisions have been made, one of the professors, kind of as a thought experiment, said here's what we ought to do. We ought to take the 230 people that we've just offered admission to and we should reject them all and take the next 230 people, and then see whether the faculty notices. Because it seemed like a fairly flat distribution.

Now, of course, I and others argued that this would be unfair and unethical and would be a waste of all the time that we had put into selecting these people, so we didn't do that. But then this guy went out and he looked at the data we had on people's ranking class, SAT scores, grade point average, the checkmarks on their recommendation letters about whether they were truly exceptional or merely outstanding. And he built a linear regression model that predicted the person's sophomore level grade point average, which seemed like a reasonable thing to try to predict.

And he got a reasonably good fit, but what was disturbing about it is that in the Caltech population of students, it turned out that the beta for your SAT English performance was negative. So if you did particularly well in English on the SAT, you were likely to do worse as a sophomore at Caltech than if you didn't do as well. And so we thought about that a lot, and of course, we decided that that would be really unfair to penalize somebody for being good at something, especially when the school had this philosophical orientation that said we ought to look for people with broad educations. So that's just an example.

And more, Science Friday had a nice show that you can listen to about this issue. So let me ask you, what do you mean by fairness? If we're going to define the concept, what is fair? What characteristics would you like to have an algorithm have that judges you for some particular purpose? Yeah?

AUDIENCE: It's impossible to pin down sort of, at least might in my opinion, one specific definition, but for the pre-trial success rate for example, I think having the error rates be similar across populations, across the covariants you might care about, for example, fairness, I think is a good start.

PETER SZOLOVITS: OK, so similar error rates is definitely one of the criteria that people use in talking about fairness. And you'll see later Irene-- where's Irene? Right there. Irene is a master of that notion of fairness. Yeah?

AUDIENCE: When the model says some sort of observation that causally shouldn't be true, and what I want society to look like

PETER SZOLOVITS: So I'm not sure how to capture that in a short phrase. Societal goals. But that's tricky, right? I mean, suppose that I would like it to be the case that the fraction of people of different ethnicity who are criminals should be the same.

That seems like a good goal for fairness. How do I achieve that? I mean, I could pretend that it's the same, but it isn't the same today objectively, and the data wouldn't support that. So that's an issue. Yeah?

AUDIENCE: People who are similar should be treated similarly, so engaged sort of independent of the [INAUDIBLE] attributes or independent of your covariate.

PETER SZOLOVITS: Similar people should lead to similar treatment. Yeah, I like that.

AUDIENCE: I didn't make it up.

PETER SZOLOVITS: I know. It's another of the classic sort of notions of fairness. That puts a lot of weight on the distance function, right? In what way are to people similar? And what characteristics-- you obviously don't want to use the sensitive characteristics, the forbidden characteristics in order to decide similarity, because then people will be dissimilar in ways that you don't want, but defining that function is a challenge.

All right, well, let me show you a more technical approach to thinking about this. So we all know about biases like selection bias, sampling bias, reporting bias, et cetera. These are in the conventional sense of the term bias. But I'll show you an example that I got involved in.

Raj Manrai was a MIT Harvard HST student, and he started looking at the question of the genetics that was used in order to determine whether somebody is at risk for cardiomyopathy, hypertrophic cardiomyopathy. That's a big word. It means that your heart gets too big and it becomes sort of flabby and it stops pumping well, and eventually, you die of this disease at a relatively young age, if, in fact, you have it.

So what happened is that there was a study that was done mostly with European populations where they discovered that a lot of people who had this disease had a certain genetic variant. And they said, well, that must be the cause of this disease, and so it became accepted wisdom that if you had that genetic variant, people would counsel you to not plan on living a long life. And this has all kinds of consequences.

Imagine if you're thinking about having a kid when you're in your early 40s, and your life expectancy is 55. Would you want to die when you have a teenager that you leave to your spouse? So this was a consequential set of decisions that people have to make.

Now, what happened is that in the US, there were tests of this sort done, but the problem was that a lot of African and African-American populations turned out to have this genetic variant frequently without developing this terrible disease, but they were all told that they were going to die, basically. And it was only after years when people noticed that these people who were supposed to die genetically weren't dying that they said, maybe we misunderstood something. And what they misunderstood was that the population that was used to develop the model was a European ancestry population and not an African ancestry population.

So you go, well, we must have learned that lesson. So this paper was published in 2016, and this was one of the first in this area. Here's a paper that was published three weeks ago in *Nature* Scientific Reports that says, genetic risk factors identified in populations of European descent do not improve the prediction of osteoporotic fracture and bone mineral density in Chinese populations.

So it's the same story. It's exactly the same story. Different disease, the consequence is probably less dire because being told that you're going to break your bones when you're old is not as bad as being told that your heart's going to stop working when you're in your 50s, but there we have it.

OK, so technically, where does bias come from? Well, I mentioned the standard sources, but here is an interesting analysis. This comes from Constantine Aliferis from a number of years ago, 2006, and he says, well, look, in a perfect world, if I give you a data set, there's an uncountably infinite number of models that might possibly explain the relationships in that data. I cannot enumerate an uncountable number of models, and so what I'm going to do is choose some family of models to try to fit, and then I'm going to use some fitting technique, like stochastic gradient descent or something, that will find maybe a global optimum, but

maybe not. Maybe it'll find the local optimum.

And then there is noise. And so his observation is that if you count O as the optimal possible model over all possible model families, and if you count L as the best model that's learnable by a particular learning mechanism, and you call A the actual model that's learned, then the bias is essentially O minus L, so its limitation of learning method related to the target model. The variance is like L minus A, it's the error that's due to the particular way in which you learned things, like sampling and so on, and you can estimate the significance of differences between different models by just permuting the data, randomizing, essentially, the relationships in the data. And then you get a curve of performance of those models, and if yours lies outside the 95% confidence interval, then you have a P equal 0.05 result that this model is not random. So that's the typical way of going about this.

Now, you might say, but isn't discrimination the very reason we do machine learning? Not discrimination in the legal sense, but discrimination in the sense of separating different populations. And so you could say, well, yes, but some basis for differentiation are justified and some basis for differentiation are not justified. So they're either practically irrelevant, or we decide for societal goals that we want them to be irrelevant and we're not going to take them into account.

So one lesson from people who have studied this for a while is that discrimination is domain specific. So you can't define a universal notion of what it means to discriminate because it's very much tied to these questions of what is practically and morally irrelevant in the decisions that you're making. And so it's going to be different in criminal law than it is in medicine, than it is in hiring, than it is in various other fields, college admissions, for example. And it's feature-specific as well, so you have to take the individual features into account.

Well, historically, the government has tried to regulate these domains, and so credit is regulated by the Equal Credit Opportunity Act, education by the Civil Rights Act and various amendments, employment by the Civil Rights Act, housing by the Fair Housing Act, public accommodation by the Civil Rights Act, more recently, marriage is regulated originally by the Defense of Marriage Act, which as you might tell from its title, was against things like people being able to marry who were not a traditional marriage that they wanted to defend, but it was struck down by the Supreme Court about six years ago as being discriminatory. It's interesting, if you look back to probably before you guys were born in 1967, until 1967, it was illegal for an African-American and a white to marry each other in Virginia. It was literally illegal. If you went

to get a marriage license, you were denied, and if you got married out of state and came back, you could be arrested.

This happened much later. Trevor Noah, if you know him from *The Daily Show,* wrote a book called *Born a Crime,* I think, and his father is white Swiss guy and his mother is a South African black, and so it was literally illegal for him to exist under the apartheid laws that they had. He had to pretend to be-- his mother was his caretaker rather than his mother in order to be able to go out in public, because otherwise, they would get arrested. So this has recently, of course, also disappeared, but these are some of the regulatory issues. So here are some of the legally recognized protected classes, race, color, sex, religion, national origin, citizenship, age, pregnancy, familial status, disability, veteran status, and more recently, sexual orientation in certain jurisdictions, but not everywhere around the country.

OK, so given those examples, there are two legal doctrines about discrimination, and one of them talks about disparate treatment, which is sort of related to this one. And the other talks about disparate impact and says, no matter what the mechanism is, if the outcome is very different for different racial groups typically or gender groups, then there is prima facie evidence that there is something not right, that there is some sort of discrimination. Now, the problem is, how do you defend yourself against, for example, a disparate impact argument? Well, you say, in order to be disparate impact that's illegal, it has to be unjustified or avoidable.

So for example, suppose I'm trying to hire people to climb 50-story buildings that are under construction, and you apply, but it turns out you have a medical condition which is that you get dizzy at times, I might say, you know what, I don't want to hire you, because I don't want you plopping off the 50th floor of a building that's under construction, and that's probably a reasonable defense. If I brought suit against you and said, hey, you're discriminating against me on the basis of this medical disability, a perfectly good defense is, yeah, it's true, but it's relevant to the job. So that's one way of dealing with it.

Now, how do you demonstrate disparate impact? Well, the court has decided that you need to be able to show about a 20% difference in order to call something disparate impact. So the question, of course, is can we change our hiring policies or whatever policies we're using in order to achieve the same goals, but with less of a disparity in the impact. So that's the challenge.

Now, what's interesting is that disparate treatment and disparate impact are really in conflict

with each other. And you'll find that this is true in almost everything in this domain. So disparate impact is about distributive justice and minimizing equality of outcome. Disparate treatment is about procedural fairness and equality of opportunity, and those don't always mesh. In other words, it may well be that equality of opportunity still leads to differences in outcome, and you can't square that circle easily.

Well, there's a lot of discrimination that keeps persisting. There's plenty of evidence in the literature. And one of the problems is that, for example, take an issue like the disparity between different races or different ethnicities. It turns out that we don't have a nicely balanced set where the number of people of European descent is equal to the number of people of African-American, or Hispanic, or Asian, or whatever population you choose descent, and therefore, we tend to know a lot more about the majority class than we know about these minority classes, and just that additional data and that additional knowledge might mean that we're able to reduce the error rate simply because we have a larger sample size.

OK, so if you want to formalize this, this is Moritz Hardt's part of the tutorial that I'm stealing from in this talk. This was given at KDD about a year and a half ago, I think. And Moritz is a professor at Berkeley who actually teaches an entire semester-long course on fairness in machine learning, so there's a lot of material here.

And so he formalizes the problem this way. He says, look, a decision problem, a model, in our terms, is that we have some X, which is the set of features we know about an individual, and we have some said A, which is the set of protected features, like your race, or your gender, or your age, or whatever it is we're trying to prevent from discriminating on, and then we have either a classifier or some score or predictive function that's a function of X and A in either case, and then we have some Y, which is the outcome that we're interested in predicting. So now you can begin to tease apart some different notions of fairness by looking at the relationships between these elements.

So there are three criteria that appear in the literature. One of them is the notion of independence of the scoring function from sensitive attributes. So this says that R is independent from A. Remember, on the previous slide, I said that R is a function of-- oops. R is a function of X and A, so obviously, that criterion says that it can't be a function of A. Null function.

Another notion is separation of score and the sensitive attribute given the outcome. So this is

the one that says the different groups are going to be treated similarly. In other words, if I tell you the group, the outcome, the people who did well at the job and the people who did poorly at the job, then the scoring function is independent of the protected attribute. So that allows a little more wiggle room because it says that the protected attribute can still predict something about the outcome, it's just that you can't use it in the scoring function given the category of which outcome category that individual belongs to.

And then sufficiency is the inverse of that. It says that given the scoring function, the outcome is independent of the protected attribute. So that says, can we build a fair scoring function that separates the outcome from the protected attribute?

So here's some detail on those. If you look at independence-- this is also called by various other names-- basically, what it says is that the probability of a particular result, R equal 1, is the same whether you're in class A or class B in the protected attribute. So what does that tell you? That tells you that the scoring function has to be universal over the entire data set and has to not distinguish between people in class A versus class B. That's a pretty strong requirement.

And then you can operationalize the notion of unfairness either by looking for an absolute difference between those probabilities. If it's greater than some epsilon, then you have evidence that this is not a fair scoring function, or a ratio test that says, we look at the ratio, and if it differs from 1 significantly, then you have evidence that this is an unfair scoring function. And by the way, this relates to the 4/5 rule, because if you make epsilon 20%, then that's the same as the 4/5 rule.

Now, the problem-- there are problems with this notion of independence. So it only requires equal rates of decisions for hiring, or giving somebody a liver for transplant, or whatever topic you're interested in. And so what if hiring is based on a good score in group A, but random in B? So for example, what if we know a lot more information about group A than we do about group B, so we have a better way of scoring them than we do of scoring group B. So you might wind up with a situation where you wind up hiring the same number of people, the same ratio of people in both groups, but in one group, you've done a good job of selecting out the good candidates, and in the other group, you've essentially done it at random.

Well, the outcomes are likely to be better for a group A than for group B, which means that you're developing more data for the future that says, we really ought to be hiring people in

group A because they have better outcomes. So there's this feedback loop. Or alternatively-- well, of course, it could be caused by malice also. I could just decide as a hiring manager I'm not hiring enough African-Americans so I'm just going to take some random sample of African-Americans and hire them, and then maybe they'll do badly, and then I'll have more data to demonstrate that this was a bad idea. So that would be malicious.

There's also a technical problem, which is it's possible that the category, the group is a perfect predictor of the outcome, in which case, of course, they can't be separated. They can't be independent of each other. Now, how do you achieve independence? Well, there are a number of different techniques.

One of them is-- there's this article by Zemel about learning fair representations, and what it says is you create a new world representation, Z, which is some combination of X and A, and you do this by maximizing the mutual information between X and Z and by minimizing the mutual information between the A and Z. So this is an idea that I've seen used in machine learning for robustness rather than for fairness, where people say, the problem is that given a particular data set, you can overfit to that data set, and so one of the ideas is to do a Gann-like method where you say, I want to train my classifier, let's say, not only to work well on getting the right answer, but also to work as poorly as possible on identifying which data set my example came from.

So this is the same sort of idea. It's a representation learning idea. And then you build your predictor, R, based on this representation, which is perhaps not perfectly independent of the protected attribute, but is as independent as possible. And usually, there are knobs in these learning algorithms, and depending on how you turn the knob, you can affect whether you're going to get a better classifier that's more discriminatory or a worse classifier that's less discriminatory.

So you can do that in pre-processing. You can do some kind of incorporating in the loss function a dependence notion or an independence notion and say, we're going to train on a particular data set, imposing this notion of wanting this independence between A and R as part of our desiderata. And so you, again, are making trade-offs against other characteristics.

Or you can do post-processing. So suppose I've built an optimal R, not worrying about discrimination, then I can do another learning problem that says I'm now going to build a new F, which takes R and the protected attribute into account, and it's going to minimize the cost of

misclassifications. And again, there's a knob where you can say, how much do I want to emphasize misclassifications for the protected attribute or based on the protected attribute?

So this was still talking about independence. The next notion is separation, that says given the outcome, I want to separate A and R. So that graphical model shows that the protected attribute is only related to the scoring function through the outcome. So there's nothing else that you can learn from one to the other than through the outcome.

So this recognizes that the protected attribute may, in fact, be correlated with the target variable. An example might be different success rates in a drug trial for different ethnic populations. There are now some cardiac drugs where the manufacturer has determined that this drug works much better in certain subpopulations than it does in other populations, and the FDA has actually approved the marketing of that drug to those subpopulations.

So you're not supposed to market it to the people for whom it doesn't work as well, but you're allowed to market it specifically for the people for whom it does work well. And if you think about the personalized medicine idea, which we've talked about earlier. The populations that we're interested in becomes smaller and smaller until it may just be you. And so there might be a drug that works for you and not for anybody else in the class, but it's exactly the right drug for you, and we may get to the point where that will happen and where we can build such drugs and where we can approve their use in human populations.

Now, the idea here is that if I have two populations, blue and green, and I draw ROC curves for both of these populations, they're not going to be the same, because the drug will work differently for those two populations. But on the other hand, I can draw them on the same axes, and I can say, look any place within this colored region can be a fair region in that I'm going to get the same outcome for both populations. So I can't achieve this outcome for the blue population or this outcome for the green population, but I can achieve any of these outcomes for both populations simultaneously. And so that's one way of going about satisfying this requirement when it is not easily satisfied.

So the advantage of separation over independence is that it allows correlation between R and Y, even a perfect predictor, so R could be a perfect predictor for Y. And it gives you incentives to learn to reduce the errors in all groups. So that issue about randomly choosing members of the minority group doesn't work here because that would suppress the ROC curve to the point where there would be no feasible region that you would like. So for example, if it's a coin flip,

then you'd have the diagonal line and the only feasible region would be below that diagonal, no matter how good the predictor was for the other class. So that's a nice characteristic.

And then the final criterion is sufficiency, which flips R and Y. So it says that the regressor or the predictive variable can depend on the protected class, but the protected class is separated from the outcome. So for example, the probability in a binary case of a true outcome of Y given that R is some particular value, R and A is a particular class, is the same as the probability of that same outcome given the same R value, but the different class. So that's related to the sort of similar people, similar treatment notion, qualitative notion, again.

So it requires a parody of both the positive and the negative predictive values across different groups. So that's another popular way of looking at this. So for example, if the scoring function is a probability, or the set of all instances assigned the score R has an R fraction of positive instances among them, then the scoring function is said to be well-calibrated.

So we've talked about that before in the class. If it turns out that R is not well-calibrated, you can hack it and you can make it well-calibrated by putting it through a logistic function that will then approximate the appropriately calibrated score, and then you hope that that calibration will give-- or the degree of calibration will give you a good approximation to this notion of sufficiency. These guys in the tutorial also point out that some data sets actually lead to good calibration without even trying very hard.

So for example, this is the UCI census data set, and it's a binary prediction of whether somebody makes more than $50,000 a year if you have any income at all and if you're over 16 years old. And the feature, there are 14 features, age, type of work, weight of sample is some statistical hack from the Census Bureau, your education level, marital status, et cetera, and what you see is that the calibration for males and females is pretty decent. It's almost exactly along the 45 degree line without having done anything particularly dramatic in order to achieve that. On the other hand, if you look at the calibration curve by race for whites versus blacks, the whites, not surprisingly, are reasonably well-calibrated, and the blacks are not as well-calibrated. So you could imagine building some kind of a transformation function to improve that calibration, and that would get you separation.

Now, there's a terrible piece of news, which is that you can prove, as they do in this tutorial, that it's not possible to jointly achieve any pair of these conditions. So you have three reasonable technical notions of what fairness means, and they're incompatible with each other

except in some trivial cases. This is not good.

And I'm not going to have time to go into it, but there's a very nice thing from Google where they illustrate the results of adopting one or another of these notions of fairness on a synthesized population of people, and you can see how the trade-offs vary and what the results are of choosing different notions of fairness. So it's a kind of nice graphical hack. Again, it'll be on the slides, and I urge you to check that out, but I'm not going to have time to go into it.

There is one other problem that they point out which is interesting. So this was a scenario where you're trying to hire computer programmers, and you don't want to take gender into account because we know that women are underrepresented among computer people, and so we would like that not to be an allowed attribute in order to decide to hire someone. So they say, well, there are two scenarios.

One of them is that gender, A, influences whether you're a programmer or not. And this is empirically true. There are fewer women who are programmers.

It turns out that visiting Pinterest is slightly more common among women than men. Who knew? And then visiting GitHub is much more common among programmers than among non-programmers. That one's pretty obvious.

So what they say is, if you want an optimal predictor of whether somebody's going to get hired, it should actually take both Pinterest visits and GitHub visits into account, but because those go back to gender, which is an unusable attribute, they don't like this model. And so they say, well, we could use an optimal separated score, because now, being a programmer separates your gender from the scoring function. And so we can create a different score which is not the same as the optimal score, but is permitted because it's no longer dependent on your sex, on your gender.

Here's another scenario that, again, starts with gender and says, look, we know that there are more men than women who obtain college degrees in computer science, and so there's an influence there, and computer scientists are much more likely to be programmers than non-computer science majors. If you're were a woman-- has anybody visited the Grace Murray Hopper Conference? A couple, a few of you.

So this is a really cool conference. Grace Murray Hopper invented the notion bug or the term

bug and was a really famous computer scientist starting back in the 1940s when there were very few of them, and there is a yearly conference for women computer scientists in her honor. So clearly, the probability that you visited the Grace Hopper Conference is dependent on your gender. It's also dependent on whether you're a computer scientist, because if you're a historian, you're not likely to be interested in going to that conference.

And so in this story, the optimal score is going to depend basically on whether you have a computer science degree or not, but the separated score will depend only on your gender, which is kind of funny, because that's the protected attribute. And what these guys point out is that despite the fact that you have these two scenarios, it could well turn out that the numerical data, the statistics from which you estimate these models are absolutely identical. In other words, the same fraction of people are men and women, the same fraction of people are programmers, they have the same relationship to those other factors, and so from a purely observational viewpoint, you can't tell which of these styles of model is correct or which version of fairness your data can support. So that's a problem because we know that these different notions of fairness are in conflict with each other.

So I wanted to finish by showing you a couple of examples. So this was a paper based on Irene's work. So Irene, shout if I'm butchering the discussion.

I got an invitation last year from the American Medical Association's *Journal of Ethics,* which I didn't know existed, to write a think piece for them about fairness in machine learning, and I decided that rather than just bloviate, I wanted to present some real work, and Irene had been doing some real work. And so Marcia, who was one of my students, and I convinced her to get into this, and we started looking at the question of how these machine learning models can identify and perhaps reduce disparities in general medical and mental health. Now, why those two areas? Because we had access to data in those areas.

So the general medical was actually not that general. It's intensive care data from MIMIC, and mental health care is some data that we had access to from Mass General and McLean's hospital here in Boston, which both have big psychiatric clinics. So yeah, this is what I just said.

So the question we were asking is, is there bias based on race, gender, and insurance type? So we were really interested in socioeconomic status, but we didn't have that in the database, but the type of insurance you have correlates pretty well with whether you're rich or poor. If you have Medicaid insurance, for example, you're poor, and if you have private insurance, the

first approximation, you're rich. So we did that, and then we looked at the notes. So we wanted to see not the coded data, but whether the things that nurses and doctors said about you as you were in the hospital were predictive of readmission, of 30-day readmission, of whether you were likely to come back to the hospital.

So these are some of the topics. We used LDA, standard topic modeling framework. And the topics, as usual, include some garbage, but also include a lot of recognizably useful topics. So for example, mass, cancer, metastatic, clearly associated with cancer, Afib, atrial, Coumadin, fibrillation, associated with heart function, et cetera, in the ICU domain.

In the psychiatric domain, you have things like bipolar, lithium, manic episode, clearly associated with bipolar disease, pain, chronic, milligrams, the drug quantity, associated with chronic pain, et cetera. So these were the topics that we used. And so we said, what happens when you look at the different topics, how often the different topics arise in different subpopulations? And so what we found is that, for example, white patients have more topics that are enriched for anxiety and chronic pain, whereas black, Hispanic, and Asian patients had higher topic enrichment for psychosis. It's interesting.

Male patients had more substance abuse problems. Female patients had more general depression and treatment-resistant depression. So if you want to create a stereotype, men are druggies and women are depressed, according to this data.

What about insurance type? Well, private insurance patients had higher levels of anxiety and depression, and poorer patients or public insurance patients had more problems with substance abuse. Again, another stereotype that you could form. And then you could look at-- that was in the psychiatric population.

In the ICU population, men still have substance abuse problems. Women have more pulmonary disease. And we were speculating on how this relates to sort of known data about underdiagnosis of COPD in women. By race, Asian patients have a lot of discussion of cancer, black patients have a lot of discussion of kidney problems, Hispanics of liver problems, and whites have atrial fibrillation. So again, stereotypes of what's most common in these different groups.

And by insurance type, those with public insurance often have multiple chronic conditions. And so public insurance patients have atrial fibrillation, pacemakers, dialysis. These are indications of chronic heart disease and chronic kidney disease.

And private insurance patients have higher topic enrichment values for fractures. So maybe they're richer, they play more sports and break their arms or something. Lymphoma and aneurysms. Just reporting the data. Just the facts. So these results are actually consistent with lots of analysis that have been done of this kind of data.

Now, what I really wanted to look at was this question of, can we get similar error rates, or how similar are the error rates that we get, and the answer is, not so much. So for example, if you look at the ICU data, we find that the error rates on a zero-one loss metric are much lower for men than they are for women, statistically significantly lower. So we're able to more accurately model male response or male prediction of 30-day readmission than we are-- sorry, of ICU mortality for the ICU than we are for women.

Similarly, we have much tighter ability to predict outcomes for private insurance patients than for public insurance patients with a huge gap in the confidence intervals between them. So this indicates that there is, in fact, a racial bias in the data that we have and in the models that we're building. These are particularly simple models.

In psychiatry, when you look at the comparison for different ethnic populations, you see a fair amount of overlap. One reason we speculate is that we have a lot less data about psychiatric patients than we do about ICU patients. So the models are not going to give us as accurate predictions.

But you still see, for example, a statistically significant difference between blacks and whites and other races, although there's a lot of overlap here. Again, between males and females, we get fewer errors in making predictions for males, but there is not a 95% confidence separation between them. And for private versus public insurance, we do see that separation where for some reason, in fact, we're able to make better predictions for the people on Medicare than we are-- or Medicaid than we are for patients in private insurance. So just to wrap that up, this is not a solution to the problem, but it's an examination of the problem. And this *Journal of Ethics* considered it interesting enough to publish just a couple of months ago.

The last thing I want to talk about is some work of Willie's, so I'm taking the risk of speaking before the people who actually did the work here and embarrassing myself. So this is modeling mistrust in end-of-life care, and it's based on Willie's master's thesis and on some papers that came as a result of that. So here's the interesting data.

If you look at African-American patients, and these are patients in the MIMIC data set, what you find is that for mechanical ventilation, blacks are on mechanical ventilation a lot longer than whites on average, and there's a pretty decent separation at the P equal 0.05 level, so 1/2% level between those two populations. So there's something going on where black patients are kept on mechanical ventilation longer than white patients. Now, of course, we don't know exactly why. We don't know whether it's because there is a physiological difference, or because it has something to do with their insurance, or because God knows. It could be any of a lot of different factors, but that's the case.

The eICU data set we've mentioned, it's a larger, but less detailed data set, also of intensive care patients, that was donated to Roger Marks' Lab by Phillips Corporation. And there, we see, again, a separation of mechanical ventilation duration roughly comparable to what we saw in the MIMIC data set. So these are consistent with each other. On the other hand, if you look at the use of vasopressors, blacks versus whites, at the P equal 0.12 level, you say, well, there's a little bit of evidence, but not strong enough to reach any conclusions. Or in the eICU data, P equal 0.42 is clearly quite insignificant, so we're not making any claims there.

So the question that Willie was asking, which I think is a really good question, is, could this difference be due not to physiological differences or even these sort of socioeconomic or social differences, but to a difference in the degree of trust between the patient and their doctors? It's an interesting idea. And of course, I wouldn't be telling you about this if the answer were no.

And so the approach that he took was to look for cases where there's clearly mistrust. So there are red flags if you read the notes. For example, if a patient leaves the hospital against medical advice, that is a pretty good indication that they don't trust the medical system.

If the family-- if the person dies and the family refuses to allow them to do an autopsy, this is another indication that maybe they don't trust the medical system. So there are these sort of red letter indicators of mistrust. For example, patient refused to sign ICU consent and expressed wishes to be do not resuscitate, do not intubate, seemingly very frustrated and mistrusting of the health care system, also with a history of poor medication compliance and follow-up. So that's a pretty clear indication. And you can build a relatively simple extraction or interpretation model that identifies those clear cases.

This is what I was saying about autopsies. So the problem, of course, is that not every patient

has such an obvious label. In fact, most of them don't. And so Willie's idea was, can we learn a model from these obvious examples and then apply them to the less obvious examples in order to get a kind of a bronze standard or remote supervision notion of a larger population that has a tendency to be mistrustful according to our model without having as explicit a clear case of mistrust, as in those examples.

And so if you look at chart events in MIMIC, for example, you discover that associated with those cases of obvious mistrust are features like the person was in restraints. They were literally locked down to their bed because the nurses were afraid they would get up and do something bad. Not necessarily like attack a nurse, but more like fall out of bed or go wandering off the floor or something like that.

If a person is in pain, that correlated with these mistrust measures as well. And conversely, if you saw that somebody had their hair washed or that there was a discussion of their status and comfort, then they were probably less likely to be mistrustful of the system. And so the approach that Willie took was to say, well, let's code these 620 binary indicators of trust and build a logistic regression model to the labeled examples and then apply it to the unlabeled examples of people for whom we don't have such a clear indication, and this gives us another population of people who are likely to be mistrustful and therefore, enough people that we can do further analysis on it.

So if you look at the mistrust metrics, you have things like if the patient is agitated on some agitation scale, they're more likely to be mistrustful. If, conversely, they're alert, they're less likely to be mistrustful. So that means they're in some better mental shape. If they're not in pain, they're less likely to be mistrustful, et cetera. And if the patient was restrained, then trustful patients have no pain, or they have a spokesperson who is their health care proxy, or there is a lot of family communication, but conversely, if restraints had to be reapplied, or if there are various other factors, then they're more likely to be mistrustful.

So if you look at that prediction, what you find is that for both predicting the use of mechanical ventilation and vasopressors, the disparity between a population of black and white patients is actually less significant than the disparity between a population of high trust and low trust patients. So what this suggests is that the fundamental feature here that may be leading to that difference is, in fact, not race, but is something that correlates with race because blacks are more likely to be distrustful of the medical system than whites. Now, why might that be? What do you know about history?

I mean, you took the city training course that had you read the Belmont Report talking about things like the Tuskegee experiment. I'm sure that leaves a significant impression in people's minds about how the health care system is going to treat people of their race. I'm Jewish. My mother barely lived through Auschwitz, and so I understand some of the strong family feelings that happened as a result of some of these historical events. And there were medical people doing experiments on prisoners in the concentration camps as well, so I would expect that people in my status might also have similar issues of mistrust.

Now, it turns out, you might ask, well, is mistrust, in fact, just a proxy for severity? Are sicker people simply more mistrustful, and is what we're seeing just a reflection of the fact that they're sicker? And the answer seems to be, not so much.

So if you look at these severity scores like OASIS and SAPS and look at their correlation with noncompliance in autopsy, those are pretty low correlation values, so they're not explanatory of this phenomenon. And then in the population, you see that, again, there is a significant difference in sentiment expressed in the notes between black and white patients. The autopsy derived mistrust metrics don't show a strong relationship, a strong difference between them, but the noncompliance derived mistrust metrics do.

So I'm out of time. I'll just leave you with a final word. There is a lot more work that needs to be done in this area, and it's a very rich area both for technical work and for trying to understand what the desiderata are and how to match them to the technical capabilities.

There are these various conferences. One of the people active in this area, one of the pairs of people, Mike Kearns and Aaron Roth at Penn are coming out with a book called *The Ethical Algorithm,* which is coming out this fall. It's a popular pressbook. I've not read it, but it looks like it should be quite interesting.

And then we're starting to see whole classes in fairness popping up at different universities. University of Pennsylvania has the science of Data ethics, and I've mentioned already this fairness in machine learning class at Berkeley. This is, in fact, one of the topics we've talked about.

I'm on a committee that is planning the activities of the new Schwarzman College of Computing, and this notion of infusing ideas about fairness and ethics into the technical curriculum is one of the things that we've been discussing. The college obviously hasn't

started yet, so we don't have anything other than this lecture and a few other things like that in the works, but the plan is there to expand more in this area.