[SIDE CONVERSATIONS]

JENNIFER GUNN: Hi, welcome to the Institute for Advanced Study. I'm Jennifer Gunn. I'm the director of the Institute. And this is the first IAS Thursday of the spring semester. And so being the first IAS Thursday, I hop you will bear with us for a few announcements.

So the first one is, coffee and water are here. If you would like to sign up for the IAS Listserv, we do not inundate you. The signup sheet is outside on the table in the hall. Today our speaker will be Sorelle Friedler. But before I introduce her, I want to tell you about a few other activities and deadlines.

So, tomorrow morning, in connection with today's event, we're going to have a workshop called "Accountability to Whom?" And it will be a panel of people who work with different kinds of data and technology, talking about the kinds of social and ethical questions that are raised for both the creators and users or subjects of these technologies. And the panel, obviously, you can see the panelists listed here. This will be tomorrow morning from 9:00 to 11:00 in Walter Library, up on the fourth floor, room 401-402. And the first part will be a panel, but the second part will be roundtable discussions with the audience. So we encourage you to come tomorrow morning for that as well.

So, a couple of announcements-- first are about deadlines. The applications for research and creative collaboratives are due tomorrow at noon. This is because, if there is a technical glitch, Susannah can fix it for you before the end of the day, so nobody needs to panic. The second is that the University of Minnesota Social Science Research Council Dissertation Proposal Development Program is soliciting both student applicants, grad student applicants for next year, as well as faculty facilitators.

So the application for faculty facilitators is due on Monday the 29th. If you are in danger of missing that deadline because you're just now hearing about it, just shoot us an email and let us know. However, the deadline for grad students is set in stone.

[LAUGHTER]

That will be Friday February 9th. If you have any questions, contact us at the IAS. And you can find more information about these programs and the links to the actual online applications forms on the IAS website.

All right, next week's IAS Thursday will not be in this room. It won't be in this building, and it won't be on this campus.

[LAUGHTER]

The NFL will be on this campus. We will be on the Saint Paul campus, hosting a panel with the Institute on the Environment on resilience. That will be same time, 3:30 PM on Thursday, next Thursday, February 1, but on the Saint Paul campus at the Institute on the Environment.

And then-- so because of the NFL, you might -- we just want to inform you of a few things. You might have seen all that construction out there on the plaza. There is no access next week to Northrop unless you submitted yourself to an FBI background check in December. So the IAS will be open. The best way to reach us is email or phone. And don't plan to come to this building for any purpose. Any questions?

All right. Well, today's event is cosponsored by the University of Minnesota Library's DASH program, Digital Arts, Sciences, and Humanities; the University of Minnesota Informatics Institute; the School of Social Work; the Institute for Research and Statistics and its Applications; and the departments of computer science, and electrical engineering, and computer engineering. I'm going to turn the program over to Ben Wiggins, who is the director of the DASH program. And he'll introduce Sorelle Friedler, our speaker for today.

BEN WIGGINS: Thanks, Jennifer. So, Sorelle is an assistant professor of computer science at Haverford College and an affiliate at the Data and Society Research Institute. Her research focuses on design and analysis of algorithms, computational geometry, data mining, and machine learning, the application of such algorithms to interdisciplinary data as well. Sorelle is the program committee co-chair for the Conference on Fairness, Accountability, and Transparency, and is one of organizers for the workshop on fairness, accountability, and transparency in machine learning.

Her work on preventing discrimination in machine learning has been featured in IEEE Spectrum-- or "i-eee"-- [LAUGHTER] *Gizmodo,* and NBC News. And she has been interviewed by The Guardian, Bloomberg, and NPR. Sorelle is the recipient of two NSF grants to apply data mining techniques to material chemistry data to speed up materials discovery. And her work on this topic was featured on the cover of *Nature* and was covered by *The Wall Street Journal* and *Scientific American.*

Before Haverford, Sorelle was a software engineer at Alphabet when it was Google-- OK. And she worked in the X Lab there in search infrastructure. She was telling me today that she helped make the blue dot more accurate-- the blue dots on your phones. Sorelle has a name for that.

[LAUGHTER]

So she also received a PhD in computer science in 2010 and an MS in nuclear science in 2007 from-- both from the University of Maryland-- and is a 2004 graduate of Swarthmore College.

[APPLAUSE]

[SIDE CONVERSATION]

SORELLE FRIEDLER: OK. OK, OK, great. All right, so, yes, I'm Sorelle. And so, I'm a computer scientist. And normally I would give sort of a technical computer science talk. But I thought that with this audience it might be more fun to give more of an overview-of-the-field talk. So I'll be talking a little bit about my work, but also about the work of others, who I will try to remember to give appropriate credit to. It's certainly on the slide, whether I say it or not.

And so I wanted to talk to you today about a new and I think really exciting field that's springing up at the intersection of computer science and sociology, and law, and policy, and lots of other fields I'm going to not name. And I thought I would do that by taking a case study that's been very talked about within this newly growing field and talking to you a little bit about some of the work that's already come out of this, even though, if you can see at the bottom there, this is an article from May of 2016. There has already been a huge growth of work on thought that's come after this that, I think, nicely covers the areas of fairness, accountability, and transparency.

So, I saw some of you nodding when I mentioned this slide. So maybe many of you have already seen this piece from ProPublica in May of 2016, where they looked at-- they looked at a risk assessment algorithm, or really the data generated by a risk assessment algorithm that they received through very painful FOIA requests and legal battles. And this is for data from Broward County in Florida. And it includes data about individuals who were arrested and then there are resulting scores based on this risk assessment algorithm.

And so just very quickly, what do I mean when I'm talking about a risk assessment algorithm? What's going on here? So, these are algorithms that are used to try to determine, try to predict whether someone is likely to appear back in court if you grant their bail, or whether someone is likely to recommit a crime if you let them out of prison. And they're used to help to determine sentences. So if somebody is given a risk score, and that's supposed to be advisory to a judge. And then the judge may use that to inform their decision.

Sometimes they'll see these for parole decisions. I'll mostly be talking about them in the context of predicting recidivism for sentencing. All of this comes with the caveat that I'm not a criminologist and I'm not a sociologist. So I'm sort of a computer scientist trying to understand the social context enough to do a-- so, be forgiving.

So, just so that you have an idea of what these things look like a little bit from end to end, this is a copy the ProPublica got of one of the-- this is just the first page of a long survey that they would have somebody fill out when they were arrested. So, it asks questions like, what are the current charges; how many charges have there been previously; how many are pending; are there any juvenile arrests; all these types of questions that go in.

This composes with the data that's going to be used as the input and just some algorithm. And then this is what the algorithm spits out. Again, it's just the first page of a much longer readout. The important part is this stuff on the top in red. That's the overall risk potential. This is a poor visualization, because red, in this case, would not actually be bad. It's just red.

And you can see on the bottom there, it goes from one to 10, where 10 is supposed to be, this person is really scary. And if it's in that top row-- see where it says "violence?" If that was a 10, then you were supposed to be really worried that they were going to recommit a violent crime as opposed to recidivism in general, as opposed to failure to appear if you let them go, and so on. And so this particular person had a low risk of violence and a higher risk of failure to appear.

And then the readout gives you a little bit more detail. So then this gives-- and there's a probable chance that there's a substance abuse issue. And maybe, if you were looking at this, you would say, OK, well perhaps that explains the failure to appear.

So this particular risk assessment comes from Northpointe's COMPAS algorithm, and it's proprietary. So this is most of what we know about it, in addition to a dataset that ProPublica collected that includes lots of individuals, not their answers to these survey questions, but information about them that ProPublica was able to independently gather, along with the output of the algorithm on this. And so from that, ProPublica looked in-depth at this dataset and did some statistical analyses of it and also generated some explanatory stories out of this.

So I wanted to start with just an example of what this might look like. So here are two people who were part of this dataset that they collected and their associated risk scores when they were arrested for both-- in both of theses cases, it was for a petty theft arrest. And so the man on the left was given a low risk, specifically a score of three. And the woman on the right was given a higher risk, specifically a score of eight.

And part of what ProPublica was highlighting here is that actually the scores were wrong in both of these cases. Right? So in both of these cases, we're looking at a case where the algorithm made a prediction and was wrong. In the case of the man on the left, the algorithm predicted that he was low risk. But he did go on to re-offend. He had one subsequent grand theft in the two years following. There was an arrest for grand theft.

And then the woman on the right was labeled a high risk, but did not go on to re-offend. So again, the algorithm predicted she would re-offend and was wrong. So she had no subsequent offenses in the two years that ProPublica looked at the data for.

And, in fact, what they found in their analysis-- so if you haven't read the story, it's a beautiful article. You should go read the whole thing. They were a finalist for the Pulitzer Prize for this work and others. But I would summarize the important statistical fact from article of being this chart.

So what you'll see here is two rows. And both of these represent cases where the algorithm is wrong. So in the first case, we have that somebody was labeled a high risk but did not re-offend. Right? So we would call this a false positive, we falsely thought that they would re-offend. And in the bottom one, we have that they were labeled a low risk, but the did re-offend. This would be a false negative.

MALE SPEAKER: So I had a score of eight, another person has a score of nine. I had a degree in statistics; I see that as a probabilistic statement of my chances to re-offend. I mean, I'd see it as a 10% greater chance. I do not see it as a 100% greater. What I'm hearing from you is it failed to predict, as if there is some god throwing dice, and you are accurate as to the prediction of that. Would you say that that's how this company sold their data?

SORELLE FRIEDLER: So there's a whole bunch of things that you're bringing up there, right?

MALE SPEAKER: Yes.

SORELLE FRIEDLER: So one of them is that I am sort of lumping them together. They have risk score between one and 10. I'm calling them high-risk and low-risk. That was the way that ProPublica did the analysis in order to, I think, more easily summarize thing for a broad audience, right? So I don't remember exactly what threshold they used in terms of the actual score to categorize the low-risk/high-risk. I think they took everybody who was both medium and high and called that high-risk, and took people who were low and called that low. And I think the threshold for that was either three or four. But go through their details to find that one.

So that's one thing. The other thing that you're asking is, I think, a potentially more complicated social question. Because part of what you're asking is really how do the judges interpret these scores? Right? And there's been a whole lot of really interesting research done into exactly how they interpret the scores. And the short answer is that they do not interpret them the way that a statistician would interpret them.

MALE SPEAKER: [INAUDIBLE]

[LAUGHTER]

MALE SPEAKER: But that's immaterial. You can't fix their idiocy.

SORELLE FRIEDLER: So, yeah, so, I'm going to be analyzing the tool, not totally independent of that social context, but focusing on the tool and whether we can understand it and whether it's a good tool. But, yes, there is then a whole other layer of how that tool is then used, which may or may not match up with its actual purpose.

Right, so I guess the thing that I was mid-pointing out is that we have false positives and false negatives. And when they broke that down by race, they found that black defendants were far more likely to be labeled high-risk, but not re-offend, right? Whereas white defendants were far more likely to be labeled low-risk and then re-offend, right?

And if you were going up before this algorithm, what you would want is to be preferentially labeled low-risk, right? So white defendants were getting the benefit of the doubt, and the black defendants were not. And so ProPublica wrote this article. There was a large amount of press accumulated around this. And people were saying, this is really fundamentally unfair. Look at what's going on here. This is bad.

And because-- much to their credit-- because they released all of the data despite the long FOIA struggles that they had to go through to get it, they released it all publicly. If you want to, you can still go and find it at these various links. Because of that, people have been able to go back in and sort of reanalyze and rethink about all of the different aspects of this. So again, I'm using this, I think, pretty compelling examples as a summary of the field.

OK, so one of the first questions that people asked was, are these decision fair? And a lot of the backlash was COMPAS is racist. OK, but what does that mean? And what would it mean for an algorithm to be fair? And what would that mean mathematically?

And the reason why that's an important question is because if we want to build algorithms that are fair, we have to do that to a specific definition, to a specific mathematical definition, which is, again, taking a complicated social context and trying to boil it down to an equation. So how do we even begin to do that? And again here, there's a caveat that people have been trying to understand what fairness, and justice, and equality are in society for a long time.

Philosophers have written large tomes on this. Now we're going, and as computer scientists, we're trying to come up with some small nugget that well encapsulates something. It in no way is appropriately capturing a human idea of fairness. But it's at least trying to capture a small piece of it.

So for some of that work, we've gone back to the legal literature and tried to say, OK, if we're looking at questions of discrimination, perhaps the law has something to tell us about how we should be measuring fairness. So there's a case from the US Supreme Court in 1971, *Griggs versus Duke Power.* And the core summary of the case is that there were workers for Duke Power, who were not being hired because they were black. And Duke Power was trying to hide that by saying, well it's not that we don't hire black workers. It's that we require a high school diploma in order to be hired.

But at that time in that place, that was essentially a proxy for race. And so this case was brought. And the US Supreme Court indeed said, what matters is the ultimate outcome. You can't hide your discrimination by using some proxy variable. You need to actually look at the resulting outcome. They referenced this fable, Aesop's fable about the stork and the fox, in which the stork invites the fox over to have some milk and serves it to the fox in that vase, I guess, chalice-- I don't know-- which, of course, the fox can't drink out of. And then the fox retaliates by inviting the stork over.

But the point there was, but the outcome is what's important. Even if something is facially neutral, it's not directly using race, you can tell if it's discriminatory by looking at what the outcome is. And in fact, the EEOC, the Equal Employment Opportunity Commission, then went back and codified that into a mathematical test. And that test is sometimes called the Four-Fifths Rule or the 80% Rule. It's saying that if you have members of a protected class-- so this is something like race or sex-- and they're selected at rate less than 4/5 of the majority group, then you consider that to be discriminatory.

And framing that into mathematical language, so I've looked at this from a computer science and machine learning perspective, that definition ends up looking like this. So for people who are more used to seeing it on probabilities, that's what this looks like. So what you're saying is that the probability that you get a yes, so that you're hired, given that you're a member of the minority group, divided by the probability that you get a yes given that you're a member of the majority group, if that's less than or equal to 0.8 or 4/5, then you consider that discriminatory.

And the implication here is that you would like that ratio to be one. You would like people to be hired at the same rate. But you allow a little bit of wiggle room. Maybe it's a small company, and it would be silly to cut people in half and accept half of them and not some other, and so on. So there's a little bit of wiggle room in that.

OK, so that's one definition of fairness or discrimination that comes out of the literature. There are others that have a similar flavor to this that I also wanted to show you. So some come specifically out of work that was studying recidivism predictions, specifically on the COMPAS data. So here there's actually three definitions. And they're sort of variations on each. So I wanted to show you all of them together.

So the first one-- in this case, the why is actually the outcome of recidivating. So you'll notice that in that previous definition-- there we go-- in this definition, nowhere do we have anything about the actual outcome of whether this person reidivated or not. So if we're thinking about this one in terms of recidivism prediction, we would think about the C equals Yes as being the algorithm predicting that they would recidivate or wouldn't recidivate. Whereas, here they're looking at Y equal 1 is that they actually did recidivate.

And we want the likelihood of-- we want the probability that they did recidivate to be the same if they actually were given the same score regardless of race. Essentially what's that saying is we want the score to mean something. So that's called calibration from this paper.

And then predictive parody is a similar thing, but now it's looking at the score as a threshold. So it's sort of doing that grouping of, OK, high risk, so everybody above some cuttoff. And then this last one is actually sort of two definitions. So this is, they call it error rate bounce. And this is essentially saying that we want the false positive and false negative rates to be the same across groups.

So this one is actually starting to get at the thing that ProPublica found that was unfair. So this is identifying that and saying, we don't want that type of outcome. We don't want that type of outcome, where white people are far more likely to be given the benefit of the doubt and labeled low risk. This definition would rule that out.

So [INAUDIBLE] looked at the COMPAS data across each of these definitions. So again this-- I don't know if you guys can read the code up there. But the grayish-bluish bars represent black population. And the orange-ish bars represent the white population. And she looked at this. So this is a COMPAS score, and that's the observed probability of recidivism based on this dataset.

And essentially what this is showing is that this is well calibrated. All right, so I don't know if you guys can see the black vertical bars. But those are our error bars. And you'll see that they largely overlap. So there's actually not much difference between races. And so based on this data that ProPublica collected about the COMPAS algorithm, it looks like the COMPAS algorithm is well calibrated. So it's fair under these two definitions.

OK, well, so what about the other definition? Right, what about what ProPublica found? Well, indeed, if you look at the error rates, you'll see that, for the black population, the false positive rates are significantly higher than the white population. And the revers is true for the false negative rates. The white population has many more false negatives than the black population. So largely this is confirming what ProPublica found. But now we've managed to frame it a little bit more mathematically and a little bit more precisely defined what we mean by fairness.

So concurrent work by Jon Kleinberg, et al came up with a similar three definitions of fairness. There are some slight variations, but we can largely ignore them for these purposes. So the first one is, again, calibration. And then the next two are the false positive and false negative rates. And what Kleinberg et al and also [INAUDIBLE] were starting to notice and showed in both of their papers is that while we feel like all three of these definitions are important measures of fairness, we can't actually have all three of them at the same time.

So more specifically-- and this is from Kleinberg's paper-- more specifically, when can we have all three of them? It's only in the cases where we have perfect prediction, which in a real world scenario we sort of have to rule out. Or in the case where we have equal base rates, where the two populations are equally likely to recidivate. So, otherwise we're going to have to put that extra somewhere. If the base rates aren't equal, then there's something unequal about the data that we have. And so we can't just satisfy all of these nice equal definitions. And I want to highlight, again, this is in the case where we don't have equal base rates, which is the case in the ProPublica COMPAS data. The base rates there are not equal.

So in trying to sort through a lot of these questions, along with some colleagues, Carlos Scheidegger and Suresh Venkatasubramanian at Arizona and Utah, one of the things that we've been trying to figure out is, OK, in a lot of these conversations, it seems like there are some assumptions that people are making about the world. So people say, OK, I want my fairness definition to be the following. And what they're really saying, I think, is the world looks like this, and so naturally what fairness means in that context is the following.

And so in trying to understand that for ourselves, what we came to is that we think that there is two different world views that people are coming from. One world view, which we call "what you see is what you get," is the world view where you have data, and you believe that the data is perfect. You believe that the data is accurate. And so if the data shows something, then than means it's true.

And the other world view is the world view where you believe that there is structural bias. And so you're inherently skeptical about the data because you think, well, there's reality and then I have measured the data, but those measurements are subject to the conditions on the ground. And so the measurements are flawed. And I think that they're flawed in a structural way. So I think that they're flawed in a way that may be racially inflected, for example.

So in the second world view, in this belief of structure bias world view, one of the things that we think we're seeing people assume is that, on the whole, we're all equal. And by that I mean that if you look at the distribution conditioned on somebody's group membership, you would expect those distributions to look the same across groups. So you would expect, for example, white people would naturally recidivate at the same rate as black people. You would sort of hold that as an inherent assumption, and then you would move with them. And so part of what we're trying to do with this work is try to encourage people to state their assumptions. Because then we can go forward and argue about them or not, but at least we know where the ground is.

This-- I guess I got a head of my slides. So we would like to see people state their assumptions. And I think in terms of these impossibility results from Kleinberg and [INAUDIBLE], I think that there is potentially an interesting conversation there because one of the things that I haven't yet explained about this ProPublica COMPAS data is the way that they indicated recidivism in this data. So when we talk about recidivism, we talk about somebody's likelihood to recommit crime. But when it's measured in the data, it's often measured based on rearrest within two years.

And these two things are not necessarily the same. The likelihood that someone is rearrested may have a lot to do with what neighborhood they find themselves walking around in, what the police think of people who look like them, and so on. And that may or may not be correlated w with the actual likelihood to recommit a crime. And so I guess I would argue that it's important to state your assumptions about exactly how you think those things are related and what you actually think the underlying world of actions looks like. As part of that, one of the really important things that comes up in a lot these algorithms is actually how are they making their decisions.

So far we've talked about outcomes. I think that looking at fairness from the outcomes focus point of view is very useful. That's sort of the disparate impact legally inflected view of discrimination in data. But there is also a desire, I think, for people to be able to examine the algorithm and say, OK, but how did it make its decision? And then I can reason from there about whether I think that was fair in the context. And in fact, there was a post-- post this ProPublica COMPAS thing, there was a case in the Wisconsin supreme court, where in fact somebody sued and said I don't have right to face my accuser. I got this bad COMPAS score, and I should get to know how that came about.

And so that went to the supreme court of Wisconsin. And there it was decided that the judges were allowed to continue to use the COMPAS scores to inform their decisions and that Northpointe was allowed to keep its algorithm proprietary, but that now each time that there is a risk assessment shown to a judge, is should come with a warning. And the warning-- so they spell out in the decision the four points that the judge should be warned about.

Personally, I think that goes to some of the questions you were just raising about to what extent can we expect judges to actually understand the sophisticated warning. But I think the important thing here is really this last sentence, which says that the goal is to try to give the judge enough context so that they can understand how to appropriately read the risk score in their decision.

And I think that that sort of leads to a bunch of other interesting work. So one of the recent things-- this is from January 2017, so really, really recent. You may have seen this. So what these researchers did was they gave a bunch of humans this paragraph. So they collected a much smaller amount of data than COMPAS collects. So COMPAS has that long survey thing. I just showed you the first page of. I goes on from there. They collect a bunch of data.

In this case, they only collected sex; age; crime charge; criminal degree, meaning felony, misdemeanor; and then the counts of their juvenile record. They showed these to mechanical turk workers in this paragraph. And then were asked, do you think this person is going to recidivate. And the turk worker read the paragraph and just guessed. And then they pooled those responses and found that that pooling did just about as well as COMPAS did in predicting the results, which I think is fascinating when you think about explaining it to a judge. Because I think this is actually something that you could explain to a judge. It just might take a-- right? That potentially-- how did the headline put it? The algorithm is no better at predicting crimes than these random people, when given this clear paragraph with these particular features. I think the choice of feature there is important.

MALE SPEAKER: Right. They are predicting it with nearly 2/3 accuracy for predicting recidivate results. Certainly when I read the headline, I get the implied feeling that it's a 50/50 call. If I could make my stock predictions 76% accurate, I'd own this university. I don't think this gives the real feel. If you could predict things at 76%-- they're at 67% accuracy. Excuse me-- I think you'd be pretty, pretty, pretty happy.

SORELLE FRIEDLER: I think we could quibble about that.

MALE SPEAKER: OK.

SORELLE FRIEDLER: I think, from my perspective, the important thing here is the comparison. The important thing here is explaining what the algorithm is or is not doing by the comparison to a random person deciding this, which I think gives some explanatory power that the judge can understand.

FEMALE SPEAKER: But it's not a-- personally it's a small part of it. I think that's important, that you see the-- a little bit better than just random. And the [INAUDIBLE] to come up with a good decision. I mean, still the algorithm, that's a different issue. But they're not going to be better than that. It's an issue.

MALE SPEAKER: I am curious as to how wrong they were. It's one thing to predict the recidivism. It's another thing to fail to predict. And there is value to that because that's your other side of the equation to really determine. I mean, just because some company calls this an algorithm-- I have an algorithm when I play Blackjack. It's not a very good one because they take my money, but it certainly is defined mathematically as an algorithm.

SORELLE FRIEDLER: Well, and I think that this is part of the point, is that judges and people at large, who are not familiar with algorithms can sometimes imbue them with more power than they actually have.

MALE SPEAKER: OK, I see. You can not fix that.

SORELLE FRIEDLER: OK, so this is-- I'm claiming this is one way to try to explain a little bit whether an algorithm is doing well or not. I still find this somewhat unsatisfying. I think we would still like to know in more depth what is the algorithm doing, right? OK, so there exists interpretable [INAUDIBLE] models. This is a picture of one of them. So this is a decision tree, also known as a flow chart. You have, presumably, played with them in the back of magazines, where you get a little question, and it says yes or no. And then you go to this side, and it says yes or no. And then eventually you get some answer that tells you-- I don't know-- something. Whatever.

You can automatically generate those decision trees. And that technology has been around for a while. The issue with it is that these decision trees can be very big, right? So this is not actually even the entire decision tree. This is just one branch of a decision tree that I worked with when I was working with some chemists to try to predict the outcomes of chemical experiments. And the chemist said, well, you know that's just very cool that you can beat the human in predicting these outcomes. But we care about actual hypotheses. We want to know how you did it so that we can learn something from it.

And because my colleagues were incredibly patient, when I printed out this giant decision tree for them, they went and actually read it and tried to reason about what was going on. And the highlighted portions are the chemical hypotheses that they came up with out of this. Still though, I would argue, maybe you don't want to have to present this to a judge.

[LAUGHTER]

My colleagues are wonderful, and patient, and cared a lot about this. But in a high-stakes decision that's made in a small amount of time, this might not be your best approach, even though it is still nice to be able to examine.

There's been some nice work done by a Cynthia Rudin at Duke along with some of her colleagues. And she's been doing a large body of work. I'm just going to show you briefly one of the papers on creating new algorithms that are naturally interpretable. And she has a stricter definition of interpretability that would rule out that disgusting decision tree I showed you. She would say, no, that's not interpretable. It's just too big.

And so one of the new algorithms that they have created is this rule list type of thing. So they've made an algorithm that can generate this. So this sort of simple algorithm was actually automatically generated from the COMPAS data. So this is what COMPAS could have looked like if it was interpretable and you were actually able to show it to a judge.

And I actually think that they've done pretty well in terms of making this interpretable. We can all take a minute to look at it now and hopefully understand what is going on and what type of prediction we would get if we were faced with this algorithm. And I would also argue that potentially a defendant could look at this and understand, OK, I got this prediction because I am between 21 and 22 and have two priors. That's pretty simple.

And in fact, this one was also about as accurate as the COMPAS predictions. And the argument that they make in this paper is that perhaps it would be better for COMPAS to be using something like this and allow the people to examine it, especially given that this can do just as well.

FEMALE SPEAKER: Was it as accurate as the COMPAS predictions in terms of race? I mean, did it fall out the same way? Because that's the big issue with COMPAS.

SORELLE FRIEDLER: That's a really important question that I don't know that they examined. It might be that I'm forgetting it. So I'd want to go back and look at the paper. But I don't think I saw it in there.

FEMALE SPEAKER: Are you saying this is cool?

[LAUGHTER]

Sex equals male, then go to jail?

SORELLE FRIEDLER: If you're 18 to 20, and you've been arrested once, yeah.

[LAUGHTER]

FEMALE SPEAKER: Do they say something about that? Or do they--

SORELLE FRIEDLER: Well, so, my point here is really that you can now see it and make any arguments you want to about it.

FEMALE SPEAKER: May I make one? Why are we analyzing recidivism? What does that have to-- our future prediction of whether someone is going to commit a crime, what does that have to do with how long you should lock them up for? The question we need to know is how long should we lock someone up for to prevent them from recidivating. And this doesn't answer it.

SORELLE FRIEDLER: And I totally agree. I think that's a great point. And my point here is that in order to have a more nuanced argument about what's going on, we need to be able to know something about what's going on in these algorithms. So you're actually pointing also to something that's not represented here, which is that we also need to have a careful discussion about what the outcomes are that we are predicting.

So you're saying, OK, I now understand how you are predicting this, but I still think you're predicting the wrong thing.

FEMALE SPEAKER: You're predicting. It's not the question that justice requires. We need to talk about what the purposes of sentencing are before we can answer what is fair to consider.

SORELLE FRIEDLER: And my concern, as a computer scientist, is that the tools that exist need to be intepretable enough so that that conversation can be had outside of the computer science environment by people who understand the context and can have that discussion.

JENNIFER GUNN: So, why don't we let Sorelle finish this. Then we'll have time for a Q&A. But we want her to get to the end of her presentation, too.

[LAUGHTER]

SORELLE FRIEDLER: OK, so despite its problems, I actually see this as hopeful because I think it's fairly clear. I think we can understand it. There's still a problem though, which is that that algorithm that I just showed you is not COMPAS. That's still proprietary. So we still don't actually know how it made its decisions. We know that we can do just as well as it did in making decisions. But we don't actually know it made its decisions.

So how can we try to figure that out? So unfortunately, I've-- just a little foreshadowing-- the techniques in this area, I think are less ultimately satisfying. I think there's still a lot more work to do. But what we can do so far is try to use that algorithm as a black box. So these techniques sort of assume that we have repeated access to ask it queries.

So I can feed COMPAS a defendant, maybe a made up defendant that I've created in a specific way so that I can then find something out based on that answer. And I can do that repeatedly so that I can get some sense of how it might be operating with different people. So that's exactly what this paper does.

So they take the original data, and then they randomize one of the columns, one of the features in this data. And then they send the data back in. And then they discover how much the predictions changed.

And so think about it. All right, if you randomized a bunch of information and the predictions don't change, you aren't really using that information. On the other hand, if you randomize a bunch of information and the predictions change drastically, then you were really using that information. And in fact, you can measure the amount that they changed. And you can use that as for the quantified amount of importance for each feature being used directly by the model. So this allows us to poke at it and say, OK, I don't know how you're using race. I don't, for example, know if it's positively or negatively. But I do know that you use it.

Using a very similar technique, another paper then goes on to the create graphical readouts of exactly what's going on for the model as a whole. So this is also on a recidivism dataset, though not actually on the COMPAS dataset. And so you can see that on the top graph, drug history is being heavily used. That's that bar all the way on the left. And the bar all the way on the right is census region. It's being used a little. And race is somewhere in between.

And then they are also able to, again using this modification technique, create a readout for an individual. And, OK, here's what the model does overall. Here's how I think it operated just on you, if I just change some of these features and send it back in. And for this person, Mr. Z, race is being used fairly heavily. That's that bar all the way on the left. They've applied heavy. OK, so that let's us poke it and figure out, OK, is there-- we can sort of imagination it as is there some box in this giant decision tree COMPAS is using that says, if race, then do this?

There is another technique that I've worked on, where you can actually tread here a little bit more into it by taking the original outcomes that were the original predictions and then trying to build a model to predict those. So I think of it as modeling the model. And that is actually I generated this. So the original model that we were using for the chemistry data was a more complex model called an SVM that is not human interpretable.

But when my colleagues wanted to understand what was going on, I said, well, OK, I could try to model what that model's precautions were, and then we could look at that. And it wouldn't be exactly the same. They would still need to go and test their chemical hypotheses, for example. But at least it would give us some insight into the bigger picture.

But again, it's still not quite poking the black box, right? So another thing that we worked on-- so the first that I was describing, you randomized just a single column. That gives what I refer to as the direct influence of that feature. But what about if you have a case like redlining or other proxy variables, where you're looking at sort of an indirect influence?

So the model itself, say, isn't directly using race. There's no if race equals this do this. But it's using zip code. Do we want to say that race has an influence on the outcome or no? If we do want to say that race has an influence on the outcome in that case-- I think if that is the indirect influence of a variable-- and we can also modify the inputs to a black box algorithm to figure that out.

And-- yeah, I have the slides. And the way that we do that actually comes from that earlier disparate impact paper, where what we do is we take, for each variable, we take the-- so, in my little hypothetical example, that's where we have red distribution is men. And the blue distribution is women. And we take their hypothetical SAT scores, and we map them to that central black distribution. It's a media distribution over a certain measure.

And essentially by doing that, we can't distinguish between the men and the women in this example, because both of those distributions are in the same place. But we haven't collapsed them down to a single point. We haven't just assigned everybody the same SAT score, which is what would happen if we just got rid of that information. So it allows us to preserve some of the distributional information while getting rid of this information that could allow this to be used as a proxy variable.

So by doing that, we can sort of mask out the information from a variable that might appear in small parts in a bunch of other features. And then we can use that same technique of, OK, so I'm removed race from my data. And then I can feed it back into this black box, and I can see how much the results changed from the black box. And if they changed a lot, then my removal-- there was a large indirect influence from race in the model. And if they didn't change that much, then even though that information was there, it wasn't really used in the decision.

And so doing that-- so as you move right on these graphs, we take out more and more of the information. And as you look down, we lose more and more of the accuracy. So the way that you can read this is that these features that are all the way at the bottom are the ones that the model was using the most. These had the most indirect influence on the model.

And again, because we don't actually have access to COMAS, so I couldn't actually run this experiment on COMPAS. This is with the COMPAS data, but with a little model that I had put to model the outcomes. So there's two different models. In these two different models in both cases, the number of juvenile, misdemeanors, felonies, and other infractions has a large influence on the model. Those are the blue-greenish lines that go all the way down.

In this model, age-- well, actually, I guess in both of them age also has an influence. That's the maroon line. And in this model, race also has a fairly significant influence. That's the pinkish line. Whereas, in this one it doesn't. In other words, we'd need to actually have access to COMPAS to play with it to get more information about that.

OK, so that's what's happened since 2016 in terms of thinking about this. And it's actually only a small sample of what's happened since then as the field has sort of exploded in various different ways. There's lots of other definitions for fairness that I didn't have time to get into. There's lots of other different types of interpretable models that give really exciting possibilities for trying to understand what's going on.

All of these things are going to be discussed in more depth as a new conference that's happening at the end of February. The conference is already sold out, but it will be livestreamed. So if any of you are interested, that website will have a livestream of all of all of the events. And they'll also be recorded, so we'll put up the videos for later viewing as well.

And if you like playing with code, you can do that.

[APPLAUSE]

JENNIFER GUNN: OK, so do you want to field your own questions? And let's try to make sure, if we have a lot of questions, that we distribute the opportunities to ask.

SORELLE FRIEDLER: Sure, yes.

FEMALE SPEAKER: Hi, thanks for your thoughts. I had a question about the different ways people are playing on with the data. Do you know of any studies where people are trying to correct for overpolicings of black and brown communities and also the disproportionate drug arrests for black and brown folks a well as disproportionate juvenile counts because of the school-to-prison pipeline, right? So seeing those juvenile arrests as being really heavy proxies for race, right? Has anybody tried to do a model where you try to control for that to help manage that?

Because it seems like the genie is out of the bottle. The courts are all over this. So, just curious.

SORELLE FRIEDLER: Yeah, so, again I would say, I'm not a criminologist. But what I can tell you is that there has been a lot of interesting work on thinking about predictable policing, which I think touches on some of those. So a lot of this work, the work in this are is trying to understand algorithms that were in the field and sort of reason about, reason about them. So some of the work has been in trying to identify the ways that existing predictive policing systems cause feedback groups. So the course idea being that you tell police to go to a neighborhood. And so they go there. So they arrest someone.

They feed that data back into the system. So the systems says, aha, there are people to arrest there. You should go back there tomorrow. And just sort of does that. There is some nice work out of Human Rights Data Analysis Group by Kristian Lum and William Isaac, demonstrating that if you run PredPol, which is a popular predictive policing system used in LA, if you run it on drug data from Oakland, then very quickly it will converge on sending all of the police to a very small number of minority neighborhoods.

And there's some caveats to that work. For example, when you are sending police to a neighborhood, their model assumes that you're only getting data from that neighborhood, when in fact, you might also be getting data from 911 calls. I've done recent work that shows that even if you add in 911 calls, you can theoretically model that you would still expect to see some of the feedback loop if you do certain [INAUDIBLE]. Don't quite answer directly your question, but it give you sort of a sense of the landscape of the work that is going on.

Yeah, in the back.

FEMALE SPEAKER: I just had a quick question. So that survey that you showed us in the very beginning, is the data that they're actually putting into the COMPAS system, is that self-reported? Or who is that coming from? Who is filling this out?

SORELLE FRIEDLER: Does anybody in the audience know this better? I believe that what happens is that there is somebody who interviews someone. So it's like a little bit of both. Right? I don't know if other people know this.

FEMALE SPEAKER: Because I think that, if we think about categorization of the ways-- first of all, the categories that we're using and the ways in which we're interpreting these categories, who is being put into those categories and whatnot, I think that in and of itself is something that ought to be explored, which I don't know if that's something that could be modeled or anything like that. But by version is--

SORELLE FRIEDLER: Yeah, I mean one of the meta questions about this area is the extent to which it relies on categorization.

FEMALE SPEAKER: And self-reporting or not, who is reading what--

SORELLE FRIEDLER: --just sort of inherent in computer science. database systems sort of beg for categorization, but I think that there is real questions about how that should work and what that means. Yes, I'm sort of going back-to-front. Yes, go ahead.

MALE SPEAKER: So a few things I-- a lot of great stuff there. I really appreciate it, and appreciate especially the emphasis on the importance of transparency. And what I take is, in a sense, your attempt to kind of reverse engineer the algorithm, which I think is really important. To a certain degree, it seems that the patent problem is what happens when algorithms meet capital. And the proprietary nature of this in the justice system is, to my mind, what's most fundamentally appalling about the whole thing.

But putting that tirade to the side, what background assumption, in some ways, I just want to throw out there to think about, is the idea that fairness-- it seems implicit here-- that fairness means having the algorithm that doesn't use race. But, as in the legal world-- in the legal world, that would be color blindness. And there's a lot of issues with color blindness. Color blindness is not necessarily always fair.

And so to think about whether in the world of algorithms, is it about erasing race? Or is It about figuring out when and how it's appropriate to take race into account. Perhaps to serve, right-- and it's hard because if it's now focused on recividism, this goes to the question of why are we looking at recividism? There's so much to look at in sentencing decision. But to think about it, maybe we should be taking race into account in sentencing, just not the way it's been taken into account in the last few hundred years.

So I just want to throw that out there that maybe it's not necessarily about trying to control for race. It could be in some cases. But maybe in some cases, considering disproportionate policing and things of that sort, maybe it should just be taken into account differently.

SORELLE FRIEDLER: Yeah, and there's somebody recently was categorizing it and found that in this field, there's now 21 definitions for fairness. So I showed you--

MALE SPEAKER: Of course you're counting.

[LAUGHTER]

SORELLE FRIEDLER: So I showed you four, right? So some of them are starting to try to get to that. I think that some of that is incredibly nuanced and is going to be different per scenario and might be hard to appropriately capture in a formula. Some of the interesting takes that people have include thinking about what they call process fairness, where they surveyed people to find out if they thought that it would be fair to use a particular feature in the resulting decision, and used that to capture whether the decision was fair.

So there are a lot of interesting, more out-of-the-box ideas that are being included. And the person who categorized these 21 definitions of fairness is also making the argument that actually that's good. It's good to have a variety of definitions so that they can be catered to different situations. Yes.

MALE SPEAKER: Yeah, taking the previous question a step further, what are you thoughts on-- this may not be a case in the recidivism model, but just kind of examining, it seems like one of the ways you could correct a lot of that bias would be to actually include race in the model as a variable so that those outcomes would be corrected for it. However, there's a strange social contradiction here, where if you're doing that, I imagine that the COMPAS database takes great pains to avoid including it because of how bad that would look.

How do you respond to questions on when it seems actually appropriate to include what might be viewed as a discriminating variable, when it can actually correct for something?

SORELLE FRIEDLER: Yeah, and so most of-- so there's a large body of work now on algorithms that attempt to correct for bias, so fair, high, and numbered them, that sort of thing. And largely those algorithms require race, or sex, or whatever bias they're correcting for as an input. And then use it to make the correction and not to make the decision. Though, depending on exactly how the algorithm works, it's a fine line about whether you think it's being used to make the decision or not.

So you can think about a machine learning pipeline as having the input, the algorithm, and then the output. And interventions have been made at all three stages. So some algorithms change the input to the algorithm and then use the algorithm alone. Of course, changing the input is itself sort of an algorithm, right? Some algorithms say, instead of using that other algorithm, you should use this one that corrects for bias. And then some say, OK, run the pipeline as it exists so far, and then I'll tack something on at the end and change the results.

And I think legally probably each of these different interventions has slightly different outcomes in terms of what the legal system thinks of it. In terms of what a thinking person thinks of it, I don't think that's going to differ based on context and based on [INAUDIBLE] the algorithm. Yes.

MALE SPEAKER: So there's a lot of focus on the algorithm itself and the case of COMPAS and the proprietary nature. So just hypothetically, if we went to [INAUDIBLE] tomorrow and found the source code for COMPAS, found it to be a beacon of fairness, that would probably lead to looking at the data itself and the bias being more present in the data. And I can think of two ways that bias can feature in data. The algorithms that people use to-- police, et cetera-- and the people entering the data, and also the systems that collect the data could have inherent bias type of funnels in their data collection. And I'm just wondering how to balance algorithm and data when you're searching for fairness.

SORELLE FRIEDLER: So a lot of work on fairness around machine learning algorithms essentially assumes that the data is biased. And so then the fair algorithm's job is to correct for that bias somehow.

MALE SPEAKER: So then what happens when the data becomes less biased and then fair algorithms would actually be unfair if it was correcting for--

SORELLE FRIEDLER: So depending on how it does the correction-- it sort of depends on exactly the mechanisms involved. So if you-- so here-- if you imagine that this separation between men and women is biasing the data and should be corrected, then what we're doing is we're moving those two distributions to the same place. If the data over time became unbiased, and men and women had, on the whole, the same SAT scores, then we would start off by being given that black distribution. And we wouldn't do anything. So in that way, the algorithm would sort of self-correct as the data changed.

This is assuming that you appropriately retrain the model as you get new data, which is, in practice, a thing that is not always done and should be. So there's always implementation details.

MALE SPEAKER: Do you advocate opening source algorithms for something like COMPAS? It was hinted at.

SORELLE FRIEDLER: I mean, personally, I think that would be great. Because then we could examine them and have an open debate about we want justice to look like in our society.

MALE SPEAKER: Reverse engin--

SORELLE FRIEDLER: Right, to reverse engineer it and sort of guess at the extent to which it actually reflects what they're doing. Yeah.

FEMALE SPEAKER: Since your presentation is a lot about giving us a survey of the field, I'm curious if there are other pieces than prediction in criminal justice that have been the focus of a lot of the work in this area. I mean, yeah--

SORELLE FRIEDLER: Yeah, so most of the work in this area is sort of domain agnostic. So any algorithm developed can be applied to any dataset. And so the work, when it's done in context, is done based on available datasets, which is why ProPublica releasing their dataset drove so much work. Because there was a dataset to look at. The thought context for some of the work is also, in terms of hiring, there are now a lot of startups doing automated hiring. There's a whole lot of legal context there that says that you should be doing not discriminatory hiring. So then, you might wonder, OK, if you've automated it, is it fair?

It also comes up in loan decisions and loan scores or credit scores. And I was referencing the predictive policing work earlier, which brings up some interesting new questions because of the repeated nature, whereas because it's not a one-shot decision. I'm probably forgetting some important ones, but those are the big ones that people have been talking about.

FEMALE SPEAKER: Cool. Thank you.

SORELLE FRIEDLER: Yeah. Yes.

MALE SPEAKER: Is there any copyright or patent protection for those algorithms?

SORELLE FRIEDLER: For-- I think that depends. That's like per person who created it, what license they have decided to produce it it under. Most of these people are academics, so it's-- and it's online.

MALE SPEAKER: Well, if COMPAS has a proprietary logarithms, could they protect them, and could they-- is there any way the could protect them by patent or copyright so that they could release them?

SORELLE FRIEDLER: Oh, I see what you're saying. I have not idea. Patent attorney would need to answer that. Yeah, I don't know.

MALE SPEAKER: But do COMPAS and the other that have produced these programs for judges to use, are they constantly feeding back the results of their-- testing the correctness of their projections?

SORELLE FRIEDLER: So, we hope so, but don't actually know. So I've talked to people who have worked with local jurisdictions to try to develop algorithms, and have heard various horror stories about how the algorithm was deployed and then it was never retrained, and then everybody in the city with locations. And things sort of gradually broke. And that is something that will happen if you don't maintain your data and your context for the algorithm.

But we don't, as far as I know, know what is going on with Northpointe and how they're, how often they're retraining it, or also how much it's tailored to a specific jurisdiction. Because you can imagine that things in New York would be very different than LA, would be very different than-- I don't know-- from Nebraska, right? I would imagine that you would want the algorithm to look slightly different in each of those places. And I don't know if it does.

MALE SPEAKER: I just wondered why you're calling your talk "fairness." Really what we want is predictability or accuracy.

SORELLE FRIEDLER: Well, so I think that part of the goal is predictability and accuracy. But part of the goal is non-discrimination, which is no necessarily the same thing. And I think the definitions try to do a lot of work to disentangle these things. Yeah.

MALE SPEAKER: I was wondering, if you did have the ability to query COMPAS any number of times that we wanted, and it was unable to distinguish those queries from distinct cases, would the tools that we have now be able to determine whether or not it was fair or not, looking at the black box?

SORELLE FRIEDLER: What it would be able to determine is the extent to which it used different features. How it used them we wouldn't really know, And whether that was fair or not would depend on how you interpreted fairness.

MALE SPEAKER: So it's our definitions for what is fair are very important and our tools for figuring out if a black box is, with what it uses, are the two important things over here.

SORELLE FRIEDLER: Yes, definitely. There was another question over here.

MALE SPEAKER: Yeah, I was particularly impressed by the incompatibility results of the different formalizations of fairness. I think there's probably a different way of seeing the same sort of incompatibility regarding the framing of the issue in terms of a utility versus a deontic issue. That is to say, whether we should maximize, as we were referring to before, the predictive ability of the algorithm versus whether or not the algorithm implements rights for the groups to which it applies in the appropriate sort of way. So it's nicely captured in the comparability of the type I and type II error rates that you mentioned.

I was wondering if you knew of any sort of work that proposed ways of finding a balance between these two sorts of criteria. Because it's quite plausible, as least to me, that not all cases, if not all cases in which we should have let the balance of type I and type II error override any considerations of total error rate. But on the other other hand, it seems clear to me that we shouldn't let total error rate completely override considerations in balance of type I and type II errors.

SORELLE FRIEDLER: Yeah, so I'm pretty sure that the Kleinberg paper includes sort of a-- basically a knob that you can turn to tune a little bit of that. And I think there's also some more recent work by Cynthia Dwork, her group out of Harvard, also thinking about further impossibility results and what the nuances there are. Yeah, so it's certainly possible to tune that a little bit. Largely what I appreciate is what you were also alluding to, which is the way that having the impossibility results allows you to move the conversation a little bit. Because it allows you to say, OK, we're taking the ability to have all of these things off the table.

And you now have to actually discuss what it is you want and what it is you believe about the base error, the base rates.

MALE SPEAKER: Right, right, because the primary thing that I'm interested is not merely having a technical tool that allows you to adjust between them, but work on interpreting, as it were, what that knob is supposed to mean. So from a point of view of utility theory or strict decision theory, t this is weighing about how valuable fairness is in different degrees. And that seems like that's a bit conversation that we should have about how valuable different types of fairness is.

Do you know if anyone in this literature has discussed about what, as it were, the appropriate calibrations of the knob balancing between the two ought to be?

SORELLE FRIEDLER: I don't think that's in that, but I might have missed it. The other thing that I would say is that a lot of these conversations are happening right now. So those paper are about a year old. So, yeah, there's also just a lot of conversations happening right now about the questions that you all have been also raising about whether prediction is even the right goal and what it means to be framing things in this way. Yes.

MALE SPEAKER: Going back to your idea that, say, COMPAS should be retraining all of the time [INAUDIBLE]. What happens when the outcome on a particular dataset affects the database? So for example, your COMPAS might say X number of people in a particular community should be in prison. That affects a particular community, but then that data goes back into the retraining. Is that a problem? Is that something that anybody is looking at?

SORELLE FRIEDLER: So, I actually have a recent paper on exactly that, which the purpose in the predictive policing context that I was describing earlier but applies to any situation where essentially the data coming in is determined by your previous predictions. And, yes, it skews the data, and it skews the resulting predictions, unless you appropriately correct for-- it's essentially a sampling correction.

You have to understand that you are sending people there at a certain rate and sort of down-weight it appropriately.

MALE SPEAKER: I would question whether you can achieve that feedback communication based on the life impacts of sentencing and how that can lead to future crime. And so if you have you algorithm that says, oh, this person should be sentenced at this certain punishment. And then that sentencing impacts their life. It doesn't just impact your data, it impacts their life and then their future. Because they're put into a situation where future criminal behavior comes out of that, not because of the data, but because of how it impacted their life.

SORELLE FRIEDLER: Yeah, and you're right. And should say that all of these results are in nice laboratory settings. They are where you've simplified the problem enough so that you can actually reason about it. What you're describing is, of course, more of the real-world context that's hard to incorporate.

Kristian Lum, who I mentioned earlier does have some nice work showing the way that a rise in recidivism rates in one neighborhood has a epidemiological spread.

MALE SPEAKER: Yeah, I would wonder, though, about that. It's great to have things be laboratory, but the goal is to use it in the real world. And So that's kind of a problem when you have, OK, we've got this algorithm that's based on these ideal conditions, but our intention is to use it in the real world.

SORELLE FRIEDLER: Yeah, I mean , I think--

MALE SPEAKER: It seems unfair.

SORELLE FRIEDLER: So I think that there's-- this is a much longer conversation than we can have right now. Right, I think that there is-- I would argue that there is value in trying to understand what your model does in an idealized situation, that you do learn something from that. And if, for example, you learn that even in an idealized situation you will have problems, which is what we were able to show, then perhaps that implies that in a far messier situation, you would really have problems. Perhaps that tells you something.

I agree that it doesn't necessarily mean that if you fix those problems in an idealized situation that you've fixed them in the mess real-world situation. But I still would hope that you learn something from it.

MALE SPEAKER: In terms of science, that all seems great. In terms of ethics and whether or not you should--

SORELLE FRIEDLER: Yeah, so I think the broader question of the implementation of the tool in society, I think is also a super tricky ethical question. So if you think just about risk assessment instruments-- so I talked to defense attorneys about this, and said, what do you think about this? And I've gotten varied answers.

So some people say, well, the judges are racist already. And we haven't been able to do anything about that. So maybe this is actually good, because this we could have some control over. And we could-- it's not a perfect tool yet, but maybe we could try to move it toward something more fair. Others of them say, yes, the judges are racist, but I know how to deal with my current racist judges. I've found ways around that. And now you're going to put in something that I can't interrogate and that my current mechanisms for dealing with this don't hold up under.

I think that both of those are right. Right? So I don't see it as a cut-and-dried let's just throw the tool out. I think it's a far more complicated question.

MALE SPEAKER: At any rate, that's not what you're trying to achieve. You're trying to achieve more algorithmic design, working on the tools though, and not necessarily understanding how it it should be.

SORELLE FRIEDLER: Right, I'm trying to say, if you're going to use a tool, here's a bunch of better options for you to choose from.

MALE SPEAKER: Sure, yeah, that's something.

MALE SPEAKER: Yeah, I think there's another use that hasn't been mentioned, which is what the use of a dataset and an algorithm would be for a judge who wants to detect his own bias and counter it. And so, under those conditions, it's a-- how big a dataset that you want. You want it really big. And how much would you want to know about the algorithm to raise a red flag to prevent you from doing something that you might not otherwise realize was a mistake?

SORELLE FRIEDLER: Yeah, and under the Obama administration, there was also a group at the White House that was helping local police departments essentially use a predictive algorithm on their own police force to try to find police who might then use excessive force and should be rerouted to further training. So, yes, there's all sorts of fascinating ways to turn this in on itself and use it to examine the system from within.

MALE SPEAKER: So defense attorneys can bring in one of those algorithms--

[LAUGHTER]

MALE SPEAKER: --the editor of COMPAS system, or their--

MALE SPEAKER: There was actually and *LA Law* about that like 25 years ago.

[LAUGHTER]

JENNIFER GUNN: OK, I just want to say thank you to Sorelle and remind you that if you want to consider these conversations again and for new conversations, come tomorrow morning to Walter Library at 9:00 AM. Thank you.

[APPLAUSE]