THE BROOKINGS INSTITUTION

THE ETHICAL ALGORITHM

A CONVERSATION WITH AUTHORS MICHAEL KEARNS AND AARON ROTH

Washington, D.C.

Tuesday, January 14, 2020

PARTICIPANTS:

NICOL TURNER LEE, Moderator
Fellow, Center for Technology Innovation
The Brookings Institution

MICHAEL KEARNS
Professor, National Center Chair, Department of Computer and Information Science
University of Pennsylvania

AARON ROTH
1940 Bicentennial Term Associate Professor,
Department of Computer and Information Science
University of Pennsylvania

* * * * *

P R O C E E D I N G S

MS. TURNER LEE: Okay, we're going to get started. Okay, welcome everybody to Brookings. Good afternoon. Okay. I go to a black Baptist church and when the preacher says good afternoon you gotta say it a little louder. Good afternoon, everybody. I know there are more people for our webcast audience than the five or six that sounded that soft.

I'm Nicol Turner Lee. I'm a fellow in the Center for Technology Innovation here at Brookings. I'm happy to host my first event of the year with these two great people who are sitting next to me, both of which have an individual story about me that they will not tell for any money. (Laughter) Today's topic actually continues a series that we actually launched last year at Brookings on artificial intelligence. My colleague and boss, Darrell West, and I have put together a series around AI governance, as well as bias, and Brookings overall has also dealt with the issue of national security. For those of you who have not yet seen it on our website under the tab of AI, our series of papers in the area of governance and bias -- and we actually just encourage all of you to go to the website because there is actually just great content, including the content of a recently released paper by the two panelists here, which we'll talk about today. So, again, under the AI section on the Brookings website you'll find two streams of paper on governance and bias that I encourage you to actually not only look at today, but to continue to track it as we actually put more content from reputable scholars across the country and the world on that site.

I'd also like to remind you, because we obviously have an audience that is in person as well as webcast, that you use #AIBias to have the conversation while the panel is going on and to continue it when it is over. These conversations are valuable, but if the only people who are actually talking about are us, right, then it's not really doing much. So we ask that you continue the conversation.

I'm really excited about this conversation today on bias and for a lot of reasons, but the first one most importantly being that I'm sociologist sitting with two computer scientists. Doesn't happen that often. (Laughter) And so I'm excited about the fact that my two friends have decided to join us today to talk about a book that that they recently put out called "The Ethical Algorithm".

So sitting right next to me is Michael Kerns, who is a professor and national center chair for computer and information science at the University of Pennsylvania. And next to him is Aaron Roth, who is the 1940 Bicentennial associate professor of computer and information science at the University of Pennsylvania as well. Let's give them a round of applause. (Applause)

And I will shamelessly say it one more time that they put out a book called "The Ethical Algorithm" -- that again, as I promote my book and others here, we want people to say it repeatedly that you should pick up -- that we'll go deeper into the content that we're going to discuss in this next hour.

So my plan is to ask you guys questions. And then again we're webcasting, so there are people who are watching us on line and we have people here in the audience. At a certain time in the discussion we'll open it up for Q&A and the move forward from there.

So I'm just going to jump right into it because I want to make sure we save enough time for Q&A. Let's start with what is the ethical algorithm? So give me something to work with there. I did read parts of the book, didn't get through everything. Also read a lot of the commentary on the book, but I would love for you to unpack, what do you mean when you say the ethical algorithm?

MR. KEARNS: so our book is really mean to be a popular science book. So many of you are aware about kind of algorithms gone wild situations in which algorithms inflict violations of people's privacy or make discriminatory decisions in lending and other domains. And we are basically representatives of a sub community of the AI machine learning community that basically is trying to design better algorithms in the first place. So, you know, to say it very briefly, we're part of a community of people that try to pose precise definitions, precise mathematical definitions for concepts like fairness or privacy. Then once one has committed to such a definition or definitions, to walk down the road of actually implementing those definitions in algorithms, like literally putting them in your python code, for instance, and then exploring what the consequences of that are, both from a quantitative kind of mathematical perspective, like tradeoffs between things like fairness and accuracy, and also more generally of the societal consequences as well.

MS. LEE: Aaron, you want to jump in?

MR. ROTH: Yeah, so I think the sort of starting observation for a lot of the work that we do and that we talk about in the book is that many of the examples of algorithmic misbehavior -- so examples of discrimination, examples of privacy violations that many of you in the audience will be familiar with because they've been featured in the popular media so much in the last few years -- are not the result of some malicious software engineer hiding behind a curtain who is intentionally encoding, for example --

MS. LEE: Oh, shucks.

MR. ROTH: -- racism or sexism into the algorithm. You know, in some sense that would be easier to deal with. We have sort of a regulatory understanding for how to deal with malfeasance.

The problem is worse than that, which is that these are the unintentional side effects of the standard machine learning pipeline, where you specify some reasonable sounding objective function, some proxy usually for accuracy or profit. And then you use an automated technique, a machine learning algorithm to search over a vast space of different models and algorithms to optimize for that narrow objective. And it's hard to predict what you're going to get out of this pipeline, except that you're likely going to get something that's very good as measured according to your narrow objective. But what we've been seeing is that there are often unanticipated side effects that take the form of things we think of as unfair or violations of privacy.

And so now what you need to do to avoid this is, first, think very hard about what it is you mean when you say that you don't want unfairness. You know, it's not enough to talk about it in English, you have to think in a mathematically precise way if you're going to embed this as a constraint into an algorithm, which is a mathematical object, and then you've got to solve the algorithm design task.

MS. LEE: So I want to go into that, because I'm a sociologist, I'm not a scientist. When you talk about these precise definitions of fairness and accuracy and ethics, even so, what do you mean by that, right? Because part of the challenge that we have, we're actually looking at

algorithm fairness, is there are many different definitions of what that means.

So I'd love to hear from you all, like what is the preciseness of that definition that you're referring to?

MR. KEARNS: Yeah, so maybe one comment worth making early is that, you know, when we talk about precision we don't necessarily mean that there's a single precise definition.

MS. LEE: Okay.

MR. KEARNS: There might be more than one precise definition. So to ground this a little bit, often definitions of fairness start with first of all a group that you want to protect, like a racial minority, gender, et cetera. And then what you consider constitutes harm to that group in a particular application. So let's take consumer lending for instance. So first of all, you know these days things like consumer lending decisions are largely becoming automated and algorithmic, and I think everybody knows that, and so in the context of algorithmic lending decisions I might decide like well, I want to protect a racial minority and the harm I'm trying to protect them from is false rejection for a loan. So I want to prevent the even that somebody who is credit worthy and would repay a loan if I gave it to them is actually denied a loan. And because when we talk about algorithmic decision making these days, we usually mean more specifically predictive models trained using AI and machine learning methods. So it's not like that a human programmer sat down and wrote python code to decide who gets a loan and who doesn't, they instead wrote the code that searches based on historical data for the model that then is the actual algorithm that makes the decisions.

So because of this, because it's statistical, you're going to make mistakes, right. And so in lending there's kind of two types of mistakes you can make, right. You can give loans to people who default or you can deny loans to people that would pay them back. And so maybe we decide that well what constitutes harm here is being falsely denied.

And so one specific definition of fairness that you could ask for is you could say, look, I don't want to just find the predictive model that makes the overall error as small as possible, I

want to find a model that makes the overall error as small as possible, subject to the additional condition that the false rejection rate on the minority group and the majority are the same. Okay?

MS. LEE: Okay.

MR. KEARNS: Now, just let me just use that to go one step further. I could ask that they literally be the same. So whatever the false rejection rate on the minority group is, it has to be identical to the overall rate. But I can relax that also. I can say like well they don't have to be identical, but they have to be within 1 percent or I could say within 5 percent or 25 percent. And importantly this gives us a knob or a parameter, as we would call it in machine learning, that lets us basically say how much fairness we want or how much unfairness we're willing to tolerate. And why would I want to do that? Why wouldn't I want to always say like well, no, make these error rates identical? Well, when I do that, that's going to come at the cost of overall accuracy, right, because it's just a stronger constraint. And so this knob lets us quantitatively explore things like the tradeoff in a particular data set that we face between accuracy, which in general we want, but also the amount of fairness that we're demanding.

MS. LEE: Aaron, you want to chime in? Because I've got some follow up questions.

MR. ROTH: Yeah, maybe just briefly. So along the lines that, you know, we're not proposing that there's one universal definition of fairness, rather there's a thought process that you go through when you think about what you might mean in a particular circumstance. The kind of definition that Michael is talking about is one of many that follow the following form. You think about in your context what is the action that sort of causes harm, what's the kind of mistake you can make in your predictive technology that causes harm. And then you ask that the harm that are caused by your algorithm and your algorithm is inevitably going to cause some harms, because again it's not perfect, shouldn't be disproportionately borne by one population. Somehow the premise here is that if these harms are sort of -- if these harmful mistakes are sort of made at random, spread across the population, maybe that's okay, but the thing that we don't like and the thing that is often flagged as objectionable in the popular media, is when we find that these sort of mistakes of the

harmful variety are disproportionately concentrated in one group. And if you can identify what -- you know, in your context, in your problem, what are the harmful mistakes, you can try to design your algorithm so that those harmful mistakes are not disproportionately concentrated in one group.

MS. LEE: So I want to kind of unpack that little bit, right. So I've sat on panels as a sociologist where scientists have suggested that 1 percent in accuracy means that the automated decision has done a better job than if something was physically -- like mortgage approvals. So there was an article that came our recently that said automated decisions do a better job in approving loans, primarily because you take out the potential discrimination by the lender if they see the person's race, gender, et cetera.

The challenge I'm having with some of the comments -- and I want you to help me with this, because it's I think the tradeoffs where people like myself get nervous. That though the accuracy is improved by 1 percent, it doesn't necessarily stop the fact that in a criminal justice algorithm you're still over incarcerating 1 percent of the African American population because they automatically start out flawed in that data base, or you're still rejecting 1 percent of people of color who are denied loans because they already start out flawed before they even enter the system.

So speak to me a little bit about those tradeoffs, right. And there's been some conversation in some panels that I've spoke about where some people say well sometimes you've got to take a tradeoff because the human -- the reduction in the human error turns out to be better because more people who may be denied loans, you know, actually goes away or there's a decrease in that number.

So help me unpack that a little bit in terms of the conversation that I have at the dinner table often times with scientists around accuracy and tradeoffs.

MR. KEARNS: Okay, let me first try to unpack what you said -- it had a lot in it. So first of all, you know, I think at the beginning of your question you were sort of alluding to the feeling that some people have that because algorithms are involved decision making that somehow things like racial, gender, or other biases are reduced or eradicated. And I think this is now demonstrably false. I mean you can -- I'm sure there's a new news article somewhere today about an instance of

demonstratable racial or gender bias in some predictive model.

Now, you know, so we tend to think of racism, for instance, as a human trait and, of course, before computers were used for things like lending decisions, there was such a thing as racist loan officers. But just because we replaced them with computer programs doesn't mean that we eradicate these problems.

And there's at least two important ways that these biases can creep into algorithmic decision making. One of course is just that the data that you feed to your predictive model in the first place is biased. So, for instance, if police officers are racist and selectively stop and frisk a minority group more often than the majority, they're just likely to find more contraband, for instance, in that particular group. And if you just naively feed data like that to a machine learning process, not surprisingly it's going to learn okay, this minority group seems to be carrying contraband more than the background population. Even if you eradicate that kind of bias, that doesn't mean that you won't end up with a biased model, even if the data is kind of unbiased, whatever that might mean, again because machine learning is essentially an optimization process. And if I ask a machine learning algorithm to find the predictive model with the lowest error full stop and it happens to be that, you know, there's a model that has infinitesimally smaller error at the expense of huge racial disparities, it's going to go for it, right. And I phrased it as this quantitative optimization problem, I didn't ask for any notion of fairness, so I'm not going to magically get it for free.

I think you were also kind of alluding to things that I think we think are not well addressed in the scientific literature right now, which have to do with kind of counter-factuals and kind of thinks I think that start to bleed into affirmative action issues, which is like, okay, maybe it really is the case that in this -- you know, that in some particular community that's lower income, small business loans are a higher risk than in some wealthier neighborhood. And so there's actually a legitimate argument to like not give loans in that impoverished neighborhood as much as in the wealthier neighborhood. But then you can sort of say like well, but maybe this neighborhood is impoverished in the first place because of your history of not giving small business loans to this community. And if you did that for a while then you would raise their fortunes. And, in fact, now

everybody could win because now here's a whole new community of people that are good lending risks for the lender.

And this kind of sort accounting for feedback loops where some algorithm or process or human organization is making decisions of time and their decisions in turn are affecting kind of the nature of the population that they're making decisions about, is something that you could in principle incorporate into machine learning methods, but very little of it has been done so far.

MS. LEE: Did you want to respond?

MR. ROTH: Yeah. So maybe to the point of tradeoffs. Maybe people like us talk about tradeoffs sometimes a bit more than we should. So if you're really in this idealized world where the data you have is perfect, the thing you're trying to optimize for really is the thing you're trying to optimize for, then it's true that if you want to additionally impose constraints like fairness or privacy, you're inevitably going to have to pay something in what you're optimizing for.

But when you look at real case studies, we're not even close to this idealized world. To use a recent example, there's a recent study showing racial bias in a united health product that tried to target medical interventions at patients. And what it was trying to do is -- you know, what it was supposed to be doing is it was supposed to be predicting health outcomes. But they didn't actually have data for health outcomes. Instead, what they used as a proxy was future medical costs. And it ended up having sort of demonstrable bias against African American because sort of equally sick African Americans had lower healthcare costs, not because they had better outcomes, but because they had less access to healthcare.

So this is an example where there's not necessarily a tradeoff. You could simultaneously improve fairness and accuracy by simply improving the data, trying to optimize for the right thing instead of this incorrect proxy.

So these tradeoffs really only manifest themselves after you've solved sort of all of the other problems and in the sort of -- in many practical cases you can simultaneously improve accuracy and fairness simply by getting better data.

MS. LEE: And that's why I asked the question, because I think in reading about the

ethical algorithm, what intrigues me about it, it's sort of placing ethics on what's already imperfect, right. Because if you look at the case of criminalization of African Americans, it has little to do with how much contraband that they have, it just has to do with a system which disproportionately makes more arrests in black communities, which allows them to be overrepresented in data sets. Or in the racial healthcare bias, it has to do with the fact that there is differential access to disparate access to healthcare, which made that feature ineffective for black patients because we're not in the system.

So I guess when you look at that, you look at this ethics framework for bringing some type of remedy to these imperfections, help me understand the correlation then between ethics and algorithms in your book and how do you actually see how everyday people sort of operationalize that.

MR. KEARNS: Yeah, I mean I think on these points -- I mean let me just point out that even in the example that Aaron gave where like look you can really go do something that would improve both the accuracy and the fairness of this predictive model, there's still costs involved, right.

MS. LEE: Right, right.

MR. KEARNS: So the reason that this product in the first place used health costs as a proxy for outcomes was because that was expedient and they didn't have the data on outcomes.

MS. LEE: Exactly, exactly.

MR. KEARNS: So to get that data, they would have to go do something that would cost them money. And, you know, we think they should do something like that, but we should pretend like, oh, you know, they just didn't realize that they had that data sitting on a different server and that it's just a matter of like pointing their model over there than over here.

I also think that one of the things we try to be careful about in the book is that we are first and foremost scientists, right. We learned a great deal in talking to people about these issues that aren't scientists. But we're quite careful to, you know, sort of identify what science can

do for us and what it can't do, okay. So we're not here to propose, for instance, that algorithms pick definitions of fairness or privacy. We view that as firmly in the human domain, in the societal domain. And so we try to very quickly sort of say like look, we're not imagining that an algorithm is going to kind of tinker around with the definition of fairness. We have to specify it. And that's a hard problem, right. In some ways that's harder than the science that we talk about in the book.

And there's also a lot of things that -- you know, even setting aside for precise definitions that you can explain to a computer -- there are some things that are just like hard social problems, right. So if, you know, your police are biased in who they decide to arrest or stop and frisk, if your criminal justice system is biased in the sentencing decisions that it makes or the bail decisions or parole decisions that it makes, you know, there's no algorithm to fix that.

MS. LEE: That's right, that's right.

MR. KEARNS: We're still left with hard social problems of like retraining the way people view these issues and changing their behavior, right. And so it's an interesting area to work in as a scientist because there's quite a bit to say about the algorithmic issues. But then there's this whole pipeline. I mean there's sort of the world, then there's some process that collects data from the world that might have problems. That produces data, that data is fed to an algorithm, that algorithm produces a model, that model makes decisions that might then in turn affect the world. And we're kind of talking about the middle part of that pipeline for the most part, but some of the hardest problems are like on the other sides of it.

MR. ROTH: So I just want to add to that. I mean when we talk about these issues in fairness and machine learning, it's always a little bit in the abstract, mostly because people have only really just started studying this. These aren't widely deployed technologies. But you can get a sense for where these things might be in 20 years by looking at sort of the parallel field of data privacy, which is maybe 20 years ahead. And we talk about it quite a bit in the book as well.

So data privacy followed a similar scientific pattern where maybe 15 years ago a particular concrete like mathematical formalization of a kind of privacy called differential privacy was proposed and for a decade people like us wrote mathematical papers about it for other people like

us. But it's recently become a real technology and once that happens these tradeoffs become very real.

So the census, you know, right around here is going to release all of the statistical products as part of the 2020 decennial census, subject to the protections of this thing called differential privacy. And there is a committee of people who are going to have to literally on a quantitative privacy parameter by sitting in a conference room and looking at plots about how privacy is going to trade off with the accuracy of the statistics. And there's not clear answer how to make this decision. I mean privacy is very important, the census is required by law to protect privacy. But on the other hand, the statistics released by the census are incredibly valuable. They're used to distribute federal funds, they're used by demographers and other social scientists to study important questions. And in this case, now that the technology is developed and you can actually like quantitatively plot out in your use case how privacy is going to trade off with accuracy, there are difficult decisions to be made. If you want more privacy, you know, you're going to have to give social scientists and federal agencies less good data and vice versa. And I think that, you know, although we're only talking about these things in the abstract now, you can imagine that 20 years down the line as some of these sort of fairness supporting technologies are put into practice, real decisions about tradeoffs are going to have to be made. The science is not going to have an answer as to how those decisions should be made. Of course it will be context dependent. What we can sort of most aspire to do is to make the tradeoffs transparent so that stakeholders can make these decisions with their eyes open.

MS. LEE: Right. So let me ask you about then -- this is how I actually met Michael, when he gave a really great talk around inferences though, right. And I want to kind of bring this into the conversation. Talking about data privacy, it probably was easier to talk about privacy when we're talking about rules of the road in privacy, you know, what particular things do I want protected, what can be anonymized, what can be de-anonymized. Now we're talking about this culmination of different micro factors that people know about us. And when you look at algorithms, as much of the argument around -- and I want to keep driving back to this ethics right, of

what happens with algorithms is that the data it's trained off of is not just me as a woman of color, but me as a woman who likes black boots, me as a woman who wears my hair in braids. That then makes assumptions about what I might purchase, what movie I might watch, where I might go on vacation, et cetera.

So the question becomes what do we do about that, right, because there's not necessarily a framework and that's in the ethics space of what is on limits and what's off limits when it comes to making certain eligibility determinations and other kinds of decisions?

MR. KEARNS: So I think this touches on both fairness and privacy. And just to make sure we're all on the same page, the kind of phenomenon that Nicol is talking about is -- and there have been recent scientific papers on this -- but the basic intuition is that these days we as individuals generate so much data that we might think of as innocuous and unrelated to who we are, that in fact acts as something similar to a fingerprint. So to make this concrete, once you know that let's say I'm an academic and that I use a Mac and that I drive a Subaru Forrester, and a couple of other things, maybe you have a pretty good guess as to my political affiliation already, for example.

And there have been very striking demonstrations of this. We discuss in the book a paper from now maybe five years ago that showed in a convincing fashion that just from your like behavior on Facebook -- so nothing about who you are, your age, your gender, just from the content that you click the thumbs up button on on Facebook it's possible, to some level of statistical accuracy, to predict things like whether you are the child of divorced parents, your drug and alcohol use, your sexual orientation, and the like. So these things that you may think are just kind of digital exhaust, you know, you give me enough of that and it correlates strongly with who you are. And this has a couple of implications, both for privacy and fairness.

So the implication for privacy, as we discuss in the book, is it basically means any definition of privacy that's based on notions of anonymization or removing so-called personally identifiable information, which unfortunately is by far and away the standard notion of privacy used in commercial practice, to the extent that it's articulated at all. Those notions are all fundamentally

broken, right. And there have been many instances whereby even though they remove your name, your age, your social security number, all the things that we think of as personally identifying, if I just know enough about your purchases on Amazon or your Facebook likes, I can pretty much reconstruct many, many other things about you. And this is why, as Aaron mentioned, this notion of differential privacy, which kind of eradicates the -- it basically is very different than notions of anonymization.

MS. LEE: Right.

MR. KEARNS: The implication for fairness, one implication for fairness of these kinds of correlations or inferences, as were calling them, is that again -- so in consumer lending there are laws in the United States that forbid the use of race as an input to a lending model, okay. And the idea is that oh, somehow by not allowing race to be a variable in the model that you guarantee racial equality. So nothing could be further from the truth. And, in fact, you know, in our book we give a very concrete example of a simple setting where in fact by refusing to allow race as a variable in your model, you guarantee that you will harm the racial group that you were intending to protect by forbidding that, okay.

And part of the problem is that if you try to restrict the inputs to a model, there are so many proxies for anything, right. I mean unfortunately in the U.S. your zip code is already a rather accurate proxy for your race. So notions of fairness that are predicated on forbidding certain inputs to the model are broken in the same way that notions of privacy that are based on anonymity are broken. And we advocate in the book that like the fairness definitions we were given before, what you should really be doing is specifying the behavior you want at the output of the model rather than fooling yourself into thinking that you're getting some notion of fairness by controlling the inputs.

MR. ROTH: Yeah. And this issue of, you know, all of these correlations out there in the world, I think there's really two distinct things that people might think of as privacy violations. And differential privacy I think helps in disentangling these two.

MS. LEE: Right. And just for people who are not following the privacy debate,

differential privacy, from what I understand, is when you put more noise in the model that do not allow a person to sort of pick up your attributes. Is that right, Aaron, on that?

MR. ROTH: That's right. Yes, maybe it's helpful just to say briefly what differential privacy is. It's a particular formalization of privacy that corresponds to plausible deniability. And it sort of takes the position that if I did some data analysis that didn't involve your data at all, then we shouldn't think of that as a privacy violation for you. And what differential privacy requires is that even if I do use your data, there should be no statistical test, nothing any outside observer can do that can let them distinguish better than random guessing whether we're in this idealized world where I didn't use your data at all and whether I did use your data.

So this is a strong guarantee that I can't learn anything idiosyncratic to you from whatever is published because I can't learn anything that I couldn't have learned even if your data wasn't in the data set at all. So the kind of thing that this can protect against, for example, just to use sort of a silly example, you know, I don't know about you all, but the things I type into Google search when I'm like alone in my room are embarrassing, I wouldn't want people to know what those are. And you might worry that Google's auto complete feature that suggests what should be completed when they start typing.

MS. LEE: Are we going down that road, Aaron? We don't want to know what you do in your room.

MR. ROTH: But would reveal what I'm -- so differential privacy prevents against that kind of thing.

MS. LEE: Right.

MR. ROTH: Now, there's the separate concern that from the kinds of things that I make publicly available, for example, I've set my Facebook setting so that, you know, people shouldn't be able to see things I think of as sensitive -- maybe my political affiliation, my sexual orientation, but I allow them to see the things that I like because I think that's innocuous. I might also think of it as a privacy violation that people actually can correlate the things that I like with these things that I thought of as my secrets and can guess accurately my political affiliation, for

example.

Differential privacy does not protect against that, and nothing could because I have just sort of -- the problem there is I have inadvertently revealed to the world by setting, you know, my Facebook setting to public all of these facts that happen to fingerprint these sensitive attributes about me. The problem was I didn't realize that these correlations were there. But preventing that I think requires education. Once people make that information public there's nothing technical or even regulatory you can do to prevent these kinds of inferences from being made, because they are just facts about data that anyone can see in the world.

MS. LEE: But I think what you're -- and I want to talk a little bit about it and we'll go to questions soon -- what you're talking about in terms of just this fairness model.

My colleague, I think he's out here, wrote a paper on lending -- Aaron Klein -- around people being able to determine if you're a good credit risk, if you're using an Apple versus a PC, right. So there are certain things that people can get.

I mean the question is yes you can use those features, but again, going back, is it fair? And as developers how do you get people to sort of weigh that? I mean you're scientists, you're always in the lab creating these models that get deployed in different contexts that may or may not have unforeseen consequences.

So the question I have for you before I go into the policy part of this, what do we need to tell developers like yourself to actually employ this framework that embodies fairness and ethics and makes a determination that maybe this feature is not a good idea to use or is thinking ahead that this proxy is going to create certain levels of bias?

MR. KEARNS: So first of all, I think the right party to be focusing this is on is sort of not developers in the traditional sense, but what we would now call data scientists or machine learning people at let's say large tech companies, or lenders, or the like. And, you know, I think the good news is that the prescription for what they should do different, for instance, to enforce fairness considerations is only minimally different in a scientific sense from what they're already doing.

MS. LEE: Right, right.

MR. KEARNS: So a lot of what these people do for a living is essentially solve optimization problems, data driven optimization problems. They have a historical data set of loans that were granted. Some of those loans were repaid, some of them defaulted. You know information about the applicant in each one of those cases and are using that historical data to say like, find the neural network which does the best job on this historical data of predicting who would repay and who wouldn't, okay. Really, the only modification they need to make is to instead solve what we would call the constrained optimization problem of find the neural network that minimizes the predictive error, subject to the constraint that the false rejection rate on black people and white people not be greater than 1 percent, okay.

And so like conceptually these things are right next to each other. Now, that second problem is algorithmically more difficult, okay, and so you need to think about algorithm design to solve it. But it's in the same field, it's not like asking these people, okay, you need to go off and study for five years. It's really completely within their wheelhouse to make these kind of changes. And, by the way, many of our colleagues in industry know this and many of them in fact do the kinds of research that we're talking about in these areas. It's just a matter of those companies having the appetite to make those changes because as per what we said before, this will have real costs. If I'm Google and I want to make sure that in the ads that I show there is not racial bias, for instance, is who is shown lucrative ads for STEM jobs. Enforcing that will mean lower accuracy overall in predicting click-through rates and this is how Google makes money. So they will literally lose revenue by enforcing this. And they need to decide whether they want to do it or not.

MS. LEE: Right. So I mean but I want to stay on that for a minute because I look at it like this, permission-less forgiveness -- or permission-less innovation has led to permission-less forgiveness, where we hear a lot of “I'm sorry’s” after something has been broken. And so I love what you're talking about in terms of getting people to realize that they have these tools in their toolkit, but the reality is one company that is now a tech company is rushing to market quicker than the next company and the next company. And the often don't have the data necessary to actually

make these kind of decisions. So do you stifle innovation or do you still operate in permission-less forgiveness, or do you find something in between when it comes to the data science community?

MR. ROTH: Yes. So one of the things we talk about both in the book and this policy paper is that we do think that regulatory agencies have to become substantially more quantitative. Not always that the data is not there. You know, there was this recent sort of assertion that maybe the Apple card was exhibit gender bias and we sat down with some of the New York regulatory people looking into this. And what they said is that at many of these sort of lending institutions in principle they have the data to check, for example, whether their system is systematically giving lower credit limits to women versus men. But they don't want to look into it because this would open this fact up to discovery if there were a lawsuit. And right now there's sort of lots of uncertainty about what regulation actually requires them to do at a quantitative level.

And so I think that in some sense like there's clear things they could do and they're -- because of ambiguity in the regulatory environment right now, they're reluctant to do it. I think that if we are going to start trying to productively regulate the algorithms, which are rapidly making decisions that we can quantify, we should be precise about what regulation requires them to do and we should think about how to audit them.

MS. LEE: Right. So I want to go on the regulation side. In a few minutes I'll answer questions. So I'm going to give you a little secret, there aren't a lot of quantitative people in government. That's not a space that people decide to go into government because they're -- I mean there are agencies that have quantitative researchers, but just nearly not enough to keep up with this innovation economy.

So then what do we do? Like how do we begin to employ regulatory frameworks that make sense given I think some of the concerns that we've talked about so far?

MR. KEARNS: Yeah. I mean so I think this is one of these things that falls into that category of hard problems with no easy scientific fix. You know, we understand the science, as Aaron said, and I think as we argue in the Brookings brief, that the kind of thing that we're proposing, while scientifically possible, is going to require real change at the regulatory agencies,

whether it's the current regulatory agencies or new regulatory agencies that would be created. And the composition of the personnel at these agencies will need to look very different. You know, the gap between technology regulators and the companies that they're regulating has grown I think dramatically in the last 20 years with the advent of the consumer internet and the rise of AI and machine learning in these companies. And I think it's not a level playing field.

And so until we find ways to get more people that think like the companies that they regulate do in the agencies and know -- that have the same skill set, I think it's going to be very, very difficult.

Another thing we discuss in the brief that's less about let's say fairness per se, but things about competition law. So in my experience in regulatory cases involving technology competition law, I've noticed that regulators are forced by law to view things through a lens that's entirely at odds with the companies they're regulating. So if some tech company is thinking about acquiring a startup in some particular subarea of tech, the first thing they have to do is identify the market for that particular piece of technology and then decide whether it's nascent or mature and who the competitors are and whether the acquiring company is kind of an incumbent or whether the playing field is sufficiently level. And then you talk to people at tech companies and they just don't see the world this way. They think like -- when Google thinks about acquiring a company, they're not thinking like oh, I'm like getting into this new narrow segment of technology, they're like, wow, if we had that data we could combine it with our data from Gmail, from search, from advertising, from Waze and Google Maps. And they don't think about it in this compartmentalized way. They think the whole thing is a single piece kind of bundled together by the data that they're gathering.

And when you force regulators to look at the world that way, it fundamentally handicaps them in sort of doing sort of things that have teeth in their domain.

MR. ROTH: Yeah, I don't have too much to add to that. And I'd say what we as scientists think the state of the world is right now is that, you know, we're at the place where in principle, in an ideal world, there are sort of concrete things you could do to audit and regulate algorithms for desiderata, privacy unfairness. But we have no particular expertise into how to get to

a state of world in which there are regulatory agencies that are actually prepared to do this and how difficult it would be.

And I think that what's needed is long-term collaboration between legal scholars, regulators, and technical folks, computer scientists, and hopefully conversations like this are the beginning of that. But I don't think we have a silver bullet in mind for how to get from A to B.

MS. LEE: Yeah, I want to just add to that. I love -- and, again, their paper is available on line at the AI bias website link. If you bring that down.

I mean I think, to your question of should we have more quant scholars at regulatory agencies -- of course, right. But I also think we also create another disadvantage barrier. Particularly we look at diversity, right. How many people of color are actually getting STEM degrees to be able to compete for those jobs? So then you have something that is now a tool that is widely distributed that affects everyday people with people who are either regulating or designing them that don't look like the people that are the subjects.

And so I think the whole thing is what Michael said -- it's a big complicated mess that we need to really think carefully around how do you mix the design, the regulation, and the other civil society aspects to create, particularly in the use cases that matter, right, algorithms that are ethical and are fair and have some level of precision that can be explained in the end.

So I don't know if you guys want to make a couple of final comments. I'm going to open it up to questions in just a moment, but any comments on that?

MR. ROTH: Yeah.

MR. KEARNS: I'd like to you to take questions I think. Yeah.

MS. LEE: Okay, perfect.

All right. So we're going to go to questions. We have a microphone coming down the middle. State your name and keep your question brief and not necessarily a commentary until I get as many as possible because I have 15 minutes.

MS. ASHE: Yes, yes, absolutely. Thank you guys so much for being here today, number one. This was very enlightening.

Number two, I would like to introduce myself. So my name is Keshia Ashe. I'm a former AAAS Fellow at the National Science Foundation with the Computer & Information Science and Engineering Directorate, education workforce development.

MS. LEE: That's a long title. (Laughter)

MS. ASHE: That's my former job. Right now I'm a free agent. I actually just launched my fifth company today called Treewaterconsulting.com.

MS. LEE: So I know you just want (inaudible).

MS. ASHE: And so -- no, no, no, so -- no, I do, I do. So I have -- I actually have an answer, a question, and then an answer. Okay. And so the answer to your question is you must encourage your industrial friends to employ more people who are women, people of color, people with perceived disabilities, all of those things, okay. We need to diversify the tech workforce. Okay.

You all are leaders. We need your leadership, okay. So that's the answer. But the question --

MS. LEE: And Keshia, we have a paper on that coming out, too, so you'll see that as well.

MS. ASHE: Excellent, excellent. The question is what are you going to do? The answer is I would strongly recommend contacting Barbara Whye, W-H-Y-E, at Intel. She has developed an entire system with her engineering degree -- I'm an engineer by the way and I'm getting my computer science Ph.D. very soon. Barbara Whye, look her up.

MR. KEARNS: W-H-Y-E, okay.

MS. LEE: Okay, they got it.

MS. ASHE: W-H-Y-E.

MR. KEARNS: Thank you.

MS. ASHE: Alton Coleman, James Madison University Ethics.

MS. LEE: All right.

MR. KEARNS: Okay.

MS. LEE: Thank you, Keshia, I appreciate that.

MS. ASHE: You're welcome.

MS. LEE: Thank you. All right, next question. This lady in the red.

MS. HELLERSTEIN: Hi, Nicol, it's Judith Hellerstein.

MS. LEE: Oh, yeah, hey.

MS. HELLERSTEIN: Question for you is we have so much in the algorithms. There is so much also bad data that individuals cannot correct. And many of these they say up to 30 percent of the data is not correctable. I would like to see more of how individuals can get the data and correct the data, because as you said, one of the problems with the algorithms is the bias. But what about the inherent incorrect information in there? That is creating more bad -- as they say more bad information in, bad information out. And how do we go about correcting that data?

Thank you.

MS. LEE: Right. So I'm going to bring that question over to you two and just sort of think about it before we answer that. What Aaron does in his room is what he does in his room. That's accurate data for any of us, but how do consumers what they think are not what's accurate about them?

MR. ROTH: So that's a good question. And just to sort of elaborate on it, it's true that right now, like if you go and you apply for a credit or a credit line increase you're going to get a decision in maybe less than a minute. But you don't have any transparency into what data was pulled in about you to make that decision. Whether it was just your credit score or whether it was using social media data, or the correctness. And that's a problem when there are data errors.

So what I would like to see is standards by which you would have the right, as you do at the moment with credit reporting agencies, to have transparency into the data that is being fed into these algorithms so that you could take a proactive stance to correct it when there are errors.

MS. LEE: Do you think that's a legislative action in terms of explainability of algorithmic decisions -- because you just triggered me -- much like credit reporting decisions?

MR. KEARNS: Yeah. So first of all, I mean I agree with Aaron on kind of what the

nature of the solution needs to be. Unfortunately this is one of these things that's sort of not -- you know, it's not a technical thing, it's a legal and policy things requiring companies to provide certain mechanisms by which you can see what they've got on you and you can change it.

And one obvious thing that I would point out is that we live in an era where the efficiencies of algorithmic decision making are such that allowing that kind of deliberative access and revision is at odds with the efficiencies that have been gained by technology. So there are going to be some hard frictions there.

On the topic of transparency and interpretability and explainability, these are also things that we talk about a little bit in the book towards the end. The reason we don't talk about them a lot is that we think that in relative terms the sort of scientific maturity of research in those fields is very, very nascent. So in particular, even though there might be multiple competing reasonable definitions of things like algorithmic fairness, you know, we don't think there's even a single good technical definition of what it means for a predictive model, for instance, to be interpretable.

And part of the reason for that is that if you think about the word interpretability, it implies an observer, right. It's sort of interpretable to whom. And do we mean interpretable to people who work in data science? We probably mean something broader. Do we mean interpretable to people who sort of have a very low level of numeracy, and then that would be a different notion of interpretability.

And so we think that the research that's needed here is almost more behavioral and that tends to not happen in areas like computer science, although there have been a couple of nice recent pieces of work.

MS. LEE: Yeah, no, just on that I think that this is really great because I think as a policy maker it goes back to people being able to audit algorithmic eligibility determinations perhaps that could have some feet to it.

MR. KEARNS: Yeah.

MS. LEE: So thanks for bringing that up. It's something to look at.

Over here, I've got a question.

MR. CHIRUVOLU: Hi, everyone.

MS. LEE: Hi.

MR. CHIRUVOLU: Hi, I'm Vikram Surya Chiruvolu. I'm a computer scientist and a psychotherapist.

MS. LEE: Wow.

MR. CHIRUVOLU: Not too many of me. Thank you, Nicol, for pulling this together.

MR. ROTH: Dangerous combination.

MS. LEE: Right, that is very dangerous. (Laughter)

MR. CHIRUVOLU: I'm also the CEO of an organization here in DC called Technotherapy.org.

MS. LEE: Wow, so there's -- I'm just going to not say anything else.

MR. ROTH: Found your niche.

MS. LEE: Niche, right.

MR. CHIRUVOLU: It's been an interesting course of life that got me to this point.

I just had to be here today for this conversation. One thing I want to put in the room is I worked as a computer scientist in industry for many years. I never had to sign anything that had anything to do with ethics. There was no expectation. When I became a psychotherapist and I became a licensed clinician I was subject to ethical codes of the American Counseling Association. My professional identity that was as a counselor. And so one sort of question I have is is there any way we're going to get out of this if we have a workforce that isn't actually themselves like committed to ethical practice in their work? And I think a place to think about that is obviously computer science and electrical engineering.

My second question has to do with the business side of this, but I think that's a -- the first question I asked.

MS. LEE: That's a great question. Should computer scientists be taught ethics?

MR. ROTH: So I think -- I mean that's a good idea, but I think that having people

sign a piece of paper is not going to solve the problem. And part of the reason is because I think it's very hard to anticipate the effects that -- say you're a software engineer and you're pushing some small update to the Facebook newsfeed for example, it might be entirely non obvious what effects that's going to have. And lots of the things we think about as possibly problematic in social media, like the creation of filter bubbles leading to maybe a less deliberative society, I think that those effects were not very easily predictable at all at the moment that updates to the Facebook newsfeed that ultimately led to these effects were deployed.

And so I don't think that the problem is so much that we have unethical software engineers, although there is some of that.

MR. CHIRUVOLU: Let me just interject here that I think you're exactly right, there's the (inaudible) of an unintended consequences. You build something, you have no idea how it's going to sort of go into the wild, right. But as soon as we know, for example, that we have this scenario with automated discrimination with mortgages, right, I as a computer scientist at that company, if I don't have any ethical obligation to pay any attention to that fact, I can go on doing what I'm doing. As a psychotherapist, I'm subject to licensing boards, I'm subject to things that have, as you say, teeth. Like I won't be able to make my living if I continue to ignore something that's a clear ethical result that's out there.

MR. KEARNS: So I mean I don't think it could be a bad thing to have people sign such a piece of paper, but I think in our view, the areas where there are problems for which we think there are scientific solutions, as we said that at the beginning, the problem is not among rank and file scientists and engineers who are inflicting their biases in the models that they build. These are corporate problems, right. So I think where the pressure is needed is not from the bottom up, but from the top down. It's to really make tech companies, for instance, or lending agencies have a policy or a law or regulation binding them to audit for things like gender or racial bias in their predicted models, and eradicating it. And that requires, as we've said, the regulators getting precise about what they mean and what the requirements are so that the companies don't have an incentive not to look at things that they don't want to see, okay. So you have to force them to look

at it and see it and report on it and let it be audited.

And so while I agree, yes, I do think engineers should be taught ethics. In fact I teach a course at Penn on exactly these topics. And, by the way, it's mainly a technical course. It's sort of how do you implement machine learning in a way that avoids fairness violations, for example. It's not --

MR. CHIRUVOLU: Is it required?

MR. KEARNS: Not yet.

MS. LEE: Not yet. Not across a lot of --

MR. KEARNS: They're thinking about making it a requirement. I mean, you know, that also limits the audience because then people from arts and sciences are more reluctant to come over and take some required engineering class. That's completed academic dynamics. But I think people need to be exposed to this, but I think where the real action is in the near-term is not that we kind of need to train better people in the abstract, is that there are concrete things we need to make companies do.

MS. LEE: Right. I --

MR. CHIRUVOLU: My question there is how? And forgive me.

MS. LEE: Okay, now you've got that mic for a long time.

MR. CHIRUVOLU: I know.

MS. LEE: Are you almost done? (Laughter)

MR. CHIRUVOLU: I'm going to let it go, I'm going to let it go.

MS. LEE: I'm trying not to have you analyze me in any respect in terms of your background. But you have been holding that mic for a minute.

MR. CHIRUVOLU: I'm giving it back.

MS. LEE: We have one quick 30 second wrap up because we're running out of time.

MR. KEARNS: Well, sir, just to your point about that, you know, I've worked in a corporation where we knew that there was an ethical issue, right, but just exactly as you described,

we knew competitors didn't care about it either. And my job as a computer scientist in a company is to do what the investors say needs to be done.

And so how do we actually -- because I think at a regulatory level, here in DC, knowing exactly to your point, they're not techies. Like we're not going to get into the weeds of these algorithms. Like does there need to be something with significantly more teeth. Like one concept, this is something that Elizabeth Warren has floated, is making every corporation a social benefit corporation by default so that a person who works for the corporation isn't in breach of fiduciary duty to make a socially beneficial decision.

MS. LEE: Right. Well, I want to add to what you said though. I think the questions that you're bringing up are really critical, particularly as the technology has evolved into this space where we have so much more data about people. And it's actually not just going to be a corporate effort, but it's also going to be the effort of individuals to understand what's happening to them as well.

And so I think going forward you're going to need that consumer education, the same way we moved from HGTP to an understanding that you don't put your credit card there, you put it in the HTTPS or all of these variables are going to have --

SPEAKER: (off mic)

MS. LEE: Yeah, right exactly. (Laughter) I mean not being shameless about this at all, at Brookings we're working on an algorithmic rating system that actually incorporates a lot of what you're talking about. But I do believe that ethics at the undergraduate and graduate level matters.

But I do want to add one more thing because I can, because I can hold the mic and say goodbye to everybody at the same time. Unconscious bias is still real. So you can teach ethics and people do come with their values and assumptions, but they also come with their assumptions and values of the world or are influenced by who they are, whether you come from the north, the south, the east, the west, you're a woman, you're a man you're short, you're tall, it all factors in there.

And that's all very important as we move forward with this debate that none of start with a blank slate.

So, with that, he is the reason that we ran out of time for another question. With that I want to say let's give a round of applause to Michael and Aaron. (Applause)

We want to again thank you all for coming out to participate in another event series on our AI work here at Brookings. Follow our paper series. And on your seats are evaluations. We want to improve the quality of the events that we give to you here at Brookings. Be sure to give us some comments so we can do better the next time.

Thank you very much and buy their book. (Laughter)

* * * * *

CERTIFICATE OF NOTARY PUBLIC

I, Carleton J. Anderson, III do hereby certify that the forgoing electronic file when originally transmitted was reduced to text at my direction; that said transcript is a true record of the proceedings therein referenced; that I am neither counsel for, related to, nor employed by any of the parties to the action in which these proceedings were taken; and, furthermore, that I am neither a relative or employee of any attorney or counsel employed by the parties hereto, nor financially or otherwise interested in the outcome of this action.

Carleton J. Anderson, III

(Signature and Seal on File)

Notary Public in and for the Commonwealth of Virginia

Commission No. 351998

Expires: November 30, 2020