THE BROOKINGS INSTITUTION

COULD RATING SYSTEMS ENCOURAGE RESPONSIBLE AI?

Washington, D.C.

Thursday, October 1, 2020

Welcome and Introduction:

NICOL TURNER LEE
Senior Fellow and Director, Center for Technology Innovation
The Brookings Institution

Panelists:

MARK MACCARTHY
Adjunct Faculty, Communication, Culture, and Technology
Georgetown University

FRIDA POLLI
CEO and Co-Founder
pymetrics

ELHAM TABASSI
Chief of Staff, Information Technology Laboratory
National Institute of Standards and Technology

JOHN VILLASENOR
Nonresident Senior Fellow, Center for Technology Innovation
The Brookings Institution

* * * * *

P R O C E E D I N G S

MS. LEE: Good afternoon everybody. We are live with you today and I'm so excited to actually facilitate this conversation among some great friends and some really smart people today here at Brookings.

I'm Nicol Turner Lee. I'm a senior fellow in governance studies and the director of the Center for Technology Innovation. And I'm really excited to actually have this conversation because in addition to the fact that I am writing a book, for those of you who follow me on Twitter, and at us, and on LinkedIn, and perhaps on Facebook, my book is forthcoming next year. It is on the U.S. digital divide, and it's about how the Internet is essentially creating a digital swath of people simply because of the systemic inequalities that overlay their experiences. That's coming out of Brookings Press.

And if you know me really well, you know the other thing I'm quite interested in is this intersection of race, technology, and social justice. And we've been talking about that through a series of conversations on racial equity and technology.

And if you know me even better, you'd know that artificial intelligence, particularly algorithmic bias and how we actually deploy these models in the real world matter to me. And over the course of the last couple years we at Brookings, we've taken this quite seriously. At the institutional level we have a great program that is designed around effective scholarship and leadership in the space, whether it's on AI governance, AI bias, in which I lead, or national security. And we even have added to TechTank, which is at the core of our blog at the Center for Technology Innovation, a podcast. The TechTank podcast which really looks at these issues in more detail. My cohost, Darrell West and I have been really get it in lately with these issues that matter the most.

And so today were going to do something that I think all of us who are on this call today, and watching us by video today we're going to look at algorithmic bias in algorithmics in general, but we're going to look at something that I've been talking about for a very long time, which are ratings systems, as well as what were seen getting done across the globe in terms of certification or labeling.

And why is that particularly important? We are now in a state where all of this information

that is being collected about us, not just the billions of variables before, but now, even more than that, right? Because we all have been home or we've been enjoying our movies, or shopping, or reading and doing all these things. They're now being collected in a way where the optimization of that information for bias is likely. And I thought before the pandemic that we should sort of think about how do we develop a feedback loop that actually allows developers as well as those who license algorithms to really think about how effective the performance is of that particular model.

Now, I'm a sociologist, I'm not a quantitative modeler, nor am I a scientist. But I do know one thing, and this is something that's been out -- I'm so happy that my friend Frida Polli is here today because we sat on a panel maybe a couple of years ago where our guy, Jason, made a comment. And he made a comment that the universe of commerce, for example, has changed to the extent that it is upon companies to win back the trust of people like you and I. And so this rating system idea, this EnergyStar rating that all talk about more later in the panel is really driven by that.

So I'm excited today to have some scholars who were actually out there, whether at the legal level, the design level, the government level, really trying to unpack what would this look like, and maybe offer critique on what we have seen happen in the EU and other places. I'm going to introduce them in the order in which I see them. They may look different to you, but I'm going to just ask them to raise their hand.

I'm joined today by Mark MacCarthy, who is a senior fellow for the Institute for Technology, Law, and Policy at Georgetown. I'm joined by Frida Polli, who is the CEO and founder of pymetrics which I'm excited to have. I've been trying to get her for about a year. She had a baby last year so I couldn't get her. This time, she's here. Elham Tabassi, my new friend who's actually Chief of Staff of the Information Technology Laboratory at NIST who is actually the governing body of standards as we look at artificial intelligence. And my colleague, John Villasenor, a nonresident fellow of Governance Studies who is doing absolutely fantastic work at UCLA. Thank you everybody for being here.

I want to jump in first, Mark, with you. Because I know you follow a lot of the stuff. In

fact, you and I are aligned when it comes to things like differential treatment, or disparate impact. We've seen this before. So before we jump into a conversation that really begins to unpack what we think we're seeing, why don't you take us back historically on how certification, licensing, labeling, and perhaps ratings actually existed in the past.

MR. MACCARTHY: Sure. Thanks very much, Nicol. And thank you for having me. This promises to be a really fun conversation, and I'm glad to be joined by all the other wonderful panelists who carry on the conversation with us.

So I'm not going to give you any kind of comprehensive history, but I do want to cite a couple of examples, and maybe we can learn some lessons from these examples about the possibility of rating systems and labeling for machine learning programs and algorithms more generally.

So the first one I want to remind people of the something called P3P which is a privacy operation that started maybe 20 years ago where people were worried about online privacy and I thought maybe we could get a rating system where websites would be matched with the preferences, the privacy preferences of their users, and a system was attempted to be developed and it didn't finally get into place, and will talk about why that might have been in a moment.

A second one, in a similar vein was a do not track proposal. It was the idea that company should get together in a standard setting operation and set up a technical standard that would allow users to opt out of online tracking. And once again, it didn't come to anything.

And the third example actually got off the ground, and it was called a V-chip and it was a requirement where the government said you've got to put these chips in the TV sets. And it required broadcasters to transmit rating codes for violence, sex, and language. And that would allow parents to block certain programs. The broadcasters used their own judgment about what the ratings were, but the rest of it was a mandatory program.

Now, in AI, in machine learning a vast number of systems have been proposed for ethics. And there's a whole cottage industry of doing ethical impact statements. But the idea of a Seal of Good Housekeeping approval, or nutrition labeling, or a rating system, that is not well developed, and it really is

worth pursuing. And I hope this conversation can help it move off the dime and get to be more widely discussed.

I have two reservations though. One is the element of coercion. The P3P and the do not track systems fail because of the lack of industry buy-in. This kind of thing would really need to be made mandatory. Otherwise, industry has really no incentive to come to agreement on the system, or to use it if it happens to be adopted.

But the second reservation is that even with coercion, the objective of an AI rating system is to give consumers and users more information about these systems. And to be honest, that might not be enough. Think about content moderation for a moment. Social media platforms, they label their content often so people can avoid it if they want, or discounted. But obviously they should do a whole lot more on their own. They shouldn't be carrying hate speech, or white and nationalist content. These content rules are obviously set by the private sector, not by government agencies, but they should do more than just label objectionable content. They might also want to be responsible enough to restrict it entirely.

Now, most uses of algorithms don't raise these kinds of expression questions, so government agencies might usefully raise questions about the intended use of algorithms, not just whether they're doing what they are doing in a fair and unbiased way, but whether what they are doing is worth doing at all. This is the kind of question that Frank Pasqual he raises in what he calls the second wave of algorithmic accountability.

And that idea of assessing the actual purposes and uses of algorithms, that's worth pursuing as well.

MS. LEE: You know, I love what you're talking about. And it's so funny, and I'm going to give it over to John. And oh I -- before I forget, please Tweet this conversation hashtag AIbias. And if you have questions, we will have time during this conversation to go to them. Please send them via email to events@Brookings.edu. Events@Brookings.edu.

So John, I love what Mark is talking about, right? Because I have gotten pushback and

some people have heard me talk about this EnergyStar rating, right? Which I'll talk a little bit more about later. And part of that came from, I don't know she's watching, but we were doing a paper a couple of years ago. A woman by the name of G.D. Barton, who was at the Better Business Bureau was actually talking about the BBB certification process. And as we delved into the papers that we actually put out at Brookings with another author, Paul Resnick, an engineer, it appeared to me that we don't have anything that sort of suggests, because as Mark said we don't know what's under the hood, you know, whether or not this algorithm is performing in the way that it's supposed to. And that particular time I was writing that paper, my dishwasher broke. And I went into a big box store and the first thing I was attracted to was the big yellow label that is certified from the FTC that essentially tells me about the efficiency of that particular appliance.

So I'm curious to you, when we think about rating systems, and we think about what we heard Mark talk about, talk to us about what is under the hood of these models and why they have sort of limited transparency that might require, I think what Mark is suggesting, a different conversation in addition to the suite of resources like impact assessments and audits, et cetera.

MR. VILLASENOR: Yeah, that's it -- gosh, it's such an important set of questions.

MS. LEE: I gave you so much. I'm so sorry.

MR. VILLASENOR: No, no, it's nothing to be sorry. It's just a great set of questions. So I guess a couple of things. So many, not all, but many of these algorithms are designed and produced by companies, commercial enterprises that are in there in the business to make a profit. And they do so, in part, because they have a proprietary, or they believe they have a proprietary advantage in keeping -- in their algorithm. And obviously, they would lose so that if they just sort of -- or some of them anyway have a model and they would lose that if they just sort of opened it up to the world.

And so there's a trade -- and, you know, to be fair, the companies do have trade secret rights. And, you know, the need -- the absolute need, which I would agree with, to hold AI accountable does need to be addressed with, you know, do respect for trade secret rights. So many of these companies intentionally, not to be devious, but just because they're trying to preserve their commercial

advantages, don't dump their source code on the Internet for the world to see.

Another challenge is that just complexity, right? These are very, very complex algorithms. And so even putting aside trade secret concerns, suppose the company said okay, here, have at it. Here is, you know, 500,000 lines of code that's deciding who gets a loan and who doesn't. You know, it may not be easy to figure out, you know, what the algorithm is doing. It can be really, really hard even when you have access to everything that's in the code to know what it's doing.

Now, one way around that is you could say well hey, if you're going to market this you've got to give us a flowchart or something that tells us exactly how it works. The challenge of say with that sort of thing, not in this domain but in others, and I have quite a lot of experience actually comparing code to documents that purport to describe what the code does. And, you know, the person who does the flowchart isn't necessarily the person who writes the code. And they do the flowchart in, you know, 2017 and somebody in 2019 writes the code and then they leave and someone in 2020 updates it. And so the problem with looking at the code is it's hard to understand. The problem with looking at the documents describing the code is that they're not the code.

And so it's a really complicated question. The other thought I'll just -- you know, a comment I'll close with is that there's a lot of dimensions to describing an algorithm. If you look at something like, I don't know, dishwasher, you know, you care about things like how much power it uses and how much water it uses, and maybe how much noise it makes, right? But there aren't that many kind of parameters that you need to know to know kinda, you know, what it's going to consume, or use, right, or what its impact is going to be?

And with AI, of course, there's the dimensionality, the problem is much higher. So if we do think about a rating system then there's a really important set of questions. How do we get there, right? What components go into that rating system? But it's a fascinating question and I look forward to continuing to discuss it.

MS. LEE: Oh yeah, we're gonna get there because I told these folks that when we get to this EnergyStar rating that I reported -- I know there's a couple of people who are out there, like, Nicol, I

don't know about that. We're going to talk about it, and just see because this is part of a project that were working on here at Brookings.

Elham, I actually want to switch to you because I mean you're listening, you're at NIST, which is the entity that is responsible for standards, right? So you all don't do necessarily rulemaking that would sort of suggests that there's something that's deceptive or out of the norm for an algorithm. But you actually do a lot of testing, right? And you think about the variables that are part of the inputs and the efficiency as well as performance algorithm.

What your opinion? Because one part of the question is should we just stick with standards, right? And sort of trust the standards within the lab versus trusting the standards how they operate in the real world? Or, you know, what's your opinion about imposing certifications or labeling, or rating systems on algorithmic models?

MS. TABASSI: Yeah, Nicol, thanks for having me here. And good question. So we, as you said, NIST is a nonregulatory agency and we have a broad portfolio of research. And in terms of AI we are focusing on trustworthy AI. And picking up on what John said, you know, what constitutes trustworthiness and different aspects and different technical requirements for trustworthy and how to be noted as trustworthy. But you use the word conformity assessment standard and certifications, and these are -- have special meanings, particularly to us. And I want to go through that, and in going through them answer some of your questions.

So concerning the assessment program models vary based balancing risk and risk (inaudible). The first thing I want to bring up here, and then I'll bring up again at the closing is that one size fits all is not going to work. Conformity assessment is defined in our risk standard and ISO standard, I think, 17,000 as the demonstration that specified requirements relating to a product process system, person, or body are fulfilled.

Now, each of these words are, you know, standard processes. So I think very carefully crafted. But requirements would mean how should it perform? Requirement defines characteristics of the object of the conformity, and it can be expressed in values where a standard is one of them. So one

challenge right now is, for example, you brought up the algorithmic bias. You want to go and -- how to write the requirements for bias? Well, what do we mean by bias? How to measure it and what data to use to measure that? So these are the things that are lacking.

And again, once size doesn't fit all. You measure bias with different use cases, application domains is going to be different. And then, in the finish (inaudible) as Nicol said, so it's demonstration of specified requirements. So if you talk about requirements what do we mean by determination? Determination is that how do we know it performs? So you need some sort of testing mechanism, inspection, or some way knowing that it does what it says.

Again, talking about standard, those things are lacking right now. We don't know how to test for existence of bias or lack of that in a standard way. We don't know that for vulnerabilities of AI. We don't know that for explain-ability of AI.

And then, there is another part at the station that who says that the performance has been demonstrated? And it can be done -- so at the station it's usually a statement made by an organization, it's usually that the requirements have been fulfilled, right? If the manufacturer makes that attestation, it's called the supply of the correction of the conformity. If attestation is done by a third-party then it is a certification. So the first party they're all, you know -- I think these are the details that never have been discussed. So the first-party is when the conformity assessment is performed by the person or organization provides the product or service. This is the seller or the manufacturer. The second party is the purchaser or user, and the third party is an independent entity that has no interest interest in transactions between the first and second party.

So in terms of a certification or conformity assessment the things that are missing are the requirements, or the testament linked somehow to the testing. And then the -- another big thing is that not only do they have a big dimensionality of the different things that has 2B assessed, but also the specificity to the application domain and the use case explain-ability may not be important for many applications but very important for some other applications.

I also want to say something about Mark said about the history of that. So I'm glad that

Mark also -- I wrote it down about successful examples of rating because it all depends on what are you trying to get out of those ratings? Is it confidence? Is it a change of the user behavior? So I think the intent is really good, but it's a very difficult problem to address.

MS. LEE: Well, and I think that, Elham, what you actually demonstrated is why it's so difficult for us to do what John said, no what's under the hood, right? Because you have to run these series of tests. Sounds like what you're suggesting is at the production level, you know, at the user level, and potentially on the independent audit level where there's somebody who actually sees how the algorithm performs, generally, right? It just appears to me -- and Frida, let me go to you because you're on the tech development side. Like, you're a scientist, you're also a person who has a company that is doing employment algorithm, that performance matters, right? And that goes back to what Jason Oxwood said, that if something breaks down on the car, most likely people aren't going to buy it again because they've lost trust in the product. I would love to hear a little bit more about, you know, when you think about ratings, and labels, and certifications the extent to which as a company, that's doing this that it actually matters, or differentiates your product from others?

MS. POLLI: Yeah sure. So we are 100 percent believers that tech companies have 2B trustworthy and earn trust. It's not a given, right? So I'll just start there. And again, I spent 10 years in academia at Harvard and MIT as a before founding pymetrics. And so I come to the problem of AI and machine learning with a different lens, meaning that I have studied the human brain for a decade prior to then thinking about algorithms.

And, you know, ultimately at the end of the day, you know, a human brain has lots of similarities to the algorithms, right? It is a system, it has -- it's a hardware system. It has operating systems to it. And so on and so forth. And so whenever I think about the systems we develop, I also think what's the human corollary to this, right?

So I would start by saying this. We, 100 percent believe that tools like ours and companies like ours should subject themselves to some sort of certification standards and audits. And actually we've already engaged in a third-party audit of our system which, to your point, John, you can,

you know, not reveal your source code, but you can still let a third, disinterested in there to attest that, you know, yes indeed they are doing what they're doing. So I think there is a way around this. It's not something that you can -- it's a -- there's a way around it.

The second thing I would say is that I believe, we believe, that it's really important to audit outcomes, probably more so than process because at the end of the day, you know, and not to say you shouldn't -- there may be some process issues that you want to think about. But -- and again, we are explainable. We provide information about what we're looking at, so I don't -- it's not that I don't believe in explainability, or anything like that.

It's just that I can explain all day long, but at the end of the day it's sort of like having a nutritional label, but then I don't know well, what is the recommended daily allowance of protein and fat, right? The nutritional label is not very helpful unless I have some sort of outcome that I'm looking for, right? So that's just the second thing I would say. So we are big fans of auditing, including third-party audits of the outcomes more than the process. The other thing I'd like to say, which others have brought up, is that I don't believe there will ever be an AI certification system for all AI because it can be used so differently.

And, you know, let's take employment. You know, there are certain things that are best practices among machine learning experts that are actually illegal in employment algorithms. So there just isn't a way that I think you can ever design a tool that says all AI must do this, and it really needs to be sort of application specific which is what, you know, Elhama raised.

And then the last thing I would say is that I think we have to -- so if you think about audits, audits have also been done of human behavior and Sendhil Mullainathan, Cass Sunstein, Jon Kleinberg, and Jens Ludwig wrote a great paper called AI as discrimination detectors. And they sort of start out this paper by saying, look guys, we've done audit studies of employment, we've done audit studies of all these different human decision-making processes and guess what, humans are not doing so great. You know what I mean?

So I think that when we think about auditing and we think about comparisons my favorite

analogy for what we might end up needing in certain cases for AIA is something similar to an FDA like structure. And actually Joy Buolamwini recommended this in her FRT paper, so I'm stealing it from her but I'm going to give it props because I think at the end of the day if the FDA just said well, let's look at drug A versus drug B, but never considered versus -- if we never gave the patient a drug what would that look like? That would be kind of a useless rating system, right?

And to your point of dishwashers, you know, we are comparing dishwashers but everyone has sort of agreed that a dishwasher is better than handwashing. We haven't gotten to that place with AI. We still think that the human process could be better. So I think that if were still having that debate phase, and we absolutely are, we need to include the base case which is like, hey this is what humans do, right? Whether it's in giving out insurance, or hiring, whatever, that's the base case. And then, we can look at AI system 1, 2, 3, 4.

And again, our recommendation is that it be -- for our thinking is that it's best to focus on outcomes more so than process. So I'll pause there, I've said a lot. Thank you for this.

MS. LEE: No, I mean, I love what all of you have said because I think this is where I got stuck as a sociologist because I'm concerned, like, policymakers on the outcomes, right. So there's a lot of debate around should you worry about what goes into the model, like garbage in, garbage out. But at the end of the day, you know, I think people want to see high-performing algorithms, but they also want to see in certain use cases where those determinations do not discriminate or profile, or survey of certain characteristics.

MS. POLLI: Correct, yeah.

MS. LEE: Which brings me to, like, what we've seen in Germany, right? I mean there AI label is a basically being applied to particular use cases which I think they're trying to apply a lot of the stuff that we're talking about. And I'm just curious from all of you, I mean the Europeans seem to get ahead of us on a lot of this stuff, the extent to which, you know, are we -- are they giving us a pathway to follow? Or is this just to prescriptive, you know, in terms of that medicine as to the labeling?

Go ahead, Mark.

MR. MACCARTHY: Yeah. I think that it's a good model what the German authorities are calling for. And it's similar to what the EU is calling for in its in its proposed AI regulation, where they say the high risk algorithms need a kind of premarket clearance. And I don't think a single agency can do this. This is, speaking to Frida's point, about all the issues are really algorithmic specific. So I do think it has to be focused on the particular application, not on algorithms in general, that's much too broad.

And, of course, compared to what is the key question. If you want -- one of the alternatives is humans, you've got to use that as, perhaps, the baseline. And I agree the process has to be like drug approval, and from the industry perspective that really is a much better way to do it because the industry gets to conduct the analysis, right. And that is then subject to a further regulatory review.

For those of you who know something about the conduct of legitimate interest assessments under GDPR, it's similar to that. Where if you're going to use legitimate interest as your legal basis for data processing you've got to do a kind of cost benefit analysis comparing what you're doing with the rights and interests of the data subjects. So I do think that's a good way to go forward. And let me tell you, if you don't do things can go wrong in pretty dramatic ways.

One of my best anecdotes is the software that is widely used to allocate healthcare to hospital patients. One of them got on the market and it was used widely, hundreds of millions of people were allocated healthcare, and yet because of the way it was constructed it was biased against Black patients. Black patients accounted only for 17 percent of the patients that were recommended for health care, but using an unbiased algorithm would have given 45 percent of the patients assigned would have been African-American.

So, you know, conducting these analyses before the fact is crucial. The developer of that product never did it and once they worked with the researcher to fix it they improve the product dramatically. But that should have been done at the development stage. And so there needs to be a kind of risk assessment before these things go on to the market so you can check to see if there's a better way of doing it than what the developer has actually developed.

MS. LEE: You know, I love that. And actually, I want to put a pin and what you're talking

about because that algorithm, every time I think about it, it was like a front-page story, because the variables that were use was the amount of money people invested in healthcare and it kicked out Black patients as a result of that. But I want to hold that thought because I want to see if anybody disagrees with Mark in terms of the German model of actually looking at particular high risk algorithms and labeling them? Do y'all have agreement that that's the way to go, or is there some pushback on parts of that? Does it restrict innovation if we actually go that way? I'm just curious to hear from others.

MS. TABASSI: I just repeat what I said for that one. Well, the first thing is that how to decided something is high risk. And go back to the point of one size does not fit all. You may think that face recognition algorithm is high risk. Certainly, it's high risk for use of face recognition in law enforcement. But if I'm using face recognition to unlock my phone, maybe it's not high risk, right?

So we cannot just -- coming up with the label of high risk and then applying it correctly is difficult. And then, the other thing is that -- the other big question is that whether the needed knowledge base and infrastructure to come up with a robust labeling scheme for AI currently exists. And then, the consequence of that is that how we can ensure that it does not create a false sense of confidence.

So it, again, goes back to what you really expect to happen with these things. If you want to get user confidence and then we ought to be -- to have some good sense of how we're going to do it so that we are not creating a false sense of confidence, or a lack of confidence. So that was just my point about do we have the right scientific foundations to do this at this time?

MS. LEE: Well, I love this. Anybody else? And wait, before we actually go -- Tweet, please folks. AIbias because this conversation is really, actually really interesting. AIbias is the Tweet and if you have questions, Events@Brookings.edu.

Go ahead Frida.

MS. POLLI: Nicol, I just want to say something around innovation. I think I made that point when I was hugely pregnant on the panel that we did together. Look, I definitely think that regulation is an issue. But quite frankly, employment is already highly regulated. So if you're developing algorithms and employment you're already subject to all sorts of regulations. So that's one thing.

The second thing is I actually think that if you look at the Fair Credit Reporting Act of 1974 that is an excellent example that I -- so I'm married to somebody in finance who thinks regulation is, you know, the spawn of the devil. But you know, if you look at that it actually showcases an instance where industry blossomed up because there was regulation put in place that basically then, you know, yielded public confidence in a product, in a whole industry that had not been there before.

And quite frankly, I think we're there with AI. I think there's very low public confidence. I actually think that if we had some sensible, and it has to be sensible, and I've seen instances of what I would consider not sensible, you know, regulation I actually think that it could benefit both consumers, employers, and makers of AI. So I think there's a path forward. It just, you know, has to be done well I think.

MS. LEE: No, no, I love this so far. John, I want to go back to you in respect to -- as I told you I have questions but the way this conversation is going. And it seems to me that -- and I think both Frida, Mark, and Elham are sort of bringing it up. Are we talking about rating systems? Are we talking about labels? Are we talking about underwriting, right? And you've done a lot of great work on underwriting of algorithms just to determine, you know, if any type of bias liability or something that comes out of the model is unpredicted, right? It has an unintended consequence. Do you think that actually should be part of this conversation in terms of the underwriting of the algorithm or are we in the right lane when we started talking about these (inaudible) and basically it's that. You know, the performance of a model?

MR. MACCARTHY: Right. Well, it's a complicated question. I guess one of the things I would say is upstream from all of this, you know, we would all, everyone, or at least all reasonable people would agree that we don't want our algorithms to be unfair, and to be biased. But, you know, one immediate challenge, once we all agree to that is well, what do we mean by biased, right?

And there is different mathematical definitions. There's equalized dods, there's predictive parity their statistical Perry, there's any number of these measures. And one of the challenges -- and reasonable people can make reasonable arguments in favor of or against any of these, any of these

measures. And so, you know, the challenge is sort of how do you make the decision about the sort of metric you're going to use to assess these algorithms and then how do you educate the public?

Like I said, if I'm going to a big box store buying the dryer, you know, I just want to know -- it's simple. I want to know how much power it uses, right? If I'm buying an algorithm I don't want somebody to say well, here's seven pages of equations that tells you -- I just don't want to know that, right. So we've got to figure out on the one way to kind of balance just the inherent complexity of the space, and the lack of simplicity that these questions force us to confront with the need.

I mean the whole -- I would -- and Nicol, you correct me if I'm wrong, but one of the whole point of your sort of desire, which is totally reasonable is you want to kind of communicate. You want to convey, right? You don't want people to have to have a Ph.D. in machine learning to understand if the thing is working or not. And so it's the right thing to do to have these conversations. I just want to make sure that when we do that we sort of look at the full chain. And then once we do that if we can agree that predictive parity is the way to measure presence or absence of bias then it's actually not too hard to do that and you might not even need to get inside the code. You might just be able to put a bunch of inputs in and look at a bunch of outputs and say hey, it's working for these 2, or 3, or 10 groups of people.

MS. POLLI: Yeah, and Jon I would just add to that though, within employment there is a legal definition of bias. So we don't actually have to reinvent the wheel. And actually you would not want to do that because you might actually conflict with the law. So -- and again I don't --

MR. MACCARTHY: Well, yeah -- can I respond to that?

MS. POLLI: Yeah, sure.

MR. MACCARTHY: That's a super interesting point. So in discrimination law I agree with you. Like there's disparate treatment and there's disparate impact.

MS. POLLI: The only thing that's legally regulated by 1007 is disparate impact, just as in a class. So like it's very --

MR. MACCARTHY: Well, I guess -- you know, the courts have traditionally, for all sorts of discrimination law looked at disparate treatment and disparate impact. I mean if --

MS. POLLI: Disparate impact is really by CCNPELOC standards disparate impact is the standard again. We can talk (inaudible) policy who came from EOC but anyway --

MR. MACCARTHY: I guess my point was that it -- point I was trying to make is that as a hypothetical if there is disparate -- if there's disparate treatment one of the ways you can rectify that is you could go and turn a dial to fix it. But in turning that dial you might generate disparate impact, right?

MS. POLLI: Yes.

MR. MACCARTHY: And vice versa by the way. And so I'm just pointing out that these legal frameworks, you know, can collide with kind of the mathematical kind of paradox as you can find yourselves in when you sort of try to fix one problem and find out that all of a sudden jeopardy despite having --

MS. POLLI: but all I want to say, right is that the definition of adverse impact, which is bias in employment is all around whether disparate impact has failed. Disparate treatment is something else you might want to look at but there's no -- Title VII doesn't really talk about disparate treatment. And I just think it's important because there is a lot of confusion out there. Oh, if we do this, then that. If we do this, then that. Actually, in hiring it's fairly clear. So I don't think there's that much confusion in hiring. In other spaces there may be. But I think that there are some clear definitions as per the sort of federal regs.

MS. TABASSI: So if I can jump in. So it's good that you have a clear definition for bias as a disparate impact for the hiring. And I think that's another challenge that that cannot be applied, or expanded to other domains because sometimes the algorithm is designed to not have been an informed impact. You know, car insurance. So you don't want everybody to pay -- so that, again, goes back to the problem of if you don't have one-size-fits-all then that makes it a lot more difficult.

And one other thing that I want to add to what John said about this. A lot of things to consider. Another issue is that they also can take to each others. You improve explainability, you make it a -- key it on accuracy, or security and bias goes hand-in-hand. So how do we do it that it's right. It's -- it's the interpretable and it doesn't create a false sense of confidence?

MR. MACCARTHY: So you --

MS. LEE: So I want to -- Mark, I knew you were going to come in so go ahead and then I'm taking my -- because that's --

MR. MACCARTHY: All right. A quick point. You know, sometimes these different definitions, and they have to be context specific, it can't be one size fits all, as people have been saying over and over again. And sometimes the kind of information that has to be revealed either by a rating system or just some kind of labeling system is very, very specific to the use of the algorithm. My favorite example comes from confounding treatment variables which can be hidden in machine learning programs. The best example is to lower risk of fatal outcomes for heart attack patients who are also asthma victims. This correlation, it arises precisely because asthma patients are at a higher risk when they get a heart attack. And so they get the hospitalization. Now it would be fatal to use that algorithm for treatment decisions, it would kill people, literally. So for some -- for this reason that the researchers really have to, you know, be made aware that that kind of confounding treatment variable might be hidden in the formula, and some researchers say don't even use those kind of formulas for treatment algorithms.

Now, this could be dealt with by labeling. You could have something that says not recommended for treatment purposes. But also, for expert you might need to go to the next layer of detail and tell them what the statistical technique that was used in that system would be. That would be useful for ordinary people. It wouldn't even be useful for doctors. But for researchers, would be essential if they wanted to get a handle on what that algorithm really was all about and how it might be improved.

MS. LEE: Now John, I know you're working on, and I want to go to my model, but I know you are working on a paper like this for Brookings and for your general research in terms of disparate impact treatment. Do you want to just talk a little bit about that? And for the people who are watching, I'm sure it will be available soon. But I promise to all of you, and Frida will (inaudible) back. This part of that conversation is so important that we have to really disentangle that as we look at AI bias.

MR. VILLASENOR: Well, I have a paper in another area, housing, the Fair Housing Act. HUD has just issued a final version of a rule on interpreting disparate impact liability under the FHA. And

notably, for the first time,, I think probably ever, in the rulemaking, certainly in HUD, it has a specific section about how cases where there's alleged algorithmic bias in the types of transactions that are protected under the FHA loan approvals, and marketing and advertisement, and how those are going to be addressed.

And frankly, I find it personally a very concerning framework because it puts a very significant hurdle in front of the plaintiff who basically requiring, as I read anyway, a plaintiff who wants to move forward in a case to sort of put up more information than they would likely have prior to discovery. And so -- but then the problem is if they don't cross that hurdle in the pleading stage then they would, perhaps, never get to the discovery stage where they could actually get the information they, in theory, need at the pleading stage.

And so it's just a -- it's a very important topic both in the housing sector and generally because we see now the government moving, as far as I know, for the first direction to lay down through the rulemaking process rule specifically addressing algorithms. And in this case, one that is not particularly conducive to people -- plaintiff's in disparate impact housing cases to have legitimate concerns about discrimination.

MS. LEE: Yeah. I --

MR. VILLASENOR: And there's a law review article that's coming out in a couple of months on that and I'm sure I'll do a little poster that on the Brookings site at that point.

MS. LEE: It's all good. It's all good.

Go ahead, Mark. And then I mean I want to talk about my model to, guys. So Mark, you have the floor, okay. And then we'll go --

MR. MACCARTHY: One quick comment; is that HUD's final rule went through some of that stuff on sort of a free pass or standard algorithms. And all to the good. I mean the previous thing was just an insult to the cause of racial justice. So I think what they finally came out with might have avoided the worst aspects of what they initially proposed.

MR. VILLASENOR: Well, I made a polite rebuttal to that.

MS. LEE: Yeah, there's going to be a challenge to that one, though. Go ahead.

MR. VILLASENOR: I mean, a polite rebuttal to that is that one of the -- is true there is sort of a carveout, or a get out of jail free card for, you know, if you're algorithms to the standard in some form. But the question is who defines that? And so -- and it also -- and interestingly it sort of allows people to sort of duck responsibility if they can kind of point upstream in the supply chain. And if a carmaker sells a car and the brakes failed and there's an accident and somebody sues the carmaker, the carmaker has a well, not our issue because we bought the brakes from somebody else.

But in the algorithm it appears, at least, that there is a little bit of move to sort of formalize a framework that can allow people to kind of kick the responsibility upstream somewhere.

MS. LEE: Yeah, so I mean -- okay so let me go on real quickly. I think I see another panel coming out of this because I think that's why I'm doing this panel and that's why we're having this discussion. So when I started this work, you know, again going back to my colleagues on that first paper that we did a couple of years ago, it was about well, do we need a housekeeping seal? Because obviously laypeople can't look under the hood. Companies have to compete for the interest in the marketplace.

And at the end of the day you do not want algorithms that are illegal or unlawfully discriminated against you. Nor do you want algorithms that if they cannot perform, for example, facial recognition technologies cannot recognize dark skin complexions, or people like me who last week the hair was curly, and today my hair is straight. Tell me, because that actually impacts the performance.

And I was thinking with that dishwasher example, and I see people on the monitor picking up on this, we have every single rating system when it comes to commerce. We have systems that we comment on in terms of hotels. We let people know what we feel about X. And they'll pass it on the Internet there is no way for you to know whether or not an algorithm performed the same way if you were a woman, if you are a man, if you were African-American, if you were Latina, if you're light-skinned or you were dark skinned.

And to me, that's why I think there's a couple of things that that have come out of it so far.

I think many of you have said we might want to consider a rating system if we are looking at companies to sort of put in their best practices, and we sort of fair them against each other. That's one step.

Some of you have said well, maybe it's good to have a rating system or labeling system because at least you know what's in the ingredient. Maybe it's disclosed that certain facial recognition will not pick up on Black women whose hair changes, or women whose hair changes. Maybe it goes back to what Elham said, right? It's okay for opening your phone, but it's not okay for driving a car, right?

MS. TABASSI: Right.

MS. LEE: And then maybe the third piece which actually I'm trying to dig into a little bit more, what's the feedback loop if you actually recognize that algorithms are generally biased in some way because it represents the values and norms, and the assumptions. If you understand that there'll be some algorithm that will have conformity to civil rights laws which are really clear what you cannot do and others that do not.

But then, you also understand that the algorithm outside of the lab will perform differently on different subjects how can you have rating systems that are maybe generated by individual users, which is my EnergyStar rating idea, that may be more about process. That you're actually able to increase the performance because you're getting feedback that this algorithm is not optimized for diversity. Or it's not optimized for women's pain thresholds over men because the sample is not well designed.

So I'm just putting it out there. I mean, I in no way -- I'm a social scientist. I'm not a scientist, I'm not a neuroscientist, I'm not any of those things. But what I do think matters is that we have to have some way to assess both our private sector developers as well as academics, and users that this algorithm is going to be designed to do what you think it's going to do, or it's not so that people can make choices in the marketplace.

So Frida, I want to go to you. But if you had something like this, I mean, obviously want people to know that pymetrics is like the boss algorithm of employment, and you want to keep competing for that; that's that first level. But what about giving a feedback loop for people to chime in once they

know that you've done all you can, right, to make sure the performance satisfies?

MS. POLLI: Sure. The one thing I will say, if you had to guess how many people think that they're less biased than other people, what do you think the percentages?

MS. LEE: A lot probably.

MS. POLLI: It's 80 percent. It's 80 percent. So my point is that, you know, is a person that studied the human brain for 10 years I -- my only concern is that we don't judge things accurately often times. Especially when the -- yeah, if my dishwasher breaks totally, I get it. My dishwasher breaks, right. If this algorithm treated me fairly or not, you know, I think that might just fall prey to some of the challenges of really understanding whether something -- so I think -- so don't get me wrong. I think there absolutely should be a feedback loop and I absolutely think it should be fair. I'm not trying to say that. But, what if, I don't know, like some algorithm does something that intersects, interfaces with a user in a way that, for whatever reason, they don't like. And therefore they don't like it, but actually the outcome is a better than some -- I'm just imagining, algorithm that goes out of their way to be pretty, nice, and that this, that, and the other. Like to the, you know, visually, but actually is user discriminatory.

So my point is just, I think, that in getting these ratings systems, and I'm not trying to denigrate users, or people as being somehow not capable of writing things. I just worry about what they would be rating. What feedback would they actually -- what thing would they be rating it on. And that's what I would worry about because then I think people would fall prey to deceptive marketing, quite frankly. And we've already seen it. Oh, my algorithm is unbiased, or this, that, and the other. Well, prove it. Well, I can't prove it to you, but I'm telling you that it is. And what if a lot of people were to believe that, and then just give that company good ratings, you know, with no proof. So that's my biggest concern around it would just be like let's make sure that there are they can definitely have objective data that they're looking at.

MS. LEE: No, I mean that's fair, right. We've litigated that on the heel in terms of ratings of our products and services. So I appreciate that.

Who else? I mean when you think about that, and we've got tons of questions, so I don't

want to spend too much time on my research. For those of you that are watching it's in the beginning stages. I've done about 12 or 15 interviews with different companies and developers trying to come up and nuance of this. So please send me feedback at my twitter account or email. But I want to make sure get to everybody's questions. Does anybody else have a response to where we sort of land up on the ratings question or where it might be more necessary or useful?

This is the first time this group has been this quiet. I think everybody's pleading the fifth, which is okay. So I can jump right into questions.

Let me go to questions then. And again, I appreciate the feedback because I think everything that we talked about also has myself as a scholar just sort of thinking about where is the rating most effective, right. Is it a process, is it enforcement, is it at the civil rights side? Is it use cases, so thank you.

So I want to go to Lori Dowling, my friend Lori who always comes to our stuff on AI. And she's particularly interested in employment. Her question is, I do love the idea of a rating system. Thank you, Lori. And I assume, like the dishwasher, you can list the key characteristics and have manufacturers confirmed that they meet them. But as to bias and our use and how they rate, how can we know what's under the hood in that instance? And that's of importance to HR.

MS. POLLI: Yeah. So again, I mean I think that this is where it's all about looking at outcomes, right. And I actually think that, and you know, John, I'm sure in looking at the law you'll appreciate this. I actually think that if you have an AI system in place, it actually is leaving a paper trail of how decisions are being made which, for better or worse that's not actually happening a lot of the times when human beings are making a decision. So, you know, there are definitely management side attorneys that that will say, hey, putting in an AI system, even if it's less biased than your human system is more risky because you're leaving a paper trail. But I actually think that's the advantage. And that's in the claim that this paper makes. You know other than as discrimination detectors is that it can actually be used as a social justice tool because they are leaving paper trails of decision-making which we don't currently have when we rely solely on human judgment.

So to the question just asked, there is the ability to examine the paper trail. I think what is in the way is that it's a very legally fraught area and therefore you might get into an argument between plaintiff and management side attorneys as to if we were to design some regulation that made transparency more at the forefront I think that that -- that we've seen, you know, sort of a clash between two groups of lawyers that tend to want different things. And therefore, how do we kind of get around that?

But I think there are -- I mean, if you look at the equal pay act in Massachusetts, I think that's a good model, right, because it says look at your equal pay, if there is a problem you have a grace period of a couple of years to be able to fix it and I think unless we have something similar to that where you are encouraging people to look at this data, encouraging people to look at how these algorithms are impacting folks, and then, if there's a problem fixing it, I think that's -- I think that's a good model for how you could move forward. That's just my opinion, obviously.

MR. MACCARTHY: I guess I would just chime in. Like, I agree with you. You have really good points, and I'm skeptical of the assertion that hey, a paper trail is a bad idea because it might create more liability. I think it --

MS. POLLI: Yeah, I don't agree with that. I've just heard it, yeah.

MR. MACCARTHY: Yeah. If the paper trail can help identify and expose biases that would otherwise have gone on detected then yeah. Maybe at the beginning there would be some people who would be held liable for it. But the system would sort of recalibrate to correct for those things.

MS. POLLI: You're preaching to the choir. You are preaching to the choir. But I'm just telling you I have heard management side attorneys --

MR. MACCARTHY: Yeah, no. I'm sure you have. But their job is -- they view their job, perhaps when they're saying that is just protecting the company from litigation. Whereas I think were standing back and we're saying, how can we broker fairer systems to --

MS. POLLI: But when the rubber meets the road, it's those management side attorneys that are then giving guidance to these companies as to whether they should continue with a clearly

human -- a clearly biased human process, or potentially put in a unbiased AI system that occasionally might actually trip the biased wire and that would leave them more legally liable.

MR. MACCARTHY: A poor combination, right? I mean if you --

MS. LEE: Yeah.

MS. POLLI: Yeah.

MR. MACCARTHY: -- a combination of human informed decision, you know, informed by --

MS. POLLI: Absolutely. Absolutely.

MR. MACCARTHY: -- AI.

MS. POLLI: I'm just telling you the challenges we've seen in trying to implement less biased technologies that people -- these are the arguments we've heard.

MS. LEE: And it's hard. I mean overall is generally hard to actually have this conversation and to sort of have this conversation between technical experts as well as lawyers, as well as sociologist all sitting around trying to figure it out.

The paper that I referenced took almost a year to actually do for this very reason. So I love the fact that I'm simulating the reason why this is such an important conversation to have.

Elham, Henry Webber has a question for you. His question is, could you discuss how NIST recently introduced more principals for explainable AI could be utilized in different rating systems proposals. And he recognizes that evaluating outcomes helps to work towards accountability, but believes that explainable AI might help bolster consumer buy-in to a standard or service.

So hopefully you got that.

MS. TABASSI: Yeah. So thank you for the question. The paper that we put out, yes, we put the paper out for the principal of explainable AI and that's out for comment by October 15. So please let us know your input, thoughts, and comments about the paper. Explainable AI is a -- the explainability of AI is a difficult and complicated topic because you don't need explainables for every possible scenario for recommended systems. For example, that's telling you and telling me which movie to watch based on

my history of movie watching. I really don't need explanations. But there are some applications that you do need explanations and particularly explanations might play a role when something goes wrong, and you want to do course corrections.

Not to mention that there are also some legal requirements for explainability. So what is explainability? What expectations around an explainable AI systems? What's the expectations of the explainability to be correct, to be within the scope of the model that was trained for the particular end and particular use? The paper talks about all of these things.

And we use the word explainable, but we are also very aware that explainability and interpretability are -- that these two terms are being used differently among different communities. And so you may need -- back to what John was saying at the very beginning, you may not need to know exactly what's going under the hood. It may not be useful information to a lot of people interacting with the systems. But it might be very important to the computer scientists and the people that are developing those models.

Those are explainable. You know, figuring out what's under the hood, how the model works. And the interpretability is how to get the outcome and how to talk about how this prediction was made. How the recommendation was made. What factor was involved in making the outcome of the AI system.

So that was the paper, and this is a true consultation. It's certainly not a complete -- not completely (inaudible) of the explainability so we are really looking forward to community engagement.

How it relates to the labeling, so it, again, really goes back as it was mentioned, there is a lot of things that a label or a rating system can achieve and today a lot of those things have been mentioned. The label can just go and rate is it safe to use? Is it biased? Is it reliable? Should I use it for this particular application or for other ones?

The explainability can help with information about how the decision was made, and if there is any of this different intent of the labeling system that was discussed. It can help -- it may be help with that. But again, I think we are trying to pack too many things under the label or rating systems, and

these are all important discussions. I'm just trying to figure out who to do that.

What you said about an explainable paper of one of the principals is knowledge limit. And Nicol, you mentioned that we want to have something that don't use it for this purpose, right? Or it -- this facial recognition does not work for this particular populations. I see them different and certainly it's not rating, right? So it's just extra informations and I think those are very -- those can be very useful.

I don't know if I answered the questions, but that's what I have to say.

MS. LEE: And that paper is available at NIST, right, Elham, so somebody can get that paper at NIST?

MS. TABASSI: Yes. Yes. I can send you the link if you want to.

MS. LEE: Perfect. Now Mark, I had a question for you and then I'm going to wrap up.

Regarding your reference to the second wave of AI accountability, there are clear reasons for investment in AI safety and management. What would you say are the most concrete reasons, and how do they compare to market driven incentives for AI development and deployment; that is, sell as many units as possible and focus investment in R&D in America. AI stakeholders seem to be more incentivized to invent and sell that to test and regulate. So our friend Tom Wheeler who was in CIT Department, which I run, always says that it's about breaking things now, and fixing it later, is basically a summary of that question. So are there incentives to actually get folks to start thinking about these types of best practices before we jump to regulation?

MR. MACCARTHY: Are there incentives? I'm sure there might be. And I think we should explore them. But I'm skeptical that any of them will do any good without the weight of government action behind it. Or at least the threat of government action behind it. The example I showed before of getting the health management algorithm out into the marketplace as fast as possible suggests that marketplace incentives really are pretty powerful in this circumstance.

And the thrust of tech development over the last 20 years, especially, is really, you know, get it out there as fast as possible and then see if it needs to be fixed. Well, you know, for a recommendation engine, why not? But for a healthcare management algorithm that's just the wrong style

of development. And I do think without some sort of government management role the market place would tend to move in that direction with dangerous results and results that I think could be avoidable with a sensible and thoughtful role.

MS. LEE: Well, and in 10 seconds Frida because this actually was to you in terms of then the laws that we have, are they sort of imperfectly designed to stave off some of the threats that we talked about today when it comes to AI?

MS. POLLI: Oh, I don't know about that. I would say that. I think they can always be improved upon, quite frankly. But I would say that existing law is problematic. I think that there's always room for improvement, but I actually think it's a really good start. So no, not at all.

MS. LEE: Okay. So John has a paper coming out, we heard from a tech developer who's out there actually trying to figure out how to make her products comply with employment laws. You've heard from the person who is standardizing this technology. And then, I love the way Mark is just sitting back listening to everybody and offering his advice as he does so well when it comes to really thinking about the evolution of technology from his role in the software association to what he does now at Georgetown. And even the stuff he contributes to Brooking.

Friends, this is an issue, and it's a challenge. And I think we would not have had this had we not realized that that as we begin to see just how much we have of people's currency of data, that it matters, right. And it matters in ways that that will create either productive processes or best practices on the technical side. Or, it has potential to engender more harm or create more deceptiveness.

At the end of the day, I swear I go back to my dishwasher model. That at some point whether it's AI across industries or within similar industries, or with consumers, the people, I'm still trying to figure it out. We need some way to either gather data from individuals that this AI is optimized to their context, or we need to let people know that, or let people tell us that it's not.

Because my fear is this. It's not so much of that AI is not doing a kick butt job when it comes to human -- improving upon human decision-making, but as Frida said earlier it needs humans in the loop to ensure that there are people who are not left behind in terms of variables that are maximized

for just to include, you know, their attributes.

So with that in mind, please follow TechTank, our podcast. We actually have a podcast on that now, on racial bias in tech, as well as a whole lot of other podcasts of interest to this community. Please continue to follow us on our AI bias work at Brookings. Follow us from the homepage, hit AI, you'll see a whole host of papers and all the things that we talked about, including several people on here. And please, I think some of you heard my dog, if nobody else cares about my rating system she did.

And so with that, I would just say have a great afternoon and we will see you back at this place in Brookings space around these kind of conversations. Thank you everyone for giving us your hour. And thank you everybody.

MS. POLLI: Thank you, Nicol. Thank you. Thanks everyone.

MS. TABASSI: Thanks you Nicol. Thank you everyone.

* * * * *

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