RAW FILE

ITU

AI FOR GOOD GLOBAL SUMMIT

GENEVA, SWITZERLAND

DAY 2, 16 MAY 2018

14:00 CET

FOUR AI BREAKTHROUGH TEAMS - THE HEART OF THE SUMMIT

ROOM K

BUILDING TRUST FOR BENEFICIAL AI - TRUSTWORTHY SYSTEMS

Services provided by:
Caption First, Inc.
P.O. Box 3066
Monument, CO 80132
1-877-825-5234
+001-719-481-9835
www.captionfirst.com

***

***

>> LADIES AND GENTLEMEN, THANK YOU FOR RETURNING. Apologies to those who have been sitting here for half an hour already because it says in our booklet 1:30. Will you see the program has been updated, so hopefully you were following the one on the app or website that says we are getting going at 2:00. Or if you have been here for a half hour already, I hope you got a chance to catch up on emails.

The new schedule is from 2:00 to 2:30 we will be having the third and final session in which we will be presenting our ideas. Then there will be a coffee break from 3:30 until 4:00. Then from 4:00 until 5 back in here there will be a panel where you can hear from UN organizations who are facing building and earning trust on the front line. From 5:00 until 6:00 will be the breakout session, and that will be downstairs directly below here in a kind of open space that you will see when you come into the main building and you go through the entrance barriers, there is a big open space behind there. That's where the breakout sessions are. If you do get lost after coffee, don't make it back for the panel session, please do come to the breakout session downstairs and go straight downstairs at 5:00 for that because that's when we really hope to harness your ideas

and input and suggestions for how we can take these projects forward and develop new ones.

Also just a reminder that tomorrow at 9:00, back in the Popov Room in the plenary session, all of the tracks will be reporting on what they've done today. We'll be reporting back, but also we will be launching the TrustFactory, which is the overarching idea which will spawn hopefully not just these nine projects, but many more besides.

Right. Now I will hand you over to Francesca Rossi, who is going to chair this session. Thank you.

>> FRANCESCA ROSSI: Okay. So this is the session describing the third team, the third angle with which we wanted to look at the issue of trust in AI, and this is about trustworthy AI systems, so systems that should behave in a way that can generate the correct level of trust in the user or whoever, the human working together with the system.

So again, as in the other sessions, we will have three -- description of three projects, and as we said this morning, if you have other ideas of other projects, come to us during the breaks or later during the day or tomorrow or even on the TrustFactory website because you are welcome to start also other projects. But today we are going to hear about these three very specific projects that have to do with building trustworthy systems.

So the first one is about bridging the policy technical gap for trustworthy AI, and Jess Whittlestone from the Liverhulme Centre for the Future of Intelligence is going to present it.

>> JESS WHITTLESTONE: Thanks, Francesca. This track is about ensuring that AI systems are demonstrably trustworthy. The way that is to be done is technical, and involves ensuring these systems are built in a way that means we are able to see that they are fair and we are able to verify that they are reliable in different situations. But I also think that increasingly making sure that AI systems are demonstrable trustworthy is for policymakers too. So more and more we are seeing we need discussion around policy around AI and how AI is trustworthy.

I think what has existed in the past is there is a natural gap between the kinds of people working on policy challenges and the kinds of people working on the technical side. And it seems to me that in order to ensure the policy is really effective at regulating and kind of designing all kinds of policies around ensuring that they are trustworthy, there needs to be collaboration and the ability to build off what's been done on the technical side. So I am going to talk a little bit about why I think that this is a problem and how we might start to try to bridge this gap between people working on the policy side of trustworthy systems and people working more on the technical side.

To start with, just take a step back and sort of why do we need AI policy? One reason that we need AI policy is we can't put all of the responsibility for ensuring AI is trustworthy in the hands of had the developers and in the hands of sort of the users. We need some kind of overarching structure to really ensure that that happens

because it's incredibly complicated, and we can't expect individuals to be making these kinds of decisions for themselves.

And more specifically, I think where we can think about a couple of things that policy can do to try to ensure that AI is trustworthy. po policy can shape the way technical is developed and the way that it's used, and it can also shape the environment that the technology is kind of deployed within to ensure that that environment is able to adapt to changes. Really when we think about trustworthy AI, it's not just the systems in isolation we need to think about, it's the way in which those systems interact in which the environment in which they exist.

Just a few examples of how this is already kind of starting to take place, policy around AI, so DDPR, which we have heard a lot about recently is coming into effect in a couple of weeks. A large part of what that does is it shapes both how algorithms are being developed, it puts certain -- you know, it's this idea that algorithms have to -- that there might be a right to explanations is then incentivizing people developing algorithms to think much more about transparency. And it's shaping how they are used and the way that people can challenge them.

Then other kinds of policy, so obviously, innovation pushes the development of AI forwards, so we might want to think about whether that's going to -- it could lead to benefits but also might lead to challenges when we think about trust.

Then more on the sort of side of shaping the environment, we have

all these talk about policies around education, around the kinds of skills that need to be developed, and around the way that the economy and infrastructure needs to change in order to ensure that the impacts of AI on the environment are positive.

So then we can ask how might A shtion policy go wrong, how might we fail with AI policy to ensure that we get the benefits and mitigate the risks of AI and ensure it's sufficiently trustworthy? One way it might fail is we must trite AI too little, we might overregister it, which would then stifle innovation and other things people have talked about at other parts of this conference. I think one thing that's sort of interesting that I have been thinking about in relation to this, this is maybe particularly likely to be a concern if regulation is an entirely separate thing, separate from the sectors in which AI is being deployed, because that will mean that the people -- if your role is just to regulate AI, then for you the incentive is to mitigate all the risks. Where if regulation is done in a sector-specific way, then the people who own responsibility for regulating also own the benefits, so they have to think about balancing the two.

A second way in which AI policy might file is we might kind of on the opposite side fail to mitigate risks and manage challenges as they arise, which we can think of as trusting AI too little -- too much. Too much. And you know, in particular we might worry about things like failing -- policy failing to ensure that the technology is kind of sufficiently transparent, sufficiently

accountable, sufficiently fair.

Or we might fail to adapt our environment. It might be the technology in itself is not directly harmful, but it has all these consequences on our environment and we aren't able to think ahead, see, and adapt to those challenges.

Which brings me to this broader point and I think the biggest challenge for policy in avoiding both of these failure modes is anticipating new challenges, especially -- and I think this is a way in which AI is different from all kinds of other science and technology policy we've had in the past, which is that it's moving very, very quickly. And it's very difficult for policy to kind of stay ahead of that and to see the new challenges. So there's this real worry that we end up in a situation where policy is really being very reactive to specific, narrow things. We might also be concerned that as a result policy ends up kind of reacting to whatever is most in the public attention at the time, and this sort of relates to thinking about the narratives that we see around AI. We really don't want a policy and regulation of AI to be overly driven by the hype and the concern because then it might not be -- driven by the wrong things. And so I think in order to ensure that policy around AI is successful, we do really need to be able to think in this very broad, general way about the way increasingly as it advances AI is going to affect various different parts of the society, various different parts of the environment, and what kind of general policies and governance structures we need in order to be able to mitigate the

challenges.

And in particular, I think we need to be asking a bit more this question of where are our AI capabilities today? Where might they be at various points in the future, in the very near future, the medium-term and long-term future, and how will those capabilities interact with different parts of society? How will they affect the economy? How will they affect law? How will they affect the way that various public services are delivered, and what kinds of challenges and opportunities does that range? Obviously that's an incredibly broad question that I think requires input from a very wide range of stakeholders and, in particular, I think it suggests that policy needs to really draw on technical expertise because I think being able to sort of anticipate these future challenges posed by policy really requires an understanding of where AI is at today and where it might be in future. But on the other hand, this isn't a question that can be answered by technical experts alone because, you know, a lot of people -- machine learning experts might be working on a very specific aspect of the technology and not be able to see the wider picture of how it might impact different parts of society. So it's for this reason that I really think, especially with AI as this fast-moving technology, we need collaborative discussion from both sides to try and answer these kinds of questions.

But inevitably, there's a substantial gap between the policy and technical communities thinking about AI. There are very few senior policymakers who have technical expertise. Most people kind of

working more on the technical development side of AI aren't necessarily engaging with policy and ethics issues, although I think we are seeing this start to change. And I think the biggest problem is that the two groups don't really know how to communicate with one another. They are speaking quite different languages. We've had a few comments today and maybe yesterday about a lot of these terms that we are using, like trust, fairness, privacy, are quite ambiguous and interpreted differently by different people, and I think in particular, used differently by, like, you know, a policymaker, a lawyer, and a technical expert. And so there's a bit of sort of people talking past each other and not necessarily working on the kinds of solutions that the other group need.

So the sort of aim of this project and the thing I have been thinking about is how do we start to bridge this gap, particularly focusing on this sort of emerging group of people within policy who are designing policy specifically around AI, specifically around AI and these kinds of things. And I think the most important thing is that we need to find ways to sort of improve communication between these -- between groups of people working on policy and ethics around AI and technical experts. So we need to start asking these kinds of questions of like what are the main barriers that we face in these two groups communicating, and maybe to begin with like how much do people on each side actually understand about what the other does? Because I can imagine a lot of the problem speaking to some friends I have who work more in technical research, they don't really have

a good understanding of what policy is or what it can do. And on the other hand, I think people working in policy don't necessarily know what research day to day looks like and what kind of problems they are working on. They obviously don't understand the technical details but even the day-to-day what kind of problems are people working on and what can they each do for the other?

And I've also been talking about with some people and thinking a bit about what kinds of sort of translator roles might be helpful here because often I think when there's a gap between two very different communities with different expertise, sometimes it's really helpful to have people who sort of sit in the middle and can bridge that gap, and I think there's certainly stuff that can be learned here from -- in the UK government, which is the one I am most familiar with, there is this -- we have a network of scientific advisors throughout government who -- people with kind of deep scientific specialist expertise, but who sit closely within government departments and advise them on these things. So I wonder if there's something we could do like that but more in this kind of specific AI policy space.

I think we also need to provide more opportunities for technical experts to engage more in policymaking and with ethical and societal issues more generally. In the UK we are kind of -- there are suddenly seems to be sort of a lot of momentum around policy and AI with a few new centers and things being set up, so I think this is potentially a really good opportunity to sort of try and start drawing more on

the technical expertise and sort of navigating the challenges that that involves. And there's also a question, one of my colleagues at CFI talked quite a lot about, sort of what kinds of ethics training do we need for technical researchers? And one thing that's really challenging here is that there isn't a sort of standardized AI accreditation in the way that you might have for doctors who, you know, you have to get a standard accreditation, and ethics is part of that. So there's this big question around what we do with that.

And I think there's also this question of what kinds of technical understanding does someone setting policy around AI need to have, and how do we provide that? Obviously, a senior policymaker isn't going to be an expert in deep learning, but what kind of things do they need to be able to understand to answer the right questions I think is an important question. And there have been quite a few initiatives that I am aware of trying to provide both generally technical training or digital skills to policymakers, which I think is really important. But in the case of policy and regulation around AI in specific, I think we might need to be a bit more targeted and ask, well, who are the key players, and what kinds of things do they need to understand? Because this is like a huge wealth of things out there within sort of digital that you could possibly be trying to teach people.

So one of the things I have been doing in preparation for this project is just trying to scope out what are the things already being done broadly in this space, particularly in sort of trying to bridge

this technical policy gap very generally as opposed to specifically with relation to AI policy. So as I mentioned, there are quite a wide range of initiatives sort of providing technical and digital training for policymakers. There's -- so Laura James, who is here, who works for dot everyone, they have been doing some really interesting stuff with sort of I suppose more targeted digital coaches for MPs. So an MP will have sort of someone who works with them very, very closely and helps them to navigate digital challenges that maybe they don't necessarily no how to navigate themselves.

There are these data ethics frameworks within government, so the UK government have been developing frameworks. I suppose the idea with which is to help policymakers who are using AI in their policy area to ask the right kinds of questions to ensure that their use of AI is reliable and ethical and safe rather than them needing to sort of really delve into the technology in detail. It's like what questions do you need to ask as a non-expert? I think there's some really interesting work to be done around trying to test those kinds of frameworks further in the field. And I know that some of the people I have been speaking to in the UK government who are developing these frameworks are very interested in having them tested extensively. If anyone is interested in doing that, if anyone has a data science or AI-related project that has a sort of important ethics component and you would be interested in seeing how well this framework works, then definitely talk to me about that.

And yeah, there are all kinds of other things. I think there's

a lot that we can draw on from the existing structure of scientific advisors in government. Some of the existing ethics training for researchers, and the fact that like the -- some of the technical industry labs developing AI are increasingly very recently establishing themselves as ethics and policy leaders as well in establishing these groups.

So what's the goal of this project? I mean, at the moment, this project is very broad, and so I suppose I would say in exploratory phase, and I am still definitely looking to better understand the sort of initiatives out there, especially in a more general sense, like how with other technologies in the past and how with sort of science and technology policy in general, what are the ways of drawing on technical expertise that have worked well and that haven't worked well? And I would be particularly interested to really hear about sort of international perspectives on that because, as I say, like my expertise and my experience is very much with the UK government, but it would be very interesting to hear from other countries if you have any sort of experience with how that kind of policy technical collaboration expertise works and what things you know of that work well.

And so I think the next step with this is working with a few partners to try and start having -- maybe running some workshops or start having some conversations that bring together policymakers and technical experts, particularly to try and sort of bridge some of this language gap. And one thing we were talking about earlier which

I think might be interesting to focus this a little bit more is to take some of these words that are being used and thrown around a lot with respect to trustworthy AI at the moment -- so you have trust in itself, you have the concept of fairness, the consouthwest of privacy, the importance of transparency, and get some people from different perspectives, specifically policy and technical, together to kind of talk about what those terms mean to different people and how they are used specifically technically and how that snaps on to what policymakers care about and try and sort of, I suppose, disambiguate those terms a bit and try and come up with common I with as of talking about them.

But ultimately, I think the real aim of this project over the next year is to try and come up with concorrect policy proposals or other proposals for how we think we might help bridge this policy technical gap. So I mean, one example of that might be a kind of specific proposal for a specific type of translator role in government that works with policy teams and experts in the technical fields to sort of answer these really tricky policy questions.

Yeah, so I covered that.

Yeah, so we've got a few current partners we are sort of talking to. There is this Laura, trustworthy technologies project at Cambridge, very interested in this. We are also talking with Bennett Institute for Public Policy, which has recently set up in Cambridge too. And also the COO of a data science company that works very closely delivering solutions in data science for government. But

as I said, I am very interested in sort of more ideas and more collaborators, especially with sort of international perspectives. Even though I suppose we might start initially for this year we will probably focus this project on the UK because it's pretty broad as is, but I think it would be really useful to be able to draw on what other countries have done and what might be helpful.

Yeah, so would definitely welcome ideas, suggestions, and I put. Come to the breakout session or speak to me afterwards. Yeah. Thank you very much.

(Applause)

>> FRANCESCA ROSSI: So I can ask one of the questions that have been posted up. So one ends with a smiley, so I think it was ironic, but I think I will ask it anyway. So it just says what about using AI to make these two communities talk to each other? You know? And I put this question together with another one, which doesn't have the smiley, so maybe it's more serious, that has to do with -- it says why not pursue AI standards as the multistakeholder bridge between technology and policy?

>> JESS WHITTLESTONE: Sorry, can you say that second one again?

>> FRANCESCA ROSSI: Yes. So using AI standards to be the bridge between technology and policy.

>> JESS WHITTLESTONE: I am not quite sure I understand what AI standards means.

>> FRANCESCA ROSSI: Mark Jeffrey, do you want to say something?

>> Mark Jeffrey with Microsoft.

Yeah, in the same way that we have security standards for security, information security management systems, we have standards around quality, we have standards around all sorts of things. And those are regularly understood in government to indicate an industry consensus on what is good practice and a good way to go forward. And for things like security, they can cope with highly changing environments, where you know, we don't have to rewrite the rules for every new breach because the management system process describes how you deal with that and how you report it and how you get better. That's what I mean is there is an international consensus across multistakeholders, which governments can then recognize and use rather than trying to invent it new in every country.

>> JESS WHITTLESTONE: Yeah, I think that's a really good point and also it might also be that we don't kind of yet have those standards totally established for AI, and the task of establishing those standards is probably going to be a collaborative effort between different disciplines and sectors, so that might also be a good focus for bringing these kind of perspectives together. But yeah, I agree, if we could have these kind of agreed-upon standards, that would maybe solve a lot of these challenges so that policymakers are not having to do all of this thinking on their own and anticipating things.

>> Just getting the same terminology across the different policymakers and other industry players, so they are using the same words to mean the same thing, is incredibly important.

>> JESS WHITTLESTONE: Yeah, exactly. I totally agree. That's why I was saying I think one more concrete way forward for this that might be useful is taking some of these terms that have been used and used sort of ambiguously and trying to make sure we are all talking about the same thing. We had some discussion at the CFI a few weeks ago where one of the postdocs who is working on technical mechanisms for privacy was sort of talking about his work and presenting like very technical definitions of differential privacy, and then there is this question of what extent does that map on to the kind of privacy that lawyers and policymakers are talking about? And it wasn't clear. Yeah, there was some uncertainty there. I think taking some of these terms and taking some of the different ways that they have been used and sort of pin down more precisely, both within the technical literature and within law and how policymakers are talking about them, could be really useful.

>> FRANCESCA ROSSI: So let's take one question from the audience.

>> Kind of following up on the previous question by Mark Jeffrey, the IEEE, that is the biggest professional organization amongst engineers, is developing standards like this. But the problem with professional organizations of these standards is that they don't want them to be enforced. It's more like guidelines, development guidelines. For the ethical context. Of course we have standards like in telecommunications or all kind of other engineering aspects that are quite (?) because it's the only way companies can manufacture things in a similar way. But I think there is a lot of

aversion for regulation, if you want, and that might be an obstaclele for the link between policy and standards

>> JESS WHITTLESTONE: Yeah, I was actually talking to someone at lunchtime just before this, he was talking about the IEEE developing these standards now, and yes, definitely something to look into.

>> FRANCESCA ROSSI: Yes, so maybe while I read the questions, someone can ask a question.

>> I just wanted to add to the points that were just brought up. Of course it's really important that we can understand each other and communicate, but the drivers of uncertainty in some of these words are -- go back a very long way. People have been talking about fairness for hundreds or thousands of years deliberately being somewhat vague because you can promise as a politician fairness to everyone and people perceive it the way they want to hear it. That applies also to some of these terms we are talking about. Of course it would be useful to articulate different categories because we are going to want different sorts of trance partisansy to different people in different contexts, the same fairness to privacy. It's helpful to say we want to try to articulate these, but I think that needs to be part of the process of working together to talk about what we want. Once we figure out what we want, we can to some extent call it what we want and a label will be useful to articulate it, but we need to think carefully about how these context also apply in different settings.

>> JESS WHITTLESTONE: Yeah, definitely agree. And like you say,

one of the challenges is that the uncertainty in these terms is not -- it's not just that technical people and policy people are using them in different ways; it's that they are inherently ambiguous. It's important for us to talk about what we want without bringing these terms in and one person means one thing and another means another thing, but having these conversations where we get a bit clear about what we mean. As you say, because we all wants different types of fairness, transparency, privacy in different contexts.

>> FRANCESCA ROSSI: Okay. Let me take another one of these questions that has two votes. You want to ask a question?

>> So what I heard from your talk, which I like a lot by the way, is that governments and big players are asking the right questions, do they have ethics committees, do this kind of stuff, but we still have a lot of smaller, medium companies that also jump the AI new technology bandwagon, how do we get those guys on board? At the end of the day what we want to have is a universal norm. It's like food. We don't want to have just the big companies produce good food. We want to have everyone produce good food. So how do we achieve that so that even the medium non-Google, Microsoft software companies produce ethically sound software?

>> JESS WHITTLESTONE: Yeah, I mean, I think it's a difficult question to which I don't really know the answer. I mean, I think that if we have -- if we kind of develop policies and standards around these issues, I guess the idea of developing policies and standards and regulation to some degree is that it should incentivize companies

to behave ethically and the way we want them to. It should -- policy well designed should make it such that it is in the small companies' interest too to kind of adhere to these standards. So I think part of it is about thinking carefully about the way we design these standards, and it not feeling just like, oh, we have to adhere to this regulation and more -- yeah, I mean, that's a very high level. So I used to work -- I did quite a bit of -- my background is in behavioral science, and I did quite a lot of work sort of translating -- working on how you can use behavioral science to design better policy. I think maybe there's some stuff you can draw on there which is like what ways of sort of framing these standards really kind of appeals to the things that companies care about and their interest. So I suppose not designing regulation in this way, not kind of having overly sort of excessive regulation that just seems like it's putting all of these burdens on small companies that have, like, not large amounts of resources to deal with these problems, and maybe kind of trying to go frame the discussion of these things in ways about kind of being responsible, doing good, making sure that your technology has benefits. But yeah, it's a difficult question.

>> FRANCESCA ROSSI: So let's take one last question.

>> Yeah, hi. (Off microphone).

>> I think the human rights has to be one of the values or standards we are thinking about constantly when we are thinking about the implications of AI. So when I talked about this idea that we need to -- we need to be sort of thinking a bit more ahead and anticipating

the impacts of AI on society in the future as it develops, and I think maybe one of the questions we need to be answering -- asking ourselves and maybe kind of thinking about harder is like how do kind of current and potential future AI capabilities impact human rights and what can policy do to sort of ensure that that's positive. So I suppose the way I think about it is a large part of sort of developing policy around AI is thinking about what are the values that we think are most important to make sure our use of AI adheres to? What are the ways that AI might threaten those values, and then what do we do about that? I suppose that's how I see that fitting in.

>> FRANCESCA ROSSI: Okay. So let's thank Jess again and remember that you can come and discuss about this project again during the breakout session. Thank you.

(Applause)

Okay. So the next project will be presented by Rumman Choudhury and Sebastian Vollmer. It's about trustworthy data: Creating and curating a repository for diverse data sets.

>> RUMMAN CHOWDHURY: Hi. I am Rumman Choudhury, the global lead for responsible AI, here with Sebastian from theturing Institute to talk about trustworthy data.

Discriminatory practices and outcomes of artificial intelligence, and they have been very popular in the media, and I am just going to name a few. So one is sexist natural language processling. When trained on pretty much the Internet, a natural

language process algorithm that manage is to programmer as woman is to homemaker. So that tells me looking at the data is readily available, in other words, the Internet and being able to scrape it and use it to train our algorithms, leads to discriminatory outcomes. A second one is a paper called "men also like shopping" trained on publicly available images, the image associating women with kitchens with a so strong, if you showed a picture of men in a kitchen, it identified them as a woman. Again, training on publicly available image data led to discriminatory outcomes.

Timely, a friend of mine at the MIT Media Lab works on a project called the coded gaze, and this is about facial recognition algorithms. What we find with them is that the overwhelmingly -- they overwhelmingly work correctly for people who are lighter skinned and people who are male, but there are highly disparate outcomes, and the worst outcomes are for darker women of color, and this comes from in part a lack of diversity in the data sets. So the data sets used to train them, again, overwhelmingly white, overwhelmingly male.

So what does that mean? That means that when data scientists create projects, we often rely on what is out there in the world. I was teaching data science before I joined Accenture, and often my students would come with a question and we would scour the Internet for data. Even at Accenture, we cannot use client data for other purposes, so we, too, are limited by what's out there.

>> And maybe I come in here. I spend a lot of time trying to get

data for different projects. If you have everything lined up, you have exceptional talent, but then data access might be delayed for various reasons. And this is the case, much of the focus of AI is on things readily available, but might not be the best use of the exceptional talent of what's out there. I think open data is great, a lot of projects, software platforms where people publish their open data, but these come with caveats alluded to, and also the amount out there is limited for the problem at hand. And even finding them is quite hard, and convincing people to make their data open is very hard. You have to tell them a story about what the negative outcome is. It's not necessarily what the positive outcome is. This doesn't convince people. It's more about what could be prevented if this data would be happened to this time?

>> RUMMAN CHOWDHURY: What we are proposing is to create a public repository of what we call trustworthy data. There's a lot to unpack in that statement alone. There are massive problems in just getting data, and even when you get data, often it was built for a very specific purpose. I will tell you some of the most common places we get data, for example, Cagle, where they run competitions, the shtion CI machine learning repository, and these are useful because they have data structured a very specific way for a very specific need. However, that's also the problem. If I go to Cagle and I use shopping data that's using income and gender and race and ZIP code and I use it for something else, I am not necessarily aware as a data scientist what are the errors and biases that might manifest itself

in so our project is ultimately one thing, but it has two parts. The first is this repository of data, and what that means, as Sebastian mentioned, this diversity of data which we are hoping to gain from people in the audience and partners, so different kinds of data, not just that lives in spreadsheets but maybe also unstructured data, like video images, as I mentioned, sound, music, noise, et cetera. And also data that has labeled contextual problems and biases that may come up. So this might mean understanding the process by which the data was created and having that be transparent but also listing cultural and social biases that might manifest themselves.

For example, you could have all of the hiring, salary, and promotion data at many of the major tech companies, and it can be all entirely perfect data, but the outcome of your algorithm would most likely be discriminatory and biased because the world is a discriminatory and biased place. So giving the context of the data and the discrimination and bias that may manifest itself as a result is also quite important. So what we are trying to build is a tool for data scientists to use not just to collect data, but to start thinking about trust and ethics in the very first steps of their processes.

>> One point I wanted to touch on, even the definition of what is trustworthy data, is to be determined. There are standards in particular industries, mouse clinical trials, very clear protocols to follow. This is not the case across the industry, and I think the AI hype has the disadvantage that many people out there think

there shouldn't be taking care any more how data is collected. A basic principle of statistics, business intelligence, just put to the side with the hope that there is a magic wand of AI which will still make everything today. I think still to this day garbage in is garbage out, and this is something where careful recording of the data collection mechanism might help to correct for biases that are present in the data, so one can build observational models, one can build contextual models. The actual truth out there, but they rely on good data and also not just dumping data.

>> RUMMAN CHOWDHURY: To piggyback, a more interesting part is what is the framework for creating the data. We want to make that process publicly available as well. Myself, somebody at Accenture, I want to use this on a data set a client has given me, I want to be able to do that as well.

We are looking for partner organizations to provide this data, possibly provide funding to put this repository where we create it. We also will be needing data scientists to help us create a process for dataset investigation.

Thank you.

(Applause)

>> FRANCESCA ROSSI: Is there any question while we wait for the ones from the app?

Yes?

>> Are you going to include in the datasets current policies to

find out the biases they have?

>> RUMMAN CHOWDHURY: Current policies as in --

>> As in existing policies already.

>> RUMMAN CHOWDHURY: Policies around data?

>> No, for example, UN policies, each and every organization they have policies they work around with.

>> That is a very good point. I can sort of make an example from my own research, where we sort of looked at e-triage, so there has been a change in policy within the hospital we are looking at, which if you weren't aware of that, it has changed quite dramatically the output of the algorithm. This is, again, another iteration needed to come up with useful outcome, and so the point is exactly on the data collection mechanism, the policy is important.

>> One of the request he is has to do with when we get the pub -- one of the questions has to do when we get the public data repository, how can we predict or track the way it is used?

>> RUMMAN CHOWDHURY: We can track but cannot predict. That's the nature of anything that is public. That's also the beauty of things that are public. People who design the algorithms cannot think of all the ways it can be used, but by putting it out there in the world we are able to achieve amazing and beautiful things. Also possible to achieve terrible and horrible things. What we will do is create as many guidelines as possible. What I will say is policies that exist today have zero interest or guidelines around how to use it properly or ethically, so this is not about -- to get to one of the

other questions, it's not about creating unbiased data. Bias is contextual. So for example, I can say that we should not use gender or income to create a model about whether or not you should get insurance. But I would probably want to use that data I am creating a recommendation engine for shoes because I would not want to recommend an $800 pair of shoes to a bus driver who is male. So bias is contextual. Fairness is contextual. So we are not trying to create unbiased data sets. We are trying to give people the tools to properly think about ethics and fairness within the models that they build.

Do you want to add anything?

>> FRANCESCA ROSSI: So another question has to do with -- oh. Okay.

What if -- no. Disappeared. Went somewhere else. Anyway, it had to do with the fact that bias can come from other sources rather than just data. It can come from the scientist themselves. So how can we deal with that?

>> RUMMAN CHOWDHURY: So Sebastian had mentioned this is part of the transparency about the data collection process, the question about policy earlier. This is not information that most data scientists know. Making that available, start to talk about the collection bias and measurement bias that may occur. The other end of the bias, of course, is the contextual society bias which we mentioned a bit in our discussion.

>> FRANCESCA ROSSI: So you think that together with the data

repository of data, one should pair it with mechanism to help scientists also understand that they can introduce bias in some other ways?

>> I guess what is useful in this and tried in certain circumstances is if you publicize a particular data set, you try to along with that have a repository to keep track of -- to enable an exchange of the standard data set. One example is the mimic data set in health on ICU interventions, and there is efforts to coordinate between different people using this open data set. But I think it's difficult to eliminate all the biases, and scientists can bring their own biases being in a particular stage of their career working under pressure, and this will also affect in a human way. We are under pressure to come up with the next thing, not necessarily to come up with the trustworthy thing, and that's, I guess, the downside of what's currently out there.

>> FRANCESCA ROSSI: Okay. So another question that I see is about the project, so it says will the project also cover other areas of trust, such as the concerns required for public use and anonymization privacy of the individual?

>> JESS WHITTLESTONE: I just want to make sure I am not hogging the mic the whole time. I think that's an excellent question. Consent is a very, very difficult thing to try to understand. This is why we are relying on partners. We are hoping to work with our partners, and maybe that is another secondary thing that comes out of this is to understand how to better arrive at consent. This also

goes back to the original question about data collection polici. Again, we are not -- policies. Again, we are not the ones doing the collection of the data. We are curating data being collected by partners. This is the only way we can ensure a massive diversity of data. Otherwise we will spend the entire year just collecting narrow data.

What we are hoping for, then, is a dialogue with our partners to talk about these questions, like constent, like privacy, like anonymization. I think that's an incredibly valuable discussion to have.

>> FRANCESCA ROSSI: Okay. Any other questions from besides the app? No? So let's look at the app again.

How can people consent to uses of the data that are not even known to the researcher? How can it be informed?

>> JESS WHITTLESTONE: So I think this question is going a little beyond the scope of what our project is doing. You are asking a very basic privacy and consent question that people who create privacy and consent policy are grappling with at this moment. I absolutely agree with you, and we do not have a clear answer to that question. Nobody has a clear answer to that question. And you are absolutely right. Previously consent was something like you give me -- you know, I give you 15% off on this website, you agree to give me your email address, and I will occasionally send you emails. It has now evolved to you are consenting to give me the number of steps per day on your watch, and I am hoarding that data as a company, and I may

or may not know how I am going to use it. Does that mean you give me free license to use that data as I wish, which may mean licensing to an insurance company who may use it to track how many steps a day and say you don't exercise enough so your insurance premium may be higher. Nobody has clarity on that question.

>> I think perhaps the GDPR where things have been clarified more on this.

>> FRANCESCA ROSSI: The next question is by serge Ross. Do we need trained to not be biased label that you could get by passing a test with this data?

>> JESS WHITTLESTONE: So Serge, you mean for the scientist to be trained?

>> FRANCESCA ROSSI: The system, I guess.

>> (Off microphone). And I could actually try to use your data and compare the output on data and then kind of determine does my AI bias or not? So that would be kind of a label that my --

>> FRANCESCA ROSSI: Certification for an AI system? Is

>> Yeah, certification type of thing.

>> JESS WHITTLESTONE: So there are many ways of approaching algorithmic fairness. This is looking at the data that is simply one component in your pipeline. The question earlier about how can we predict or track the way it is, my answer is you can track but not predict. Sebastian mentioned garbage in, garbage out. We are trying to help that part. But if your algorithm is garbage and your model is poorly specified, you will still get garbage out, and that

may not be a function of our data, but we will have no control over that. We are just focusing on the transparency and trustworthiness of the data towards building transparency and trustworthy of AI.

The second half is people need training and understanding of algorithmic fairness and clarity.

>> Right, I mean just there are different notions of fairness, for example, Troying to, if you change a covariant from male to female keeping the same problemistic classification shouldn't change at all or at least within boundaries, and then you can define different metrics on that. But I think for these methods to work, the data also needs certain quality I guess is more focus of this project.

>> So my idea was turn this around. You have good quality data, and you could use that to actually see, like, if the stuff comes out right. If the algorithm is bad, it will probably come out as garbage.

>> JESS WHITTLESTONE: So I think that's a really interesting use. I think there's something kind of similar happening at Microsoft, sort of. We are working with them. There's been a paper published on transparent model distillation, which is sort of like that, when you compare a model, the output of an algorithm and the output from the real world, and any difference would be bias. So yes, there may be clever ways one might use this data towards understanding the pie I can't say in your own model. I think that's very clever, yeah.

>> FRANCESCA ROSSI: Yes, hi. Yes.

>> Hello. Hi. So I think this is a very worthwhile effort. I just wonder if you heard of two data consortiums, one of which is

called the linguistic data consort yes, ma'am in the U.S. based out of U Penn, and the other one is called LRAC, it's European. These two data consortiums have been collecting data for research and development purposes for over 20 years, and in doctor obviously, in the language and speech area, so linguistic sources, language resources, before the age of Big Data, Internet data, Web data, social media data, they were the ones who actually went out there and collected data for our research.

And there is a very important criteria for every database they collected, which is called data balance. Sampling balance. Which means that you need to have equal number of male/female speakers, an equal number of -- you know, many dimensions of balance. It's called data balance. Bias was not a concern at the time because we just thought it was best science if your data set is balanced. So it's only in recent years when people turn to this kind of data set in the wild from social media that we start to see these kind of bias.

So I think the good practice of data collection has been in place before the problem with that, though, is that we have found there's discrepancy between system performance if we train our system only on those data sets that was careful collected and then you try to apply that to the real world.

So I don't know, are you thinking of going back to that kind of practice, or are you thinking of doing something different? How are you -- I mean, other than I've heard you talk about gender bias. Is there any other dimension bias you are thinking about or trying to

address? Religious I assume, but what else?

>> I guess this is really two questions here. The different biases being ethnic, age, background, religious view, sexual orientation. There are publications, you think you can almost detect if somebody is homosexual from image classification algorithms, and what are the implications? This sort of just giving a hint, and I guess Rumman can expand on the different biases. I mean, I don't want to go -- go back to the old days we did it very limited and small, but consolidate. I think sampling, which is well known in statistics, something that is sort of -- has still meaning today, and it's something which is part of this project to highlight that.

>> RUMMAN CHOWDHURY: I want to reiterate I don't think there is one way to choose bias. We don't pick and choose what is -- bias will creep in, for example, with race being correlated with one's address or with one's income or with one's profession. And this is nothing to do with something very -- that's actually associated with the race, but just a product of society, institutional biases.

Our goal here is not to make yet another shared repository. I realize shared repositories exist, so I listed a few of them at the beginning of the talk, but none of these shared repositories go into the depth, particularly for different types of machine learning algorithms or uses, about describing not only data collection mechanisms, biases that may occur, issues that may happen with the

variables, but also with the societal context and biases that may occur, like as I mentioned, even if you have perfect data. Right? You can have the best, most representative data, but it still will reflect cultural and social biases. I actually do not know of a data set repository that does that for you and explain that to you.

It is also not -- so even when we say creating balance, you may not always be creating 50/50 balance between men and women. That's not necessarily de-biased or unbiased or most fair. That is contextual. You could be creating a shopping website for shoes and you have pretty granular data, and you need to know things like gender and income; whereas, other cases these may not be variables that should be included. We cannot say all the scenarios in which our data will be used. But it is significantly better than what young data scientists and actually practicing data scientists do today, which is scrape pictures off of of Tinder, scrape text off of websites because it's all they have at their hands.

>> FRANCESCA ROSSI: Okay. I will put together two questions which have to do more technical nature. One says what about the creation of synthetic data, and the other one has to do with do you see potential in generative models to create less biased or more balanced data, for example, creating more diverse faces by (?) generational parameters randomly?

>> RUMMAN CHOWDHURY: Sure. So the issue with synthetic data is bootstrapping data is simply a function of the data that exists today. I am a social scientist. We do a lot of bootstrapping. So we do

things like, for example, in the United States, certain districts maybe have a very low minority population, so we take the survey results of the two minorities who took the survey and, you know, extrapolate that upwards. You can see where problems arise. You have one or two people that are determining the outcome of an entire population. So synthetic data is useful in some cases, it is not necessarily a cure-All.

>> Maybe expand on this. Only half I remember. On that point you made, for instance, poor prediction. There was a case where poor prediction was really heavily influenced by one thing, a black male in the state --

>> Southern California.

>> Yes, I don't remember. So this is an example of that. And I guess generative models, how they can help to reduce the biases, is I guess if you think of (?) I think it's tricky because it's always influenced by the data set. I think I don't see how the generative model necessarily has a better understanding of representing what's fair or how the original data has been collected.

>> FRANCESCA ROSSI: Any other questions from the audience?

Okay. One has three votes. It says even if the source of data is trustworthy, what kind of a filter, cleaning, have you thought to make also the content unbiased, content which can come from different users.

>> RUMMAN CHOWDHURY: So I am going to repeat that we are not trying to make unbiased data. This is not about taking data that exists

and imposing our idea of what fair is or what unbiased is or what clean is or what ethical is. I don't think either of us are anyone to be doing that. This is about making transparent the data that is being provided for people so that they can make educated decisions on how that data is being used in their model. And it's also about providing information and knowledge about the societal and contextual issues that may arise. Both of these things are things that data scientists are not explicitly trained in necessarily and may not be explicitly aware of in the data sets that they use today.

>> FRANCESCA ROSSI: Okay. Any other question? Yes, last one.

>> It may be related to same topic. Not assuming it's trustworthy, are you looking at all at the provenance of the data and its authenticity?

>> RUMMAN CHOWDHURY: So trying to identify fake data?

>> Correct.

>> RUMMAN CHOWDHURY: So this is why we are relying on having high-quality partners. We are not trying to just crowdsource information from people, et cetera. So that will be a function of our partnership with the people we work with to think about -- and this is also what Sebastian was talking about with the data collection, the data source, et cetera. So implicit in that process would be vetting authenticity, et cetera.

>> FRANCESCA ROSSI: Okay. So let's thank them again for the description of their project, and we can discuss again later. Thank you.

(Applause)

So the third project of this team will be presented by Krishna Gummadi, who works together with Adrian Weller on this project, and it's about cross-cultural perspectives on the meaning of fairness in algorithmic decision-making.

>> KRISHNA GUMMADI: Thank you, Francesca. So what I am going to do today is tell but some recent work that we have done on actually understanding how people in the U.S. perceive the fairness of algorithmic decision-making, particularly in the context of criminal risk prediction. Our hope is that we would have other partners who might be interested in replicating or redoing this kind of a study in other countries and potentially for other decision-making scenarios as well, like for credit scoring or, say, recommendation systems.

So this is joint work with, actually, one of my PhD students, Nina, at the max Planck Institute for Software Systems, as well as (?) from the University of Maryland and, of course, Adrian weller here.

The way we are going to share our task here is I am going to present some findings that in a provocative manner, and Adrian is going to answer all the questions.

What I am going to talk today is about algorithmic decision-making, and as pretty much everybody here knows, this sort of algorithmic decision-making is being used in many scenarios that affect lives of people, they are being used in hiring, they are being used in assigning social benefits, and the specific thing that I am

going to focus here is on granting bail, where we have algorithmic -- where we have learning-based algorithms that are being used to predict the risk of someone recidivating and thereby affecting their chances of getting bail.

Now, the question, of course, is are these algorithms fair? And if you want to reason about the fairness of an algorithmic decision-making system, perhaps we could think of it consisting of three parts in the decision-making pipeline. First you have the inputs or the features of the users that you are going to use, and then you have the decision-making system that will process this in some way, and then it would generate some output.

In this talk today I am going to focus just on the inputs of the decision-making pipeline, and specifically the question we are going to ask is, is it fair to use a particular feature for making the decision? So we are not going to talk about the fairness of the algorithms and the ways in which they are processing the data, but just about the fairness of the inputs.

So I am going to focus on three questions. The first is is it fair to use a feature? And the second is why do people perceive certain features as fair or unfair? And a third thing is do people actually agree in their fairness judgment?

So to focus on this first question, we thought there could be two different ways in which you could go about this. One is to take a normative approach, where you have some intellectuals that would say here's how fair decision-making ought to be done. There are some

really smart people who have thought in through and who end up telling or defining how fair decisions ought to be made. Now, of course there are antidiscrimination lawsuits can be thought of as an example of this, where for instance, there are certain features that are explicitly identified in antidiscrimination laws.

Another way you could I this of is where we actually ask humans, people, affected by these algorithms, how they perceive the fairness. We could ask whether it is fair in using parents' criminal history in predicting whether you are likely to commit a crime. Is it fair to use your education background in predicting whether you are likely to recommit or reoffend in the future?

So the case study I am going to focus on is specifically on a tool called Compass. It is a tool that is built by a commercial company in the U.S. This tool has been -- what it does is it's supposed to help judges decide if a person should be granted bail, and the way it works is it takes as input the defendant's answers to a set of questions, so the compass has a questionnaire, and the answers that the respondent or the defendant provides are used to actually estimate the risk of someone committing a crime again in the near future, and that, in turn, is used by judges in certain U.S. jurisdictions to decide on decisions related to granting bail.

Now, let me give you a glimpse of what the compass questionnaire looks like. It has 137 questions, broadly categorized into ten topical categories, and these are the categories. Some of them are things like current criminal charges and criminal history, but there

are also questions related to substance abuse, whether the person has a stable employment history, what their personality is, what their criminal attitudes are. Example of a criminal attitude is would you steal if you were hungry? And how safe their neighborhood is. And it also includes criminal history of friends and families, the quality of the social life they have, and their education and behavior in school.

Now, of course, there are 137 questions, but each category has like 10 to 13 questions, and this question is -- questionnaire is publicly available. I would highly encourage you to check it out for the kinds of questions they ask under these categories.

One thing I want to make clear here is none of the questions are directly related to any sensitive features. There is no race or gender or even age that is actually directly asked for this n these questions. The question is is it fair to use these features to make bail decisions? So here is how we gathered human judgment. So because this is related to the U.S. criminal justice system, we did a survey primarily of U.S. respondents. The way we recruited them is one a bunch of users from Amazon Mechanical Turk. These are these category of people called master workers. Of course, Amazon Mechanical Turk is not a population representative sample of the U.S. So what we did is we also used another company called SSI, and we assembled a survey panel, which is about 380 respondents, and these are picked to be census representative.

Now, the findings that I am going to present are consistent across

both the samples. So when it comes to -- the first thing we were interested in is how did these respondents rate the fairness of a feature? So as I mentioned, we had ten different categories, ten different topics, and for each of the topics, what I am showing along y-axis is the mean fairness rating that we got from the respondents. So the ratings go from 1 to 7, and 1 is that it's very unfair to use the feature, and 7 is it's very fair to use the feature. As you can see, there is a wide divergence across the different topics. In fact, one of the interesting things is if you look carefully, the mean fairness is actually lower than 4. 4 is like it's neutral, they don't have an opinion. For a majority of the features that are being used to make these predictions, the mean fairness is on the negative side. Thatence into most people, it to be unfair. thoos are topics related to education, school, criminal history, family and friends, so on. Now, that's about how fair people think the using of the features in the compass are.

The next question is why do people perceive these questions adds unfair? We had two hypotheses. The first hypothesis is whenever you see a certain category or feature, you tend to things in the back of your mind that may be influencing your judgment. We came up with a list of eight latent properties. We just thought hard about this. We are not claiming that this is exhaustive. But I will present some results as to whether -- how widely they cover. But the feature, latent properties are things like whether the feature is relevant to the decision-making. For instance, is your educational

background really relevant to making decision about your criminal in the future? The second thing is how reliably can a property be assessed? How reliably can you assess criminal attitudes by asking the people the question "would you steal if you were hungry?" Whether a feature -- is the answer to a question like your parents' criminal history, is that something of one's choosing? What about the other one is is a feature very privacy sensitive? Is it fair? Is juvenile criminal records of a particular person fair game for estimating recidivism risk or is it a privacy hen sensitive issue.

Then we had a bunch of reasons related to causality. Does the feature cause the outcome? Coo could the feature result in vicious cycles? Could the feature cause (Inaudible) and finally, is the feature itself caused by the membership in some socially salient group like race or gender.

This was our first hypothesis. Second is intuitively when people are making a judgment about fairness, they would estimate the latent properties then do a mapping of the latent properties in determining whether something is fair or not. This was -- these were our two hypotheses.

To reason about the fairness judgment, what we did was we again conducted another survey where we asked the respondents why they thought whether a feature was fair or unfair. Particularly, when they thought that something was unfair to use, we looked at with what frequency they cited one or more of these latent properties. And this graph actually shows the results. The first point is that most

people were happy with the eight reasons that we provided. In fact, only a very, very tiny fraction, 3% of all the survey respondents, ever even mentioned a factor other than the ones that we listed here. And the other important thing is if you actually look carefully at the plots, so the Y axis is showing how frequently each one of the features was cited, so .15 means 15% of the respondents cited caused by sensitive feature for the reason as why they rated some feature as unfair to use. Notice that actually there are only two of these latent factors that are directly related to discrimination, which is causes disparity and caused by sensitive features. And there are a whole bunch of other factors, like causes vicious cycle or reliability of assessment that are actually not directly related to discrimination but that actually influence people's judgments about fairness of using the feature. So to put it in a different way, even if the society consisted of completely homogeneous population where we didn't have to worry about race-based or gender-based biases or discrimination, there are still good reasons or other reasons for why we might not want to use some features.
That's the point that this plot actually shows.

Now, next when it comes to modeling of the fairness judgment, what we did was we took the -- we did another survey where we asked people to judge the latent properties or to estimate the latent properties first, and then we tried to predictk their fairness judgment. Here what we found was that we were able to actually train a simple classifier or a simple mapping function that was actually working

pretty accurately with 88% accuracy. What that means is if you tell us about how you perceive these latent factors, we can actually predict whether you would think the feature is fair to use or not with very high accuracy. And what it also -- another way of interpreting this is that many people seem to actually use a similar kind of a heuristic or similar kinds of weights for how they would consider the relative importance of these factors when making their fairness judgment. That's the interesting thing here.

Now, of course, this was within the respondents, which is the U.S. population, there was one heuristic that worked well for across all the people, but of course, an interesting future work would be to see whether in different societies the heuristic itself -- that means how much importance you give to different factors -- might change.

That's about the second one, why do people perceive features as unfair. Now getting to the third one of do people actually agree in their fairness judgments? So what we did next was, again, we have the ten different features. Then what I am plotting along y axis here is how much agreement there is in terms of the judgments people made. Zero means complete disagreement, and one means complete agreement. And as you can see, for certain features like charges and criminal history of the person itself, there is a fair bit of agreement, but there are a number of different features where there is considerable disagreement. That means different people have different opinions on whether it is fair or unfair to use these things. The question is what actually is causing these

disagreements? If you think of our hypothesis, what you will notice is there is a disagreement in the fairness judgment, but this could be caused because of a disagreement in the latent properties or in the mapping function itself. As I pointed out before, there is a fair amount of agreement on the mapping function, and so most of the disagreement has to come from the way in which people are assessing the properties. What we actually found is when it comes to reliability of assessment and privacy sensitivity of a feature, there is reasonable consensus. But where the consensus completely breaks down is when it goes to the causal reasoning part of it. That means which feature causes what. And particularly when it relates to the causal reasoning, there is a complete sort of a disagreement or a fairly substantial disagreement amongst people.

Now, we also looked at whether these -- how this is correlated with demographics of the people. We looked at demographic factors like age, race, education, and gender, and somewhat surprisingly for us, we didn't find really statistically significant differences along these demographic factors, but then when we looked at political views of the people, there were significant differences. In fact, when you go from very liberal to very conservative -- to people who have very liberal to very conservative ideaologies, people with very conservative ideaologies tend to rate using of many more features as fair than compared to the people with very liberal ideaologies, and particularly when it comes to causal reasoning, there is

substantial disagreement between these two groups in terms of which features cause what. In general, very conservative leaning people seem to think that more features are relational, where more liberal leaning people seem to be arguing that more features are actually caused by social groups and so on.

So to quickly summarize, what I was presenting was the fairness of using a feature actually depends on its latent properties, and fairness goes beyond discrimination. This is an important point because for better or worse, a lot of ongoing research on algorithmic fairness has gotten extremely focused on just the notion of discrimination, which is an important type of unfairness, but it by itself doesn't cover all types of unfairness. And then when you look at the disagreement in fairness judgment, there is -- we found agreement in the mapping of latent properties to fairness, but a lot of disagreement in the assessment of the latent properties themselves. And especially the latent properties related to causality. And this seems to be correlated with ideological views of people in the society.

So in terms of future directions, I think one of the things that we would love to do more is redo the study in other countries and potentially even in other algorithmic decision-making scenarios.

And maybe if I can leave with a final comment or a final couple of thoughts, what is actually interesting or rather surprising is that we found agreement in mapping from latent properties to fairness, but disagreement in the latent property assessments.

Because if you were to step back and think about this, or at least when we were initially conducting the studies, we thought the latent property assessment would be somewhat more actively assessed and therefore there would be more agreement there. But where there will be disagreement is in terms of how much weight to give to the different latent properties. But what we found was somewhat the opposite of what we expected.

So what does this mean for the future of algorithmic decision-making? So here's just one part. So the moral reasoning is about mapping. We could actually maybe perhaps derive a moral reasoning in terms -- by inferring the mappings functions. But then when it comes to the latent properties, like particularly those related to causality, if you -- in theory -- this is in real theory in the sense of like in practice this is very hard to do -- in theory, if you had sufficient data, you should be able to infer the causal relationships, and we shouldn't be actually relying on people to guess what the causal relationships are. And if we ever get to the stage of doing that, then maybe we could actually settle this disagreement in the causal reasoning part. And hopefully it will lead to a more agreeable algorithmic decision-making systems.

Thank you.

(Applause)

>> FRANCESCA ROSSI: Okay. I didn't see any new questions, but anybody want to ask a question from the audience? Yes.

>> Very interesting work. One question that I had immediately, so you have picked out two subgroups. Spectrum of groups conservative to liberal. Have you discovered any other subgroups in terms of clustering or other algorithms to date (?) subgroups in fairness, it would be interesting to hear.

>> Okay. So we were doing -- we did that kind of clustering, but we are still trying to interpret the clusters that we found, meaning so we could cluster people who are thinking similarly in terms of the fairness, so you would get these clusters of people who are making judgments in a similar manner.

Now, the question is what categorizes that cluster; right? That's where you would have to use these annotations that you have for each one of the users. And there we found those clusters correlating with conservative or liberal but not clustering with other things. This is still ongoing work, and we are still wondering if there are other ways of categorizing clusters of those people thinking in a similar manner about fairness.

>> FRANCESCA ROSSI: Any other question from the audience? Yes.

>> For the part you said about the waiting, there was a fair amount of agreement in terms of the weighting of the different properties in terms of whether they are judged as fair or not in making those decisions; right? I was wondering if it's because this is like -- it's inherently hard. Let's say you have a hundred points to do weights for all of these things; right? Because this is something that's highly disparateizable, is this something that would -- a

certain equitable distribution between those attributes? Maybe I am not phrasing it in the best manner, but what I mean to say is they don't have as much of depth of understanding to be able to attribute weights to it as much, and then you could regress as much as possible to maybe make it as equal, but -- or something that's very clearly important would get, let's say, 50%, and then the other things you just distribute 50 divided by 6, whatever that is. I don't know. Is that something that you think could have happened?

>> So we did observe a clear ranking of the different latent properties. So if someone considered a feature as not being able to -- that you cannot reliably assess it, then there is a very high chance that they would label it as unfair kind of a thing.

So what I didn't actually go into the details of like even for assessing the latent properties, we actually gave them a scale of 1 to 7 where they can relate something as reliably assessable or not. So we had much more fine-grained data on that. So it's possible that some of what you are describing might have happened, but it's -- but still, I think the higher-order takeaway that people seem to be thinking about when they are reasoning about fairness that they are accounting for certain latent properties with higher priority than others. I think that trend is plenty consistent in the data.

>> FRANCESCA ROSSI: Yeah, so a question here has to do with some static features versus dynamic features. Are static features viewed as less fair than dynamic ones? Is

>> I am wondering what -- how --

>> FRANCESCA ROSSI: Maybe somebody who asked the question can say something about static versus dynamic.

>> (Off microphone)

>> You mean -- ah, I see, I see.

>> FRANCESCA ROSSI: The value of the feature changes over time.

>> I see. So I guess if something is static -- that's an interesting dimension that we didn't --

>> FRANCESCA ROSSI: Maybe that we can change, so it will be maybe unfair to focus on a dynamic feature which can change over time. In the future it can be different.

>> The only thing we thought of was volitional versus nonvolitional, and I guess it does seem like there are many things that are non -- okay. Yes, that's an interesting -- another dimension to think of. Like we didn't consider that.

>> FRANCESCA ROSSI: The class in static and dynamic, you didn't consider that.

>> Yes, because if something is very static, that means you probably cannot bend it to your will. That means it's less volitional in that sense. Maybe there is a correlation between volitionality and whether it's static or dynamic.

>> As you said, for example, income would vary beyond your control to some extent.

>> That's a good point. Thanks for that question.

>> KRISHNA GUMMADI: So another question is about what kinds of data would be most helpful to future developments of your work? So

you analyze, you know, one data, one algorithm -- one kind of data for the compass algorithm, yeah.

>> At a high level, we are really trying to understand what do people think of as fair. We mentioned before fair means different things to different people. If we want to understand different algorithms, we need to understand what do people feel. This was a starting point looking at this one specific question in one specific location, and we observed this result, which actually was somewhat surprising to us about the relationships that Krishna described. And we are very interested to see if those relationships would hold if you asked different questions and if you looked in different regions. So we are really interested to gather. I don't know if we have any immediate things to look at, but I think lots of things would be game.

>> FRANCESCA ROSSI: Okay. Yeah.

>> Can you hear me?

>> FRANCESCA ROSSI: Yeah.

>> Yeah, I might point out something which is come to last talks. It is the idea of fairness, whether it is legal or commercial or whatever.

One thing might be interesting to introduce not, I would say, (?) but for the stakeholders, for the judges, whether in fairness is (?) whether for insurance company, whether making an unfair decision toward women would cost more than making an unfair decision against a man, for example. And then you can use your data. You

don't think about whatever constraints, you can use your data, but just weight what is the cost of mistakes toward this or this group of people, and I would say scientifically, you can build some algorithms which deal with that.

>> So I think -- as you are asking the question, sort of combine the two.

>> Yeah, sure. The 88% of the accuracy, I think that is flat; right?

>> I am sorry.

>> The 88% of the accuracy that represented, it is flat; right? Is it is a flat accuracy? That means just the number of --

>> ADRIAN WELLER: Yeah.

>> Okay. Thank you.

>> So I can talk to your proposal, which is, as I understand it, what if there was such a strong reputational disincentive to have unfair data/algorithms that it was sort of a self-enforced mechanism? I think that would be a lovely world if we lived in it, but it is not. I wish there were. There's increasingly scrutiny. One may debate what the reputational outcome is. I think you need to -- for example, all this discussion about Cambridge Analytica, it's not just about data breaches, it's become a discussion about privacy. Yet Facebook and others have been using data to target people for many different things for many years. The Obama Administration -- storery, when former President Obama was running for election, it was very transparent and open that they were using data science, and

everybody applauded it. So you know, cultures are changing, but it is not changed.

The other problem with waiting for a reputational backlash is that, for example, in the Compass data set, we have very real people whose lives will have very real negative outcomes. People denied the right to see their families for the rest of their lives, for example, are we to say hey, by the way, you are just going to be part of this collateral while we wait? I think that iherently is also fair because we are putting this out in the wild. It is not an experimental structure.

>> KRISHNA GUMMADI: Maybe you can take it up later when we have the discussion of the project. In the interest of time also not to make you wait for the coffee break, I think now we go to the coffee break. We thank again all the people presenting and working on these very interesting projects, and let me just remind you so that everything is being shifted with respect to what you see here of half an hour, so now the coffee break from 3:30 to 4:00, and then there will be the panel discussion on trust in AI opportunities and challenges from 4:00 to 5:00. Then from 5:00 to 6:00, there will be the breakout sessions in a different room, just below this one, there will be one table for each of the nine projects, and people can choose which one to join and discuss.

So thank you again, everybody.

(Applause)

***

***