In Principle and In Practice

March 31, 2020

Welcome this premiere. To this premiere. This is the first entirely virtual Berkman Klein Tuesday lunchroom event. I'm Urs Gasser, and I have the great pleasure to moderate this one hour session, both in principle and in practice, I hope. As the announcement says, we will take a closer look today at the many AI ethics and governance principles that have emerged across the globe over the past few years.

In the first part of the session, we will hear from Jessica Fjeld, who's Assistant Director of the Cyber Law Clinic here at Harvard Law and at Berkman Klein, who will share some insights from her recent, principled AI report, which provides a mapping and interesting analysis of norms.

We will hear from her about some of the common themes and threats across the different principles, as well as about also differences among them and maybe even gaps.

Now we all know it's one thing to write principles, but putting them into practice is a whole other story. And I'm therefore particularly grateful that Jess and we all are joined by our colleague, Ryan Budish, who's an assistant research director at the Berkman Klein Center, who will highlight the few of these implementation challenges based also on his work as a member of OECD's AI governance expert group, which was one of the bodies that came up with the set of principles.

So I'm looking forward to both opening presentations, which will of course set the stage also for our discussion afterwards.

After these two opening statements, I'm really thrilled to invite three respondents to join us. We have Mutale Nkonde. We have Doaa Abu-Elyounes, and Vivek Krishnamurthy. I will briefly introduce them after the initial presentations.

We will have a Q&A as well, although only virtual today. Please type your comments using the Q&A function, which I will monitor. I will then select and hopefully also cluster some of your questions and share them with our speakers.

Please also note that this session is recorded. We will share it later on.

Of course, we know this is a little bit of an experiment. We will use the webinar style mode for this luncheon, but we will also experiment with other technologies and other modes going forward. And if everyone who joins could go on mute to the extent it's not the speaker, that will be great. And without further ado, I turn it over to you, Jess. Thank you so much for joining us today.

I'm delighted to be here with all of you. And I'm really excited to be able to talk about the principled AI report that we put out in January. We had been looking forward to doing this as an in-person lunch for a couple of months now. And so, in this new world that we are all in, it's good to have this opportunity to discuss it. And I'm really looking forward to all of your questions and to our discussants' reactions, as well as to the more practical perspective that Ryan brings.

As we are discussing this event, we thought it would be really interesting to kind of go at AI principles, which have come out on such a hot and heavy schedule for the past few years, both kind of from the macro view that the principled AI report takes, and then also from the micro view, from Ryan's perspective having been involved in the drafting of the OECD ones that are particularly influential.

So also thanks to Rube and Megan and Lis at Berkman for helping put this together. And, of course, my co-authors on the principled AI report, including Adam Nagy, Nele Achten, Hannah Hilligoss, Madhu Srikumar, and just the rest of the research team who helped put this together.

So Principled AI was the result of a yearlong study of principles documents that set forth standards for socially responsible AI, which seek to ensure it will be ethical and rights respecting, have a positive impact on the world.

As I noted, we released a white paper in January, which is available on the Berkman website, and also this visualization. Now if this visualization isn't immediately transparent to you, don't worry. I will be going through all of it in a minute.

I'm planning to talk for just under 10 minutes, just to give you an overview of the project methodology and findings, and then of course happy to take your questions after Ryan's presentation.

The top level finding that we have to share from Principled AI is that, in spite of all the chatter and concern over the fact that there isn't really a vision for socially responsible AI, we were able to isolate some strong themes in the 36 documents that we looked at. And we believe that they are the signs of the sort of earliest emerging consensus for societal norms around how AI can be, should be, used.

Now, of course, principles are just a piece of governance. And they should exist in a broader scope of governance that includes everything from the everyday practices of professionals who are involved in this all the way up to law and regulation at multiple levels of government.

So here's a timeline that shows all the documents in the data set. Just on a slightly different way, you can see that the earliest one we have in the data set is from 2016, the tenets of the partnership on AI. And they go all the way up to late 2019.

The data set is a curated set of 26 docs that we assembled using what's called an expert or purposive sampling method. So it is not an attempt to be comprehensive. We're aware of approximately 100 documents that would loosely fit our definition. But what we wanted was a

kind of manageable set. We knew that we wanted to build the data visualization and that we couldn't have too many documents on there, or it would go from where it is now, which is a challenging and intricate, to basically unreadable.

And we wanted, nonetheless, to be able to include the-- to include the large number of documents. So a variety of documents. So in terms of stakeholders, in terms of the timing of the publication, I will note that we were hoping for variety in geography. And we were able to achieve some. But, for example, we were not able to find any documents from the continent of Africa. We're aware of some that are in process, but there were none that fit our definition that were published at the time of the report. And so that obviously marks a significant shortcoming in our finding about the existence of a kind of global consensus.

So here, you can see one vision for, when we look at the variety within the data set, we are particularly interested in having a variety of stakeholders represented. Because it was our hypothesis that this would be a significant area of variation between the data sets.

We also included, it's worth noting, documents that looked at AI as it-- AI technology generally as it is applied in specific sectors-- for example, the justice system or the workplace. But we excluded documents that looked specifically at a particular type of AI technology, such as autonomous vehicles or facial recognition. Because as we looked at those documents, they were just different in character than the generally applicable AI-- the sort of broadly applicable AI documents.

Also worth noting, not all the documents in our data set include the word "principles." They don't all use that to describe themselves. But we are understanding of that word. Our definition of that word was, documents that make normative, in the sense that it's used in the legal community. So a sort of proscriptive statement about how AI ought to be used. We excluded empirical or observational documents. Like for example, the annual reports that come out from AI now, which have a lot of interesting insight on how AI is ethically used and deployed but don't sort of contain that normative statement.

So let's go back to the data visualization now that you have a little bit of an understanding of what is represented here and look at how to read it. So each spoke on this visualization is one document, with the exception of the OECD and G20 principles, which are represented on a single spoke. Because the G20 adopted the OECD principles more or less verbatim.

The principles themselves are verbatim. They excluded some of the descriptive text. The sector-- the stakeholders are color coded, the same as they were in the pie chart. I showed you on a recent slide. So green is government. Orange is intergovernmental organizations. Blue is multi-stakeholder. Pink is private sector. Yellow is civil society.

There are nine rings in the visualization. The eight inner ones are the themes that will be isolated. The outermost one is international human rights, where we collected data on whether or not the document mentioned human rights or explicitly noted that it proceeded from a human rights law framework.

The framework documents are indicated by a star. The documents that mention human rights or related international instruments are diamonds.

For the themes, you'll note that there are circles and that they are different sizes. The size of the circle corresponds to the percentage of principles in that theme that the document contains.

So if there are 10 principles in the theme, and the document hits all 10. It gets the largest size circle. If it hits just one, it gets the smallest size circle.

Because there are different numbers of principles within each theme, it's instructive to compare within each ring but not between the rings.

So what are these themes? Here, we've zoomed in a little bit. And you can see them better.

The eight themes, in order of how frequently we see them appear in the documents, are fairness and non-discrimination. Some principle related to fairness and non-discrimination appeared in every single document in our data sets.

Privacy is the next one, and accountability. They were both in all but one document. Transparency and explainability, then safety and security, professional responsibility, human control of technology and the promotion of human values.

We got to these eight themes by hand coding every principle in the data set and then grouping like principles together. So it was very interesting. At the same time we were working on this project, there were a couple of other sort of similar studies of principles which all came up with a number of themes that are parallel to ours in many ways, though everyone came up with a slightly different number.

So for example, the report that came out from ETH Zurich has a theme called beneficence, that sort of loosely lumps together, I think, our promotion of human values and some of the accountability and safety and security principles. So slightly differently divided, but I think researchers around the world are making similar observations.

Some people have come to us with frustrations about the themes. For example, we've heard from a few people that they wish that sustainability or environmental responsibility were sort of more of a top level item. It is represented both in the promotion of human values and the accountability principles. But because there wasn't a large number of principles under that heading, it didn't rise to the level of these themes.

So every principal has a different number-- sorry, every theme has a different number of principals within it. This just lets you see a little bit better what those are. The accountability theme has the most greatest number of principals within it-- the promotion of human values, and human control of technology themes having the fewest. And you can also scan this to kind of get a sense of the range in the themes, right? We see some that are really big sort of consistently capital letter abstract concepts, like equality, under fairness and non-discrimination. Whereas we

also see very particular policy recommendations like, under transparency and explainability, the idea that there should be a notification when an AI makes a decision about an individual.

So my last slide, just sort of what's next for this project? What's next for these observations? I like to think about it in terms of, what is wrong with this chart? Now, this is a chart that shows the documents in our data set broken down by geography.

You can see that, in spite of the fact that we built a multilingual research team with roots around the globe, we nonetheless had-- the principles were dominated by North America and Europe, with a substantial chunk, about a third, from East Asia, mostly China and a Japanese document. We had one document each from India and from the MENA region and then a handful of documents from Latin America.

I think what that means is that, while those eight themes will be important in AI governance, we have a lot to do to expand the conversation to ensure that all of those who will be impacted by AI are weighing in on this governance piece. And because these principles-- for example, the word equality, we just picked that up on the last slide-- can mean very different things in different cultural contexts and to different people. And because the people who are likely to be most strongly impacted by AI technologies are marginalized and vulnerable populations, I think it's absolutely key to continue to make this AI governance conversation accessible to a broader number of people and to ensure that the voices of a really diverse set of individuals and organizations and governments are represented in it.

So with that, I'm going to wrap up and really looking forward to your questions.

Thank you so much, Jess. I have a quick question that's coming up from Padmashree, and that is, in addition to some sort of the mapping of themes and the content of the principles, did you also map or analyze further what the underlying accountability mechanisms are? Not for those who design AI, but for those who design the principles? So are some of these principles more robust in terms of the inbuilt mechanisms for enforcing their oversight than others? And did you map that?

It's a great question. And the-- I think the initial answer that occurs to me is that it really-- it sort of goes stakeholder by stakeholder. So for example, you get something like the Toronto declaration, right? Massively co-sponsored, organized though by amnesty and access now. That is largely a coalition of civil society and individual and economic actors. There's not a lot of accountability measures for an organization like that that's circulating principles.

On the other hand, some of the government principles are adopted in the context of AI national strategies and do-- if there isn't sort of explicit commitments immediately, at least include sort of recommendations for the study and adoption of new regulations.

So I'm thinking in particular about the German and British AI national strategies there that are sort of looking closer to regulation. Worth noting, though, that the actual-- the first government to adopt regulation which actually parallels these themes in many ways was Canada, which did

not produce a set of principles first. It really did just go straight for regulation, which governs government bodies' acquisition of AI tools. I hope that's helpful.

Thank you so much. We'll have more questions, but I think this is a good moment to turn it over to Ryan to share your perspectives, as Jess already indicated, what's of course ahead are all these hard implementation questions. And I was wondering whether you could share your perspectives based on your work with OECD but also beyond. Thank you, Ryan.

Great. I am really excited to be a part of this experiment in BKC events. So thank you. As I just said, I'll just repeat the thanks for the BKC team that helped make this possible. And, of course, it's always hard to follow Jess. But hopefully, I think what I aim to cover here should follow on nicely with some of what she was sharing.

And in particular, I have two main objectives for what I wanted to cover in the next few minutes. The first is that, when you look at that really amazing visualization that Jess created and her team created, you see all of these as finished-- you see all those principles as finished products. And I actually wanted to provide a little bit of some personal insight, personal reactions, to the process that I experienced being part of the AI expert governance group that the OECD created in developing their set of principles.

And then secondly, I wanted to sort of go a little bit deeper into at least one of the principles within that document to highlight some of both the challenges and opportunities looking ahead as we think about moving from principle to practice.

So as far as the process that the OECD went through, there were really four stages. The first was that they created this group that they call the IGO-- the AI group of experts for the OECD. And I'll talk a little bit more in a second about who was part of that group.

And then-- and they met four times. I was lucky enough to be a part of that group, and we met four times between 2018 and 2019. Then the group essentially drafted a set of recommendations, but that was not the final step. That draft set of recommendations was then passed up to the OECD's committee on digital economy policy, and then that group had the opportunity to revise the draft principles and reshape them, and then ultimately voted on them, and then sent them one level higher to the OECD's ministerial group.

And then that group had a chance to continue to amend the principles. And then it was finally approved in June of-- sorry, June of 2019. That's a typo. In June of 2019, that was finally approved.

Now in terms of the composition of the expert group, there were 14-- what they called outside experts, invited outside experts. And those individuals came from academia and business. Then there was another nine representatives from other OECD committees. And those representatives came primarily from civil society organizations, like those that focus on labor issues, or those that focus on privacy issues, as well as additional representatives from businesses like IBM.

And then there were 33 representatives or so from OECD member states. And many of those people came from specific regulatory bodies within those countries that have overall jurisdiction for issues relating to emerging technologies or telecommunications issues.

Now within the OECD principles, there were really two operative sections. One was the principles for responsible stewardship of AI. And that really related to a set of values-- things like transparency, human rights, issues like that. And then there were national policies and international cooperation for trustworthy AI. And that section related to principles that governments would-- the audience for that, half of the document is really governance.

So that related to, future of labor, data sharing, making investments in AI research. And the document also contained a broad definition of an AI system. What does that mean to be an AI system?

Now as Jess noted in her comments, the OECD document is actually fairly unique in the influence and adoption that it's had. Many of the principles are important for the organization that created them-- a company puts out a set of principles that largely defines how they're going to implement AI. But those principles, in many cases, don't have widespread influence. What one company says that they're going to do doesn't necessarily shape the whole industry. What makes the OECD principles somewhat unique is that it was-- the OECD process operates on the basis of consensus.

And so in order for it to be adopted, all 36 OECD member countries had to agree to it. In addition, there were six non-member countries that adopted it. And then shortly after it was adopted, the G20 adopted-- essentially adopted the OECD principles verbatim.

So I wanted to, as I said, to provide a little bit of sort of my own personal reaction to being on the inside and seeing some of how this one set of principles were created. And there were really three things that I wanted to mention. First, the framing of the process really matters. And by that, one example that I'll give is that the OECD framed the process as, principles to advance the adoption of trustworthy AI.

And what you'll notice about that phrasing is that it has a very sort of positive pro adoption bent to it. And that meant that certain things that I think are actually important parts of the conversation when we're thinking about AI-- things like, are there spaces where AI should not be used? Are there areas in which we think that AI should not be adopted and should not be advanced? That really wasn't in scope for what the OECD was considering.

So the initial framing was important for determining what kinds of principles could be in this document versus not.

Secondly, the question of who ultimately decides, who's this sort of final decision maker, I think is really, really important. And as you saw when I showed the sequence of events leading to the adoption of the OECD principles, the group of experts who created the initial draft, we were not the ultimate audience for the principles. It was the member states that were ultimately going to have to adopt it.

And so what I saw from the process was that it was really designed, in many ways, to try to reach the end of the process where there was something that the member states would be able to adopt, and would be comfortable adopting.

And so you could certainly have imagined alternate processes that they could have gone through that would have yielded a very different document that ultimately, again, because of the consensus driven nature of the OECD process, never would have been approved by all of the final decision makers.

And so I think it was a real conscious decision to design the process in a way to get to something that could be approved. And that really gets to my final point, that the conveners and the staff at the OECD, who was doing a lot of the drafting in between meetings, and responding to comments, really had a lot of power in terms of how they structured the process in order to try to get to that end point.

And so I think it's certainly fair to have some criticisms about the OECD process, whether it's questions about who was invited to the table, or the sort of, should AI always be adopted and advanced? Those kinds of questions. I think it's totally fair to have criticisms of the process and of the document.

But ultimately, I think as measured by how I think the OECD viewed it, as reaching a point where it could be adopted and ultimately implemented by the countries that adopted it, I think it really was successful in that regard.

Next, I just really wanted to quickly give a little bit of an example of some of the challenges and opportunities by looking at one specific part from the principles relating to transparency and explainability.

And the first thing that I wanted to highlight here is that there's four subparts to this principle. But really, when you look at the-- the first three subparts are really all about advancing understanding. And you can see that the key words foster a general understanding of AI systems-- awareness, understanding the outcome. And so these first three are really not about changing AI systems, but really helping people understand how they're interacting with AI systems.

It's this last piece really, the fourth subsection, that really actually, the most challenging in some ways. Because it's about creating a way for people to challenge outcomes that may be adverse.

But when you start to read it, it actually, in many ways, raises more questions than it gives answers. For instance, it says that it talks about people adversely affected by AI systems. Well, what does it mean to be adversely affected? What if you don't even know that you've been adversely affected? Or what if the system actually performs better than its human counterparts, but compared to certain other people, your impact was less good than others?

So it's a complicated concept. To challenge its outcome. What does that mean? In the moment, is there some sort of appeals process? Is it a human review? What does it mean to challenge the outcome?

Based on clean and easy to understand information. In different contexts, different kinds of information might be more or less relevant. And, you know, for instance, is it the single factor that was most important in the decision? Is it the top 10 factors that were most important in the decision? Is it the one factor that if changed the least would alter the outcome? Any of these could be, depending on the context, the most important kinds of information.

And finally, the logic that served as the basis for the prediction. Does that mean the source code? Does that mean the training data that helped create a machine learning system?

So there is-- all of these terms raise a lot of questions. And that's not-- they're all answerable, but this document on its own doesn't necessarily help someone figure out how to answer those. And there are certainly other organizations that are out there that are thinking about these issues, but it sort of creates more work for someone who's trying to think about how to comply with these principles that then they have to start to try to answer as many of these complex questions.

So the OECD recommendations are nonbinding, but the OECD does have monitoring capacity in collecting data. And some of that data is being collected at the recently launched AI Policy Observatory. And I think the OECD themselves, they've created a new-- some new groups to help really think about, how do we move from these principles to practice, recognizing that there are a lot of unanswered questions right now? That these principles, in many ways, do provide a helpful agenda to governments and to organizations that are thinking about how to move forward, but recognize also that they really raise even more questions.

So I think that that's really where we are at this particular moment. So I'll stop there so that we can get to the questions. So thank you.

Thank you so much, Ryan. There are maybe two quick questions, and I ask for short answers. So one is, who is the intended audience or the users of these principles? Maybe Jess can take this one.

And then there is the question by Nagla with regard to the OECD process in particular, whether there have also been non-AI experts involved in the consultations? Jess, do you want to take this one?

So it's interesting, this was actually a piece of data that we collected on each of the documents in our data set, and then ultimately didn't find a great way to represent it in the data visualization. But we were curious what we could glean from the text of the principals themselves about who the anticipated audience was, right? Was it policymakers? Was it individuals-- users or affected folks? Was it companies in the private sector or others? Academics, for example?

So there was a significant variety. I think sort of perhaps most notable ones that are the private sector documents are often-- sort of have two purposes, right? One audience is internal. So organizations, like Google and Microsoft and others that have adopted these types of principles, have also built teams who are responsible for ensuring that the development and deployment of AI within the organization corresponds to the principles that they've circulated.

So the documents are internal facing to some degree, right? They're persuasive and binding, more or less, on the teams within those organizations. But they're also external facing, right? They also have a PR function. And so they're aimed at us and perhaps at policymakers and others, to the extent that, you know, that's the sort of argument that's often made. That when private sector organizations adopt these principles, it is in part to make an argument that the regulation is perhaps less necessary because the organizations are doing it themselves.

So certainly with those documents, you know, I think we see the sort of primary audience as being-- well, I suppose it depends on who you are and what your perspective is, how skeptical you are, whether you think that the primary audience is inside of the private sector organization, or whether you think the primary audience is the kind of PR or staving off policy function.

Jess, I think that's good.

That's good?

Yeah. Thank you. Ryan, do you want to respond to Nagla question please?

Yeah, so I think that the question was about whether non-AI experts were consulted in the process. And actually, I think it's a little bit more of the reverse, that the title of the group, the AI experts governance, you know, it's a little bit of a misleading title. Because I would actually say that most of the participants came from different kinds of policy backgrounds, legal backgrounds. That were certainly some people who were part of the 50 some people in the group who had computer science and AI specific expertise. But the majority actually brought very different perspectives to it. And I can talk more about that later if there are other questions relating to it.

Thank you. That's great. Ryan and Jess, there are a number of other questions coming up in the Q&A box. And if you would be so kind to respond to those that if you can respond right away, we'll return to those in a minute. But I wanted to really ask to the floor, to the virtual floor, our respondents. Mutale, you are the CEO of AI For the People. You're a fellow at the Berkman Klein center. You have done a lot of work focusing essentially on the unequal impact that these next generation technologies have on people from different backgrounds and different circumstances and in different geographies.

And I was wondering, you know, as Ryan said, there are all these open questions. And now, where do we take it from here? And how do you see your work fitting into this question of principles and practice of AI?

My first question was, the report was very rights based, in terms of how you're thinking about AI going forward. Much of my work looks at equity, which is a different type of lens. And given that we have shorter time, and I could probably speak about this for seven years, I was wondering whether there was any conversation around equity? And I'm thinking specifically of negatively racialized communities that want rights, but also have this equity deficit?

And also, very fast, do you want to add two sentences about your work? And Jess, we will collect a few statements and then open it back up.

Yeah. I can qualify why equity was so important. For me, I was doing this work. I was a practitioner in Congress, the US House of Representatives. And I spent a lot of education time really letting lawmakers in the US know that often human rights frameworks are not looking at the reality. That if we ignore a phenomenon like anti-black racism and how that impacts the deployment of AI-- I'm thinking, specifically in this moment of corona, I live in a city where the police are now thinking about using biometric technologies to figure out who's social distancing.

Finding somebody who is white and rich is very different from finding somebody who is poor and black, but it's one principle. And that's just a very concrete example of why thinking about equity in terms of impact is something that I very much dedicated my work towards. And I would love to know whether this was a consideration in the report.

Thank you so much, Mutale. And we'll return to that after hearing from a few others. But also, thank you for your important work in this field. It's really great to learn from you.

And I was wondering whether Vivek, who joined us in the meantime, who is a law professor at the University of Ottawa, where he leads the Canadian internet policy and public interest clinic, also former Berkmanite. You've done a lot of work from a human rights perspective, working with companies and thinking hard about the intersection of technology and human rights and building on some of these questions that were so nicely framed by Mutale. I was wondering where you see some sort of the rights frameworks coming into play, but also possible limitations of such a framework? And it's good to see you, by the way.

Yes, it's great to see you under these interesting circumstances. So I'm delighted to be here virtually. Yeah, so I've worked, again, on technology and human rights for a long time. And I see the value of human rights, because I've seen them actually be quite transformative inside companies with regard to, let's call it, the Web 2.0 set of human rights problems.

And I think companies are grappling with, what do we do with this very wide set of technologies? Right? Our algorithmic systems and AI, if you think about it, are a cluster of technologies. And a lot of the human rights impacts are particular to the use of that technology in a use case, right? And that presents a challenge for companies that are trying to use their, let's say, global network initiative-- and that's something I've been very involved in-- era tools to assess human rights risk and apply it to this open ended set of systems.

So just a step back though. I mean, I think there's a lot of value in human rights approaches to AI. And I'm really glad to see-- I think the report that Jess and colleagues put out is incredibly helpful, the visualization, just in sort of showing how different human rights conceptions are found in the different principles is incredibly useful as a descriptive measure to show what the landscape is.

But to me, from a normative perspective, the value of human rights, even in these challenged times, is that they do provide a baseline set of understandings. You know, they're law, for one

thing, that states have generally accepted, that they feel that there's obligation to respect these things. And there's a normative framework there-- a common way to talk about problems, even if we don't agree with what the solutions are.

Now I think the hard challenge-- and I've alluded to this before, I think this is reflected in the title of this event today-- is, how do we take those various articulations of principles? High level human rights principles, the more granular principles at the OECD and companies that your report has shown? And provide practical guidance to different actors in the AI stack as to what their human rights responsibilities are based on differential impacts and different use cases.

And I think that's a nut we haven't cracked yet. And it's really difficult nut to crack because of the diversity of the tools and of the use cases, right? So I think that it's a really difficult human rights challenge that we're at the early stages of thinking about. But I actually think that all the work that's happening makes me quite hopeful. Because we are thinking about it at an early-- well, still relatively early stage of the development and implementation of these technologies. I'll leave it there.

That's very helpful, and also a good reminder how much context matters in these discussions and maybe a great segue to Doaa, who's an SJD candidate here at the law school and has done a lot of work, not only actually at the OECD, but also studying the use of algorithms in the criminal justice system.

And I was wondering, Doaa, whether that is an initial use case, whether it's AI properly or algorithm or where of course many of these questions around fairness and transparency and bias are applied at play. And with Vivek's reminder that this is still relatively early stage. How helpful are principles like the ones we've been discussing today in the work that's front and center in your research?

Thank you, Urs. And hi everyone. I'm happy to be here. It's funny. They say if you have lemons, make a lemonade. So in the normal world, I probably wouldn't be able to join, because I'm not in Cambridge. But it's good to have the opportunity to be here.

I want to divide my observation into two. In the first one, I will wear the hat of someone who was working in the OECD part of the time where the principles were considered and the expert group was meeting. And the second is like, the academic hat of someone who is thinking about regulation of AI.

So for the first part-- so as Ryan said, the OECD operates on the basis of consensus. Reaching a consensus between so many members and other actors in the field, it is a hard job.

So the principles are broad indeed. And one of my favorite exercises is to give this list of principles to someone who is from the field of computer science and a CEO, what do you make out of it? And usually, the answer is like, not much.

But I want to emphasize that although it is hard, that principles raise more questions than answers, perhaps it's a good thing. Because we don't want to be too limiting in the approach that-

- especially if so many countries are adopting those principles, we want to give the room for either countries or companies and everyone to adapt to their needs.

But the OECD-- but the principles, so they have a very important, I think, declaratory purpose. They kind of put the important topics upfront, and then the implementation into practice will be discussed later on.

As probably some of you recall, the OECD has been very powerful in shaping regulation around privacy. Starting from the 1980s, the concept similar in the same way that AI principles were adopted in a council recommendation, there was a council recommendation of privacy. And this has been shaping the privacy regulation around the world massively. So it can have an impact.

Now as an academic who is thinking about the regulation of AI, I think the hard part is to kind of think, how do we balance between all the principles? And not just how-- I tried to look at several case studies, from criminal justice, from welfare, and to how to not just unpack each one of the principles, but how to balance between all the principles? Are they all-- is there any hierarchy between the principles? Is there any difference? Where is all the discussion that we having the legal world about checks and balances?

Let's say that, for example, I'm doing very well on the transparency in a certain case. Do I need to comply with the others? Similarly, I think what is lacking at this point, and I'm hoping that was the development of more and more case studies, is that we'll see this conversation being developed. Not just only what does fairness mean in each context, but how to kind of look at all the prints-- yeah, the guidelines simultaneously, the requirements simultaneously, and what to make out of that. I'll stop here.

This is super helpful. Thanks for sharing your perspectives both from the inside and as a researcher. It's great to have you here.

We have a number of comments and questions. And before giving it back to Jess and Ryan for some sort of a concluding remark, since you have only eight minutes left, I wanted to maybe ask Padmashree to share your observations that you also put into Q&A. But it's just also nice to have your voice live here. And then Amy Johnson will be next. And I hope I can unmute you. Padma, are you here?

Yes, yes. Can you hear me?

Yes, very well.

Great. Thank you, Urs. And thank you everybody, especially Ryan and Jessica for presenting. And I'm really happy to be part of this.

So the question I asked-- and I think that maybe-- and I think it's a relevant point, is that, whether we should reflect on if a human rights framework is not contextualized, in terms of new inequalities, considering that a lot of the human rights principles that we are considering today,

they were created 70 years ago, or 70 or so years ago. And there are new kinds of iniquities and inequalities in society.

Whereas you, yourself, have written a paper about digital rights and the different kinds of digital rights that have been actually proposed. Then I think that there is a need to have a discussion on what kind of new thinking we need about contextualizing a human rights framework to AI and inequalities and iniquities of the kind we see now. And what about the people who are normally not part of this conversation? How do we take them on board?

Thank you, Padma. And Dennis Redeker who's the lead author of the paper you kindly mentioned is also on this call. I'm so sorry we have so little time, but we will hopefully weave it all together. So Amy, do you want to share your question or your thoughts as well? And then I'll turn it over to Jess and Ryan.

I was wondering about the question around the adversely affected. The scales of these systems, and the effects of these systems, are so large that it seems odd to me that the only form of challenge would be the person who is directly harmed. And so I'm curious if there is consideration of other folks, whether it's a bystander intervention style, or some other kind of method of challenge. If that was under consideration, and if not, why not?

Excellent. Thank you. So Jess and Ryan, no small task, in five minutes, to talk about rights-based versus equity-based approaches, question of hierarchy among principles. And ultimately, also, the question of interventions and bystanders Over to you.

Oh, yeah. Well, it is a tall order in five minutes. I'll just-- coming out of this report, which of course is like a deep dive into the principles themselves, I mentioned a couple of principles that came up for me in this conversation. So one, under fairness and non-discrimination, we have a principal called inclusiveness and design. It was really interesting to see how some of the different principles documents interpreted it.

So some of them interpret that idea as basically just like we should build more diverse design teams, right? Include more women and minorities on design teams. But there are a few documents of which the IEEE, ethically acceptable design is the-- ethically aligned design is the perhaps the primary one that actually think about inclusiveness in design in more of an equity than a rights framework. So it's not about the design teams. It's actually about designing technologies such that they allow for broader participation than the present state of things.

And while Mutale brought a race-based frame to it, which I think is incredibly important, the case that the IEEE document highlights is actually disability rights. So how could AI technologies be designed to build a world that's more inclusive for folks who are hearing or vision challenged?

So that's one principle that I wanted to bring up. From the perspective of Amy's question, which is, I think, a wonderful one, and highlights a shortcoming in sort of current US law. At least speaking as a US lawyer, right, where if you have a sort of very small harm as a member of a

large group who are harmed in small ways, you sometimes have standing problems in actually bringing a lawsuit to enforce those rights.

The sort of equivalent of bystander type enforcement that we observed is that quite a number of documents-- I think almost half the documents in the data set recommend some sort of exterior audit or evaluatory function. And it's interesting to think about how that might take shape in various jurisdictions around the world, whether administratively or otherwise. So that's, I think, a sort of space to watch and a group that could function in that kind of bystander role with civil servants.

Thank you, Jess. Ryan, your thoughts on some of these questions?

Thanks. So I actually wanted to come back to something that Doaa said. Because I think it's a really important point to emphasize is that, I think when you look at, for instance, the OECD principles, it's really a stake in the ground. It's not an end point in itself. But I think it provides direction, hopefully, to nations, to other nations and to organizations that are thinking about these kinds of questions.

And I think, really, that's what a lot of these principles are. And so I think that one of-- there's sort of two interesting things. One relates to Amy's question, that when you read these principles, there is both the sort of-- the specifics of, you know, why isn't there bystander rights? But viewed two steps back, looking at it as this sort of marker, it's really more of a question of, how do we build effective accountability mechanisms?

And it provides some ideas on how to do that, but there is obviously many ways to actually implement it in practice. And things like bystander rights may be part of a more comprehensive approach.

And I think, related to that, my other observation is that I think one thing that I'm interested in seeing is, to what extent do things-- as we move from these principles to practice, to what extent can they become differentiated in different contexts?

So there may be some cases where the right to challenge as articulated in the OECD principles is actually good enough. But there may be other domains, other areas of application, where a very different approach is going to be necessary.

And so I think it'll be interesting to watch going forward whether there are sort of forks taken in the road and different approaches that vary depending on context, and how these sets of principles are used or not in thinking about those differentiations.

Thank you, Ryan. And I also want to acknowledge we have a number of additional questions and inputs on the Q&A window, and I'm sorry that we're running out of time.

I want to acknowledge, however briefly, what Julie wrote there. And which may be one of these contexts, Ryan, is talking about real time. What's happening now in the COVID 19 crisis? How

will AI be deployed to combat this particular public health crisis? How robust are these safeguards, or at least norm statements that Jess was presenting and Ryan was discussing too?

So I do think this is kind of a first dramatic real world test of what we have been discussing a little bit in the abstract today, but that gets concrete very quickly.

So I'm sorry we're running out of time, but it was a wonderful one hour with all of you. Thanks so much, Jess and Ryan. and Mutale, Doaa, Vivek, the entire BKC team who made this possible. Thanks to all the participants for listening in, and please be in touch, and stay safe and be well. Thank you.