FINAL COPY

ETHICAL DEVELOPMENT OF AI

JUNE 8, 2017

Services Provided By:
Caption First, Inc.
P.O. Box 3066
Monument, CO 80132
1-877-825-5234
+001-719-482-9835
www.captionfirst.com

***

***

>> MODERATOR: Good morning, everyone. Let's get started if we could, please.

My name is Rob Kirkpatrick. We have a really exciting discussion for you today on ethical development of AI, and we have a number of the smartest people anywhere on the subject here joining me on stage. We are going to have today two speakers, we'll present initially on some of the principle challenges around ethical development of artificial intelligence and these are Dr. Luka Omladic, who is here with us who is a member of the UNESCO Commission on Science and Technology as well as professor Lorna McGregor, director of the Human Rights of the University of Essex.

After their remarks we'll have some initial responses from three panelists, one from IBM, Francesca Rossi, a Scholar for the Interdisciplinary Center for Bioethics at Yale University, Wendell Wallach, and finally on your far right, Executive Director of Communication and Governance at the National Law from New Delhi, Chinmayi Arun. We'll have questions and answers to them and open it up to -- to the floor.

To my immediate right is Andy, Board Chair for the IEEE

Computer Society who is serving as our Rapporteur.

Our mission for today -- no pressure -- it is to come up with essentially a consensus of recommendations on how to move forward with ethical development of AI and we'll be reporting back on these after this session to the plenary and then collating them and refining them in coming days.

Andy, if you can give us an overall sense of what we need to accomplish out of this and, please, in the course of the discussion jump in at any time, interrupt, make sure that you're getting the information you need.

>> Is this work?

>> I think it is on.

>> Can you hear me all right? Great.

My job is easy, to make you all work. Today what we're trying to do is come up with the guideline recommendations. And in there I want to be more specific.

We want to know what focus will be in there and we want to bring some use cases. So if you could put forward discussions or use cases as part of the recommendations and we want to know the area of policies and how long do we think timing wise is it going to be short term, long-term, a year, two years, how many years and what kind of data we need to gather all of that information.

The other is resources: What type of resources, money, experts do we need? And who would like to contribute, partner with us, who would like to be a part of this plan going forward and last part is measurements so we know how we're progressing.

>> MODERATOR: Thank you.

By way of background, I mean, this session, you have seen the overall description, there are a number of ethical concerns that have been raised in recent years and certainly yesterday we got pretty deep into these conversations. Questions whether AI should be able to make life and death decisions, some possible harms that could result from it and how they can be mitigated. Why are we talking about there in terms of ethics, why is law not enough? I think we have to think early now not only about designing the technologies but how to adapt to living in a world inhabited by our official intelligence and how this changes us, an ethical discussion in parallel with the legal discussions around rights is critical. I think -- we came to this space out of work with big data and privacy and now of course there is a lot of interest in Artificial Intelligence that's more than about the analysis of big data we got into the conversation by looking at one of the central paradoxes around the use of big data which is that

there are right issues related to misuse of the data in terms of privacy but there are right issues related to the non-profit entities use of that data. In terms of the dependency we'll have on the Sustainable Development Goals for achieving them on the use of the data where it is not currently being used.

We have spoken over the last 24 hours a lot about automation and jobs, for example, about life and death decisions made by vehicles, autonomous weapons, and how to strike a balance over advertising, public profit and public goods, misuse for surveillance and the real risk that the technologies will accelerate and amplify in quantity around the world.

What I would like to do is turn first to Luka. If you would, take us away on your thoughts on what the challenges are that we're facing.

>> LUKA OMLADIC: Can you run the presentation?

You can go to the next one.

Thank you very much for being here. I'm representing here the UNESCO's Commission on the Ethics of Science and Technology. UNESCO is an organization that puts a lot of effort into educating in the fields of ethics and disseminating different materials.

Here in this slide you have some of the basic tasks that we do. For example, we want to serve as a laboratory of ideas, the World Commission on Ethics and the comment is a group of experts composed interdisciplinary. There are engineers and philosophers, and some of them are both engineers and philosophers. Unfortunately I'm not one of them.

Of course we want also to serve as a standard center. For example, there are a couple of UNESCO declarations that were conceived as a work of COMES or sister organization by the International Ethical Committee. We hope that perhaps someday our work can just -- can you proceed to the next slide -- our work on the robotic ethics which is currently our task, we're composing the report on robotic ethics that should be adopted later this year and maybe sometime there will be a possibility for UNESCO to draft maybe kind of a declaration on Artificial Intelligence or robotics.

At this time, for example, the UNESCO members are in the process of accepting the declaration on ethics of climate change which is based on work we have done in the last two years.

In this slide you have the basic idea what topics do we cover in this New Report on the robotic ethics.

Again, I composed this this morning. Maybe it is messed

up. I think just for illustration it is really important when we speak of the ethics of AI and robotics to have in mind that we're moving in this field of imaginary and real of realities and possibilities because the whole idea of creating artificial being is not something new, it is something that's embedded in the human cultures. We have, for example, different cultures, so on.

When you try to draft an ethical principle or recommendations we have to bear in mind that some things are real, some are imaginary and some things are conscience and some things are really unconscious in our relations to robots and AI.

Some things from our report -- we can proceed.

I want to comment briefly on that.

That's my -- the advantage point of our Commission's work on robot ethics was first we tried to define this different position and two major positions have taken place, techno-optimism. You notice there is a large tribe of techno-optimists in our conference, and in this general definition techno-optimism expects a better future through technology and so on.

Then we have a strong version of techno-optimism which is transhumanism, human enhancement, replacement and we have heard a lot about that.

On the other hand, techno-pessimism, it is very suspicious about the technology and its strongest extreme position by bio conservatism. If we proceed, what we try to do is adopt a middle ground and we recognize that in this approach of value Dynamism. This is a concept where we recognize the technologies do not only have societal -- direct societal effects that can be ethically evaluated, but they also effect the very ethical frameworks with which we evaluate them. We change -- we evaluate the technologies but the technologies change our values. It is a two-way process and it is important for this to be recognized. We can call this is techno-moral change, and an aspect of this is that we have to proceed with the value sensitivity design which I will mention briefly later.

Next slide, please.

So proceeding from that, we recognize -- we recognize some of the really relevant ethical principles that apply to robotics and AI and here I would mention the four of them, for example, do not harm principle, which is very well-known in its form and presented by many and so on and we heard yesterday how important it is to not construct the robots or the AI that will decide how to kill or not to kill human being and we all agree with that. Then again

when you think of, for example, the classic dilemma with the automatic cars that has to decide whether to kill the passenger or whether to kill the bypassers, then you have this problem immediately when car decides about killing, taking lives of the humans. Principle of autonomy, of course, we should mention, it is human autonomy. It is not obvious anymore.

We're starting to think about if you move to the register more or less the imaginary but starting to think about robotic autonomy, AI autonomy, principle of responsibility and liability is extremely important. It covers the fields of privacy, traceability and last principle of proportionality because the technology must take into account the social and cultural context. The assessment and implementation must take into account the culture and social context.

This is very, very much UNESCO DNA. Their idea.

To move on quickly, there are a couple of recommendations that are isolated and maybe we can discuss them later, recommendation of value sensitivity design, recommendation on experimentation that was drafted in the context of robotics, but it also applies very much to AI, recommendation on experimentation. It is not obvious in today's engineering environment many times robotic and also AI or IT technologists, speaking more broadly, they're applied without the standards that apply for a convention of technologies like chemical technology or other fields of engineering. That's something that we'll have to be probably change with advancement of the technologies.

Next one.

Again, public discussion on education. This is one of the UNESCO's core missions and we try to apply them to this field.

Next one -- I will skip that, maybe we can -- if there is a time for discussion. There is technology-based ethical framework that our colleague engineers are working in our ethical commission developed. Maybe we'll have a chance to talk about this later.

Next slide.

I will just end with this, I think you recognize this quote. I couldn't help it. It is a quote from Frankenstein, he finished writing exactly 200 years ago and it takes place here in Geneva. The story takes largely place here in Geneva. And was also written in Byron's Vila near the Geneva Lake and the quotation shows this ambiguity, on one side, the huge optimistic expectations of benefits that AI will bring to us, to its creators. On the

other hand, anxiety which is later in the Frankenstein story justified of completely unexpected relations of this new entity to its creator and perhaps the essence of this ethical urgency we're talking about is right here in this quotation.

Thank you.

>> MODERATOR: Thank you. Thank you.

I think it is -- like Frank Stein, all of this endeavor we're currently in is essentially a metaphor for parenthood, and one hopes that our digital offspring will love us as all parents do hope of their children -- for giving us for requiting on them unwittingly the same flaws we find in ourselves.

With that, over to you.

>> LORNA McGREGOR: I want to highlight three challenges.

The first challenge that's important to discuss is why we need an ethical development of the AI in the first place; the second is what do we mean by an ethical development; and the third, how and who develops this ethical approach.

The first thing, we need the ethical development AI, and I come at this from the same type of background as you were talking Robert, I codirect a project on Human Rights big data and technology. I have come from a Human Rights perspective, but also big data perspective. I think that it sounds like an obvious question why do we need an ethical approach.

From the perspective of my project we feel that before we can start talking about what that ethical approach looks like we really need to define what we're responding to in the first place. It may seem obvious to us but for the stakeholders we need to engage we have to define the problem.

In the Human Rights sphere the risks are often presented quite narrowly and in a fragmented way. We often see the risks presented as just a privacy risk and traditionally privacy is not seen in a really serious way or it is a security issue or we're talking about algorithmic discrimination. We see the risks, they're very true, but they have the potential to narrow and compartmentalize the potential impact of AI and big data and therefore the responses we see. We think it is really important that we understand this as a fundamental paradigm shift. That it is a risk to all rights, and it strikes really at the core of what it means to be a human being and the values that underpin Human Rights so to dignity and how society

functions. That's the baseline for us in understanding what we're talking about.

We're talking about this in relation to risks, but we also have to talk about this when thinking about AI for good. It is often easy to think because there is a good object that we don't have to take the risks as seriously or we can balance the risks because as Robert says, there also is the rights issue if you don't use AI or big data. This is really the fundamental starting point, how do we understand this massive paradigm shift and of course within that we're talking about machines that don't reason in the way we reason.

There is a big risk that we start to think about AI as ourselves when the pattern of reasoning is completely different and with increasing autonomy that AI systems have. That comes to the challenge of how do we future-proof and move at a speed that we need to move at to make sure any ethical responses are relevant to the AI and big data challenges we face now and very quickly in the future. So that's the first challenge.

When we're clear on the risks and the challenges, this very much shapes the scale and nature of the ethical development and that goes to the question of what do we mean when we talk about ethical development.

Now, from a University perspective, but also from a company perspective, we could be thinking about ethical approaches in terms of gateways to enable us to do what we want to do. Gateways for AI for good for example. As long as we have X, Y, Z in place and traditionally we think about things like informed consent which don't work so well in this area then we can go ahead. It is PSW guessing past the risks in order to do what we want or a compliance approach. Now, this approach is probably not the one we're looking for right now. We're really looking for a much more substantive approach to what we mean by ethics.

We need to define what we mean by an ethical approach. That's by understanding when the risks are too great and I think you talked about this yesterday, when are the risks too great that we start -- that we can't get off the starting line even if it is offering us really great potential. Then when there are risks, how do we actually mitigate those.

When we think about it, we can talk about lots of different approaches. There is a lot of philosophical discussion and what we mean by ethics. At least from our perspective, Human Rights-based approach has to feature within that. In a very general way we can think about

Human Rights as legal expressions of ethical principles that go together with ethics so we think about how Human Rights and medical ethics go together. For example, so they work together, but Human Rights provide that legal obligation and teeth and particularly the accountability aspects there. There are lots of forums we're hearing that the international Human Rights law framework is obsolete or isn't good enough to deal with these AI and big data challenges. I think that we would push back on that and say that it is not that we don't have the principles, not that we don't have the procedural ideas, it is no the that we don't have frameworks like corporations, but what we do have is a big challenge in terms of how do we apply the international Human Rights law framework alongside ethical principles to new and incredibly different environment and with this huge paradigm shift.

So that may mean questions like what's privacy mean now? Do we have to give heightened status to privacy because it becomes a gateway to the violation of every other right in certain circumstances. It also makes us think much more about prevention which we haven't traditionally thought so much about in Human Rights but we have seen much greater movement on that. How do we prevent and that goes to ideas of how do we design AI? How do we make sure that the design phase that we really embed in the principles? How do we think about internal and external over site and work with issues of transparency while understanding that we have to protect our algorithm secrets, how do we work on remedies? That issue doesn't come up much, and if it does come up in the discussions, remedies are spoken about in terms of accountability or how do we make sure that AI doesn't do this again.

What Human Rights brings to the table is how do we think of remedies for the individuals and the groups effected. That's a critical question as well that we need to bring in. Of course how do we understand responsibility here, particularly given the complexity of human to machine state of business, how do we think about this and how do we ensure we don't hide behind the machines and say that we can't have responsibilities, too difficult to have accountability for the humans and the states and the corporations and this process.

Again, I think principles like due diligence, the framework design will give us ideas on how to approach this. How we make sure that the prevention and the oversight principles are incorporated properly so there is a possibility to think about responsibility ahead of time

rather than retrospectively.

That goes to the third challenge about how and who develops the approach. We have many, many actors working on ethical approaches to AI and big data. That's really important because we are behind the curve, and we're behind the pace of which AI and big data is being able to impact on rights and affect us as humans.

So we see the IEEE initiatives, other initiative it is at Harvard, general level, these are crucial principles to build on we're seeing a lots of particularized approaches to ethics, we see a lot of work on how to use social media, how to use open data and what the ethical approaches are to that to enable us to document Human Rights violations or to respond to humanitarian crisis h we're also seeing, for example, things like the UNHCR guidelines on personal data that are looked at as internal guidelines but looked as models externally and all are crucial and it is fantastic we see agencies and actors thinking about this. We definitely need to a process to bring it together so that we can be clear on where we're going with ethical processes.

I think that the comment made in the panel yesterday and the final panel about thinking about AI and big data as a global challenge like climate change and thinking about global alliance platforms to bring together all of the initiatives, one so we can share approaches, so we can learn, what are we thinking about ethics and Human Rights on algorithms, how does that compare to autonomous weapon systems so we can share an infused approach to different communities but so that we can also get to shared perspective on what the legal obligations and ethics are to move forward and what Luka has said is an urgent situation.

Thank you.

>> MODERATOR: Many, many thanks, Lorna.

Let's turn to our panelist for some ideas and recommendations and examples of what they're already doing and thinking about in this space in terms of the kinds of solutions that are going to be needed.

We will start with Francesca, if you can give us some thoughts on these challenges and what you're seeing from within your work.

>> FRANCESCA ROSSI: Were you some of the things that were said resonate with me very much.

I think I try to understand how to make AI system behave ethically once deployed. I also work within a company like IBM that, you know, as a whole processing place to design and develop AI systems and as data policies already in

place, even for, you know, the use of big data for big planning. So also in that space, in that operational and transformation space, I think that there is a lot of work that I do in trying to understand how to embed and identify values to be put, for example, in the design and development process as a first-class entity and not just something that you put in at the end once the project is done and ready to be deployed.

I wonder how we can embed ethical principles into AI systems and the main features of an ethical behaving AI system, of course it has to be able to follow some ethical principles you have to explicitly identify and state, and not just identify and state but these have to be so explicitly represented so that everyone can inquire of the AI system what principles are following and what values it is showing and this should be done by an external party or also by the human using that AI system in everyday life. Then, so it should not just follow this principle of value that's been identified but, you know, people should check compliance but also I think in terms of the ethical principle the AI system should behave also in a not just reactive mode in terms of complying to the principles but in the proactive mode in the sense of in my view again it is just a system working with a human.

If the human, we think of the human as the final decision maker in this human-plus machine system then we think that the AI system should be able to alert the human if it is deviating from the ethical principle in to the decision-making process and possibly even suggesting more ethical decisions or actions to be taken. I see an ethical AI system not just being ethical itself, but also helping humans to be more ethical even just because the AI system can look at much more data so can have much more information and knowledge on what to base the decision. That already would help, you know, making a more ethical decision sometimes. Even more than that, you know, he should have a very explicit, you know, set of principles to follow. The question is which principles will you put in the AI system and everybody understands that there is no universal set of principles because they vary a lot, they're based on society and culture that you want to develop the system or the specific task, on the specific scenario, again just culture is a very big discriminator of the principle.

For example, if you want to deploy a companion robot for an elderly person and you want to deploy it in the U.S., Europe, Japan, the three societies have very different ways

of relating with robots and not so relating with elderly people. On advice, the same kind of companion can be deployed in three different countries for that to follow different social norms and possibly different ethical principles.

I agree with you that we don't have to think about being compliant to existing laws, laws change over time. Laws are different in different geographical things. We have to go beyond that. We want to have even in developing AI system, we want to be able to be very proactive and go beyond what the law requires and define some code of conducts of principles for developing these AI systems and I think that the on the initiative on the global initiative on ethical considerations for AI, it is a great example in putting together very multidisciplinary group of people, more than 100 people thinking about these topics and trying to understand what it means for developers to think about this issue and to take this issue into the development and design process, and as you may know, the initiative as generated, a very significant document of more than 100 pages with -- divided into eight, nine different topics with a long list of issues and possible recommendations to be followed, to be discussed is just version one of the document that you may find on the web and I think that's really the right approach. Very multidisciplinary environment for discussion, AI people like me can help in understanding how to overcome the technical challenges and alone we cannot solve everything. The whole feature on how to make AI behave ethically and how to ethically develop AI system. So really, a multidisciplinary discussion is really needed.

Another initiative that I want to mention again, that also is focused on multidisciplinary and the developing of best practices with discussion among corporate, non-corporate environments is the partnership on AI and again that's a different approach but I think that's very important. Corporate entities that develop and deploy AI systems and put them in the real world are those that make an impact to the real world but they need to work together with academia, policymakers and, you know, regulators and media and the public to really understand the issues.

The last thing I want to mention that resonates with me very much is this need for educational efforts to be addressed to every stakeholder, so policymakers, media, public, business executives, so on, to make them aware of the real culpabilities of AI, but very realistic knowledge of culpabilities and value of AI and how we embed the

values in AI and what values we want the AI system to be aligned with us and helping us to make better decisions.

>> MODERATOR: A question, your suggestion on the anticipating more proactive approach is interesting. Could you give a concrete example of the kind of system where and how that would be applied?

>> FRANCESCA ROSSI: In healthcare we know that doctors -- in every profession the professional should follow the accepted code of conduct, if you want ethical principles that the professionals have to follow, sometimes we're not very good at taking some ethical principles we think are good and bring them into practice. We have several biases in making decisions and sometimes the bias brought us to deviate, to bring us to deviate from our principles whether they are ethical or other kinds of principles we follow.

I think machines can help us to understand the biases that we have in our decision-making process and to alert us for deviation. For example, I know even in our company that there are people that are trying to develop tools for developers that help -- that alert them when they are maybe unintentionally introducing some bias into the use of data in developing an AI system and that I think is very helpful because, of course, humans are not perfect and not machines as well of course are not perfect, but the fact that they can explicitly represent if we succeed in doing that ethical principles, they can really alert us and help us to be more ethical.

>> LUKA OMLADIC: I think what you mentioned was very interesting, but I think it is a very subtle thing and people don't distinguish it, and we talk about AI and AS, I would like you to explore more on your thinking of the Artificial Intelligence versus autonomous systems.

>> FRANCESCA ROSSI: In my view, this system to behave ethically is not just for autonomous decisions. You need to make sure that they right the make decision, but even in systems that are not autonomous but work together with the human where the human is the final decision maker, even in that case, even more so you want the system to work following values that are aligned to the human ones. You want, again, explanation. You want to make sure that the human understands that the system is aligned with the values. Otherwise you could not have the necessary theme work between the human and the system that allows for the overall hybrid human plus machine system to make the best decisions.

>> LUKA OMLADIC: That's a good analogy.

I think sometimes we refer to thinking of if you use a

Google maps and you were driving and it has intelligence to tell you don't go there, it is the traffic, it is slow, you have to turn to the right in real-time data and be able to advise you, the intelligence, but when you have a car that drives for you, the integrity will really be different in those systems.

>> MODERATOR: Wendell, what are your thoughts?

>> WENDELL WALLACH: (Off microphone).

We're trying to maximize the benefits of emerging technologies and we minimize the risks and societal impacts of those technologies. When you really get into this, this is a vast subject area. We're talking about technologies, AI alone which is going to touch nearly every facet of modern life let alone that we're not talking about biotech knowledge, geo engineering, nano tech that will also have impacts.

Back in the 1980s it was underscored that it is easiest to shape the development of a technology early on, but early on you don't necessarily know what the problems are. Oftentimes by the time you do know the problems the technology has become so entrenched in the society that there is very little you can do. This principle has bedeviled technology policy around the world for the past 35 years. Those of it us engaged in anticipating governance, responsible research and innovation, we reject the simple logic and suggest that there are opportunities when we start to recognize the impact of the technologies before they're fully entrenched. I refer to these as inflection forms.

The difficulty is whether we recognize the inflection points as they come along and we take those opportunities to shape the development of a technology.

Taking that opportunity through any means available, whether that is techno solutionism or whether that requires ethics, whether it requires governance, there are many different methods and the other problem is recognizing the inflection point and looking at it, studying it and recognizing what is the most appropriate response when that occurs.

Now let me talk specifically about AI. We are at an interesting inflection point in the development of it. One of the books I have read that had impact in this area with public years ago, moral machines, teaching robots the difference between right and wrong, we were talking about a problem about how you may view sensitivity to ethical considerations and computers and robots would factor those in to the choices and actions. We are now just beginning

to move into the inflection point where that's a real challenge and no longer just one for us philosophers.

To me, the important point was this underscoring it early so that when we come to this stage it can begin to be recognized and we can begin to look at what the various responses are and what the various pathways are within that. Now in the area of Artificial Intelligence a few other inflection points have just come to the surface. One, algorithm bias, it can predetermine what are the outputs of the AI analysis. So now we have to look at how do we understand the implicit biases and inputs and therefore what the biases are in the outputs.

Can we do that technologically, what kind of tools do we have to develop for that? If we can't, what restriction should that put on the deployment of the technology?

The second area that's come up is whether -- if something goes wrong, can the technology itself, can we explain why that accident happened? We cannot do that with people learning algorithms at this moment. There is no transparency and it becomes really much a problem when it comes to forensics, looking after the fact why an accident occurred but perhaps more importantly is which technology should and should we not be deploying. Should we deploy technologies that we not predict what the impact will be? Do we have the approaches or standards in place for determining when or where they should be deployed? These are complex adaptive technologies. These are no longer simply predictive and not only that, the important turn that's happened is these are being introduced into social contents. That's why all of these other concerns, the other speakers have underscored, they're suddenly so important.

Perhaps some of these technologies shouldn't be deployed, particularly if we can't predict their behavior or the behavior has a potential to cause great harm. Suddenly we're talking about issues with self-driving cars, with lethal autonomous weapons, with technological unemployment, all kinds of impacts that we're cognizant of the impacts or we will miss inflection points and not give adequate attention to what measures we may take to shape the development and deployment of the technologies.

I could go on and on, but I think I have got the basis of this out there.

Just this one transition point for our next speaker, these technologies are largely being formed by the wealthy north. As in most technologies the benefits are largely going to those who are in a position to take advantage of

them and the detriment as in many areas also go to the same group. Those that have acquired the benefits are not necessarily those who are paying the price. It is often the poor or those that will lose their jobs because of some disruption that we need to give great concern to and I think we particularly need to give great concern to it of those in what's often talked about as the South or the emerging tech companies and the countries with large middle classes and they have even larger poor classes.

>> MODERATOR: Following up on one point, this issue about transparency, I mean, we heard yesterday, you know, a suggestion that we may have to be content with AIs that are as vague in answers as humans tend to be in terms of questions of why a particular decision was made. I guess in the case of humans didn't design the human mind whereas, you know, we're talking about designing an Artificial Intelligence with the idea that reflection and the ability to make explicit the nature of decision-making processes is one that one could think about.

The problem is, as Lorna said, the technologies don't -- we're not talking about something that reasons the way we do. When you ask why, you know, people for as long as there is written literature have lamented the limitation of language itself and there are famous examples in human language like knots, you can't explain in words how to tie a knot and have certainty that they'll tie the same knot that you tied. Language is not good at that. It’s good at other things, but not that.

Of course here we're talking about liabilities at the end of the day, among other things, and do you believe that if it is not possible for an Artificial Intelligence to be transparent to some standard that it should be a case of what Lorna referred to as you cannot proceed with that because it is too dangerous?

>> WENDELL WALLACH: You outlined that well.

The system can't be transparent. We don't get the benefits from them. Does it matter? Let me give you a concrete example that I think helps understand this. If a learning algorithm is helping a self-driving car understand the environments better, and we think that self-driving cars are going to lower overall fatalities and accidents in general, then this is a good thing. Whether it explains itself or not, I don't think any of us have a great deal of problem with that.

Consider the situation now with a self-driving car killed a pedestrian or had an accident where an aware human driver wouldn't have made the same mistake. Then it

becomes really crucial. Even if it cannot explain the activity that we have the forensic capability of understanding why that occurred and you have to understand what is beneficial or not and which have implicit biases in them that could have societal impacts and which could have destructive activities that we at least need to have the forensic capability, if not eventually the capability for those systems to explain their own intentions, even if they, like us, come up with rationalizations after the fact that are not true.

>> MODERATOR: Finally, over to you: Having the best AIs in the world will be available to the wealthiest in the world and those are the ones least effected by issues like job displacement, so what are your thoughts?

>> CHINMAYI ARUN: Thank you.

It is a difficult act to follow when you speak after multiple people that made solid points like all of you. Let me address the question you asked me.

Most of the conversations that I have been a part of on AI have tended to talk about it like it is developing in a borderline, sterile environment and they have not thought about it in terms of global capitalism which is more about how AI will likely impact. It is not just about developing technology and the it technology impacting sort of the uniform people but it is about who has the money to market the technology, which country's laws are they subject to, whip countries are they marketed to and whose data are they accessing and programming the technology and to give you a small example of how this works, very little things like research and medical technology, right, the global pharmaceutical company doing research on drugs is often looking around and finding that the global north does not permit certain types of experimentation, where do they go to conduct the experiments?

To the global South, to countries with poor regulation, people develop drugs and products ranging from not just medical but even to facial recognition and find that they can't use what is they're trying out in the countries. Again, where they go is to countries with lower threshold of regulation, countries that are less likely to catch them, sometimes it is marketed as Internet for the global South, free access for everyone. The question to ask and I'm really happy to see it is being asked here, it is that can the countries that are developing ethical norms and legal norms for on the development of technology disown its impact in other countries.

It is difficult to deny that the technology has impact

in other countries. As global capitalism and technology develops a very international dimension, I think it is not a question that we can run away from anymore. I'm happy to ask it at the ITU, so many U.N. agencies, it is a great place to do it. Basically the cross-border impact is something we have to think about from the start.

What form it will take is something that will develop over time. I appreciate it is almost impossible to get the countries of the world to agree on ethical norms and on any kinds of laws. I appreciate that you could get countries into a certain agreement but you always have a rogue operator somewhere that violates the law. I think everybody wants to start thinking about it that way or we'll never come up with ways to regulate even the outliers.

The other thing is major Internet platforms are discovering that this business of creating and designing a technology for a particular society basing content regulation themes in the global north is not working. You can't recognize hate speech that takes place in some corner of India if you're in California, you have no idea what particular languages mean in other countries. To do that in retrospect, it is difficult. Again, even you're talking about value built in to design I think it is important to recognize that the value built in to design has to be thought of in terms of global impact and not just in terms of local impact.

Second thing, stepping away a bit from the cross-border impact but connected, it is that I think it is also time to start thinking about who is responsible for AI? Not necessarily in terms of the actual developer. Again, the formation of AI I think involves the regulation of nation states within which it is developed and the specifics of the researchers working on AI and eventually any global business as we all know is an assembly of many parts and so then the question is what standards do you apply when you're sourcing bits of software, hardware from somewhere else, what happens if your business is split across, you know, Syria, Iraq, the U.S., Poland, India. How do you split responsibility both between regulators and between people that have developed specific component parts and that market it. These are all questions that I think we need to start sort of mapping out, create sort of a framework of the different kinds of people that could be accountable and how so that they can all start preparing now early on for ethical development of AI and the last, which is actually pretty simple.

I'm a researcher. My background is in law. I'm based in India and trying to write a paper on Artificial Intelligence and liability, and I find that actually the writing on it, just on Artificial Intelligence and the law and liability, it is not extensive at the moment. Maybe a concrete, good place to start is to start pulling together the different solutions people are identifying. For example, there was a valuable point of if AI thinks differently, how do humans police AI? There is a paper written about how we need to develop AI to police AI that -- I think that Francesca brought it up, you need tools to have the processes. It is useful to start bringing together the different ideas that people have, the different information that people have so that we can also see the holes in our thinking about the framework and identify areas in which researchers and agencies can work.

>> MODERATOR: That's enlightening. I think as -- yeah. Can you develop an AI to police an AI? Who polices the police? It is interesting.

I mean, you know, this technology as we have been discussing is a category on to itself because of its tremendous power for both good and ill and I think you know we have international regulations as I mentioned yesterday around for example nuclear material, you know, you can't go on eBay and buy a chunk of uranium almost how you can buy any other element out there. There are reasons for that. We have seen examples of potential offed weapons for example. Now we're talking about the ability to print an atom bomb. How do you prevent that problem? How do you police that?

Does that mean that we have to move into a world where security takes priority over privacy because you can't stop access to that file and printing it in some sense? Anybody could potentially write a piece of software sitting on a bench in a park offline somewhere that has the power to cause great harm. It is a challenge.

You have spoken before about the impact in the developing world. Surveillance is another area as well. What are your concerns there in terms of the impact in the South?

Surveillance is the classic example. There is a sharing of technology but honestly I global South governments are not hiding what they're doing about those. I was happy to see Brazil had the response it did to the Snowden revelations and India says everybody is doing this, why is everybody objecting to us doing it to. It is again the design technology. Countries don't take responsibility for

how it is used and it goes to the global South and used partially by the state in a way that's very easy to understand states impacting citizens' rights and also in other ways. You are sort of linking AI with big data, that's valuable, a big telecommunication company in India, Reliance, they have actually opened up this new network it is creating with the tag line data is the new oil.

Data farming is becoming a huge thing particularly in countries with no privacy protection. I wouldn't j you say go to surveillance and the harms it creates but there is a nexus between business and government and the collection of data with no accountability, the big processes that are definitely issues. India has a controversial bio metric database and it is a state database that basically links banking information, ticket information, medical information, basically the works to an individual and a single number. It has caused great concern because a law enforcement agencies can access it whenever they want.

>> MODERATOR: Many thanks.

Let's go, if we could, to the audience. If there are questions we'll take a few questions.

>> AUDIENCE: Hello. Software engineer and student.

A topic that I heard all of the panelists touch on is the fact that we built deck knowledges and the technologies shape us in turn. These algorithms we're building have potential to harvest enormous amounts of data. A question yesterday, a first question was about targeted advertising and how it has shaped our -- how it has the ability to shape our beliefs and moral opinions from many things from privacy to political opinions, so forth. How do you feel as representatives of the constituency of the moral beliefs changing overtime, many of you mentioned, Francesca about the dog wagging the tails or the tail wagging the dog and the ideas and it is seeming like so easy to sort of hijack a lot of our beliefs just using targeted advertising and these types of rules.

I guess my question would be do you feel that our values now should be fixed points in the spaces whereas we're -- people used to be very concerned about privacy, now there may be more willing to sacrifice that in terms of convenience and security and what things aren't you willing to compromise on? It is very easy to take the optimist and pessimist and land in the middle and you dilute the ideas. Where do you land in terms of what things won't you compromise on and how can you give -- I think about regulation and policing, watchdog more teeth. Where do we stand concretely in to these ideas?

Thank you.

>> MODERATOR: Here first and then here.

>> I'm a system analyst.

I would like to ask about the perspectives or views regarding ethics with respect to future of super intelligence. It may be easy to combat the ethical values to systems that we're building and test driving and can debug so the Human Rights medical ethics and other values can be embedded and test driven even in the self-driving systems, right. It is not going to be so easy when we have self-improving evolving systems in the wild for which we have ingredients already today.

For example, it may be enough to create a computer program that seeks to optimize the ability to obtain computing resources and use it to mine digital currency and use that digital currency to hire people online and to improve parts of the code to create the reliance. Basically using the human resources. I think we should really be careful about that.

Eventually a system evolving in such a way would learn some values. In all likelihood the values would not be centric but machine-centric. How can we cooperate with that? Looks like it is inevitable that this will happen with a theorem and other digital concurrences and can we operate with them and come up with universal ethical values that would be shared between different forms of intelligent life? I think yes, we could.

If we try to approach this with an ethics and information theoretically treating life as information evolution process under influence of physical entropy always trying to raise us all, not just humans but machines as well. What is your views on implicit ethics that will follow from what life is as information process?

>> AUDIENCE: Thank you. I'm from the University and developer from Alfred, a machine learning start-up.

I have kind of two questions and they're sort of -- I'll throw some things up and see if there are reactions from the panel. Some things I have noticed as themes in this panel and yesterday are really I think what we're getting at when talking about how to regulate or come up with an ethical framework for AI is a question of intentionality. Within that, I sort of have identified two sub categories, one of which I call meticulous transparency so if you can't sort of -- not just know what the model is doing but have people tell you exactly, precisely what they're developing the technology for, how they'll use it and how the regulatory framework is for that.

The second idea: Contingency planning as part of your intentionality. If we're doing X we will have -- we have to document for Y, Z and Alpha. The fact is, this is what we have to do in academia and academia before we go through the research in medicine we have to go through the research and say we plan to do this, this, this and exactly this and we'll share the data with these people and these people and not with these people and we'll use it for X, Y, Z and if we want more we have to ask. It is very simple.

The second question, data management or rights, something that everybody has alluded to but we have not gotten to the meat of, how do you work on principles for data management and rights and do we have to make a decision, a priority in the areas, medicines, there is a model being developed that that information must by law be always anonymous without question and only with extended informed consent can you have a non-anonymous database in use. On your point, do we need an AEA for AI?

Thank you.

>> MODERATOR: Second row.

>> AUDIENCE: I'm Patrick Jones, an architect and interaction designer.

I love that you opened up with the utopian dichotomy assuming mass adoption of AI and I wonder what other Rights of users not to be implicated whether you're somebody who stands accused, a Philistine, Libertarian, whatever, if you're a prepper, just because something can doesn't mean that you should do I have a right to say don't implicate me or involve me and we're talking about the development of AI and trying to look at principles for that. That's one thing we need to talk about I think.

>> MODERATOR: We'll stop there and get reactions first from our panel and speakers. Lorna sadly had a plane to catch. She has to go and vote and that's a good thing. She's heading to the airport. For those that remain, we'll start here.

Luka, we have manipulation, intentionality, super intelligence.

>> LUKA OMLADIC: Those of you that have a philosophy training, you know that that's a classical question about -- the question, how to device universal ethics that would be valid for both non-human intelligence and human intelligence because in principle if you look at classical ethical ideas, even utilitarian, they're not limited to human subjects. This is a free rational thinking as a rational subject and for example, utilitarian, something, whatever that may be. In principle you could use all the

classical tools of ethics to devise universal imperatives for rational beings that we are here. That's not -- that's not -- at least at the theoretical level it is not such a big problem. The more immediate concern, we have machines for milinea but I think we never had -- how should I say -- decision machines and now we're facing with the world populated more and more populated by decision machines.

We delegate our work to machines or to the animals for millennia and now we're delegating our decisions to machines and that's in my view the biggest ethical question, one of the biggest ethical questions, is it possible to draw a line, is it possible to draw a line and say these types of decisions should not be delegated to decision machines. That's one of the huge questions in my opinion.

If I may briefly comment on the question, is it possible to step out of this work, I think it is not possible to step out. I think it is -- we're all in the same situation and even the preppers, even going off grid or eventually you will be effected, just like climate change you will be effected no matter where you go. It is probably not a universal decision to get disconnected. It is not possible to be disconnected. You live in this general intellect which is shaping itself and it is not possible to go off grid in my opinion at this time.

>> FRANCESCA ROSSI: I would like to target the two questions. One is targeting advertising and I think that of course there is a whole spectrum. Of course, you want advertising to be personalized because otherwise you're bombarded with things that you don't care about and so that end of the spectrum, it is not useful for anybody.

On the other hand, I mean, if you do too much to the other, towards the other end of the spectrum you get into this issue of being so personalized that you get manipulated and you see the world in a very non-realistic way. I think -- I don't think that we can universally say so this is where we want to be between the two ends of the spectrum for every tool, for every scenario, for every application. I think that one of the answers can be what was -- I like this term, the transparency, that, you know, you want AI systems and targeting advertising as well, you want it to be transparent as to what values, what optimization criteria, because values are also optimization criteria, are used in order to give you the suggestion of the advertising.

I envision that in the future every AI system has an ethical AI to be third-party certified and you can locate

inside and you can see exactly which values the system is following. If the system is telling you I'm doing this advertising and the objective criteria I have is to maximize engagement and clicks that's it! That's the system it was transparent and then you decide whether you want to use that and you maybe decide to use it anyway, even if the system is telling you I'm not -- I don't care about fairness, diversity, I don't care about what other values, and maybe you decide to use it anyway.

The issue of being transparent about what criteria, what values the system follows is important and essential to give you the freedom to decide, you know, what you want to use. Then maybe the market will -- given that everybody knows what is values are used and then people make their choices, maybe there will be some business value, maybe not in another win and the companies will deploy according to some values and not just maximizing the engagements for example because maybe there is no reason for value then, maybe no one wants to use them. Maybe, aim saying in the future. You know, that's an evolution to be seen. I don't think somebody could say, you know, this is the place where we are going to be by regulations for example. I don't think that's the way to solve the issue.

>> MODERATOR: Wendell, a comment?

>> WENDELL WALLACH: We have so many different things with just four questions. I'll respond to a few of them.

First of all, we're too caught up in accentuating the value differences when there is an awful lot of value consensus out there. Particularly in things like universal Human Rights that is central to this body here and the difficulty is that there are values all the way down, values are in everything, there is consensus on the values, not always consensus on the application of values or how they're prioritized when you have emotional issues that come before, particularly issues that threaten peoples' lives.

I'm not so concerned about getting in place foundational values. I have also think we also lose sight of what the functions of values are. When I say that we lose sight I'm not sure we ever had sight but the punch in the values is they're tools that we use to navigate an uncertain world. It is nothing more or less ultimately. We don't have all the information we need, some of the information we have is false and we can't calculate all of the consequences of our action. That's the reality of human existence.

Values give us some parameters on if you go too much in this direction you'll open up another can of worms and

you'll have a worse situation than what you started ouch with. We hope that system works for us and it has worked really actively well. How that is applied in other situations, it is complicated. Should we have this data in medical research? Yes. How effective is that? Not really if you really want to figure out who is anonymous, you can. We have all of these kinds of problems in research ethics.

We have the ownership of data, which is perhaps the biggest issue when we talk about AI. Those who own data have oil. They are more powerful in this coming age than the oil companies ever were. The AP will be the AI-opoly is there a way to opt out of that? I think that's something that we have to give due consideration to. Unfortunately in a capitalist world that means all power flows upward. It flows to those that own stocks which may be everybody in this room on, you know. In that fact, we're all within the privileged part of society and obviously some people own a hell of a lot more stock than other people do. Those are implicit values that are structured into the very environment that we're moving these technologies in and it is very important that we recognize the power structure of those and recognize whether that will serve humanity as a whole as we go down the line.

Finally, super intelligence, I'll say one, two things. With the jump too quickly to -- that there will be in super intelligence, what do we view it with and I think well, there's a lot of bridges that you have to cross between here and there. If we can't figure out a way to create AI systems that are value aligned, have sensitivity to human values, demonstrate a degree of wisdom in choices and actions, then we haven't gotten to super intelligence. Super intelligence is not just going to appear out of the woods. We may have the illusion that we can create certain kinds of comp terrible systems that manipulate information in far superior ways to humans. Unless there is a solving of the value problem they will always have the Achilles heels and that's not fully recognized. If they don't have them it is just because we're too stupid that we have aggregated powers to them that they never deserved in the if user place.

>> MODERATOR: A brief reaction here, we're close to the end.

>> CHINMAYI ARUN: My big reaction is this is a great panel. You know, sometimes it is too late to ask the intention question after an entire city is destroyed depending on the nature of what we're talking about.

That's not the place where you ask what went wrong, how can we do this better. There are some parts that we have to fix early on.

Intention is linked to reasonable foreseeability and other questions like that. When the propulsion of harm that you're risking is substantial sometimes there is not time to look at the differences and there is concepts like strict liability, very controversial that's been created in areas like that.

I want to point out that AI is the new generation of technology and the old generation of technology which is online platforms, this idea, it is doing its way to -- they have all of the manipulation and the fake news and it is too late to get into the design right now. For all of the things we were advocating, free markets, free speech -- this includes me by the way -- not into questions of who you does the algorithm work, how do we look in the black box and we have a lot of destruction and no real option to opt out for the companies saying this is a product, don't use it if you don't like it, it is a part of our world. Maybe there is a lesson in there that we can use when we're getting into the design of AI.

>> MODERATOR: Very good. Thank you.

Are we close to a consensus? We have -- there are more questions from the audience. If you will come up afterward and talk to me, I can give you my contact information, love to know any thoughts that you have on recommendations and what -- recommendations and what should be in there, we're sadly close to the end of the session. I'm sure that the points you have to make are very valuable and I would love to hear from you and we're compiling recommendations to take forward to this group.

>> LUKA OMLADIC: I listened to what you said, to come up with recommendations, it is an impossible task you I like the ideas that come out in the theme, in the theme of where are we with the transparencies and the contingency plan that we had talked about.

From my background as an engineer and when we look at the systems to automate, to control nuclear reap actors, that could have happened, it is very simple. We do monetary, we do announcing and controls. We need some of that in general in place and we could agree on what kind of things we want to monitor, put in the design for this so that we can monitor what the activity is, from there we can decide what type of the window you want to initiate. It is what degrees are you worried about will happen. The third is where is the kill switch that you need to kill it? To

me, I think I like to propose that we think along the lines of -- this is a common thing, it is a good engineering approach, we could write something up together and I would love to get your input between now and tomorrow on what you think and if you want to get involved, if you want to hip us shape it up I think Robert is the right person to talk to and I'll help Robert put that together.

>> MODERATOR: Thank you, all, for joining the conversation. Thank you to our guests here, the speakers.

>> It is important to include Civil Society in this process. It hasn't been stressed enough, this is more important than experts speaking. It is the importance of Civil Society speaking, it is the only way to overcome this idea that the free market will form our values but not the values that society wants I think.

>> MODERATOR: We're talking about trying to come up with the world in which AI is used as a tool for empowerment, that's got to start by giving everybody a voice and deciding what it should and should not be able to do. That's the first part of empowerment.

Thank you all for being here. Thank you very much to our speakers and panelists.

This is just the beginning of the conversation. Let's keep it going immediately after this. We look forward to hearing more feedback and thoughts from you in the days and weeks ahead.

Thank you so much.

***

***