AI and You

Transcript

Guest: Cansu Canca, part 2

Episode 112

First Aired: Monday, August 8, 2022

Hello, and welcome to episode 112! Last week, I started talking with Cansu Canca, who was calling in from Turkey. She is a philosopher and AI ethics consultant, and founder of the AI Ethics Lab, prior to which she was on the full-time faculty at the University of Hong Kong, and an ethics researcher at the Harvard Law School, and many other organizations. She has given over 100 talks on AI ethics, including her TEDx talk *How to Solve AI's Ethical Puzzles*, and she was listed as one of the "100 Brilliant Women in AI Ethics."

Last week we talked about her journey coming from the field of medical ethics into AI ethics, what the experience of a company working with the AI Ethics Lab is, really getting an idea of what it's like on the front lines of making ethics work in the use of AI right now. Without further ado, let's get back to the interview with Cansu Canca.

So, you mentioned a couple of key words, there," control" and "privacy" and the impact of AI system, perhaps on how they influence people's thinking and I know you've done some work with search engines and their impact on ethics. Could you describe what you've done in that respect?

Sure. I mean, I think the something that is very interesting is our life, our access to information, our access, a lot of the questions that we're asking now is through some sort of search engine, which obviously wasn't the case before the search engine. So, it's don't just think about the typical website: Bing, and Google and Duck Duck Go, or whatever is your favorite. But also think about the vertical ones, like you go to, in order to find the best hotel or the best plane, you're probably using some sort of vertical search engine. Or if you go to YouTube or Facebook, you're using another search to search within their system. So, the search engines are really like everywhere, too, and they are great, of course, because without them, there will be just like jungle of information out there we never can sort through and get anything out of. So, we definitely need them. We know that we need them, we all like them - or should like them, if not. But, the way that they show information, the way that they sort this world around us, is extremely relevant to how we make decisions. So, any question that you're asking, right, if this is about, should I get a vaccination? This question will go through some sort of search engine, either social media, which is going to show you different answers, or platform like YouTube, you want to watch videos on this question, or you're going to ask us to ask this in our search engines or something like Google. And imagine the different options, right? Like, first of all, there's this engine might just show you information that are just plain out false, wrong. Which is a problem; then you're going to make decisions on those on that information, especially if you don't know what you don't know. If you don't know that there is like vaccinations, that is an easy topic in a way. I mean, it's horribly complicated, but easy in the way that you have some idea of who are the authorities, like you could figure out who to look for. But many questions, we don't even know who is the authority, like who should be asked these questions. So, the way that the information is served to you, or the nudges that are provided, as you're searching, so the autocompletes that are provided as you're searching, for example; all of these impact your decision making and the decision making, basically that's just saying that we are free individuals, we have control on our lives, on our bodies, implies that we make our own decisions. And if we are somehow manipulated in this process, it is as bad as being manipulated by force. So, we are really talking about a major issue there. And I think we came to really realize this, with the vaccination with conversations. I mean, way before COVID this was a big topic about measles vaccination, for example, but also the big event of the elections, both in the UK and the US and going back to my personal history narrative of being in Hong Kong and then founding AI Ethics Lab. AI Ethics Lab really worked out, it succeeded, because Cambridge Analytica and Facebook scandals happened, because that was the time that people said, we want to talk about this, ow, who are we going to look for? And that's when AI Ethics Lab became relevant, otherwise, I could have been doing this on the side forever and no one would know.

And that opens up so many avenues of exploration, because the search engines basically answer the question of what's the web page that you're most likely to be looking for? And the recommender algorithms show what conversation are you most likely to be interested in spending your time on. And now these things are starting to converge, and that we're likely to get search engines that are conversational, that involve chat bots, and something like LaMDA, for instance, and then the question is going to be well, what's the bias in that conversation like? Because there's so many more dimensions it could occupy than just "here's a web page." And the recommender algorithms of places like Tiktok and YouTube have been dissected and found that they basically radicalize the audience, take them in the extreme of whatever their tendencies are, because that's what will get them to spend the most time on it and that's influenced those elections, as you were talking about, and the measles vaccination and other conversations that we're only painfully familiar with. And even after this came out, that this breakage of democracy happened because of these algorithms, I haven't seen a whole lot of change - I haven't seen any visible change - in how these things happen. It still seems to be going along. Is this a conversation that's happening behind the scenes anywhere that I'm just not privy to, how these companies might be trying to fix their ethical problems?

Definitely, there's a conversation around this, a huge conversation around this, behind the scenes, but sometimes also, you know, catches the public attention. But this is a major question and everybody's aware of it and I would say actually, to my knowledge at least, most of the companies that I know who are the relevant parties to this discussion, are trying to find some way forward, some solution. I don't think the problem here is they're ignoring it. It *used* to be that, but then, it became no, this is not a problem that they are ignoring anymore, because the implications are so clear. You know, when an ethicist says here, this *might* be an implication, it's one thing. When you actually realize that the world has just taken a different course, well, that has an impact. But it is a tough question. I mean, it is a very, very difficult question. The way that I started working on this one was, we wanted to use this actually as a use case when we were designing a workshop. So, we wanted to design this interactive hands-on workshop for non-philosophers to really make them walk through an ethical decision making and I mean, walk through. So we really designed a floor game, where they physically have to like, as they make decisions, they move, like physically move in the room and there are a lot of opportunities to discuss how to solve the question, but so I'm really trying to make them dissect the question, understand the ethical push and pulls within the question, but also start connecting it to their own expertise on well, how can I technically solve this question using what I know? So, we use this as a use case, and we worked really - like anytime you create - this particular workshop, what we call the mapping is one of those that require a lot of like pre-work because we, the ones who are leading the workshop, we have to know all possible scenarios so we can keep leading the group. So, we dissected this question, really, in all pieces. It's a hard one, we don't have an answer. After all, we just don't have the answer to say "Here is the right thing to do." It's extremely difficult, we can say some of the things that are completely wrong. And a lot of the times this solution is really difficult. You know, we understand that all of us understand that we don't want fake news. Whatever your fake news radar is catching, everyone agrees we don't want fake news. We might be calling each other's news fake news, but we are still agreeing that we don't want fake news. But it's so difficult to actually say, okay, here's how we are going to implement a technological structure without causing complete violation of freedom of speech, and get rid of fake news. It's just incredibly difficult. So, it is good that we are talking about it, it is good that we are pushing the companies, we are pushing philosophers, to come up with solutions and not just think forever. But the truth is these are questions that are not going to have easy answers, we are going to go with trial and error and iteration of getting better. But that's like saying with the search engine, same with the fake news, it's just incredibly difficult questions to solve? Because there are trade-offs, right? That's the thing. Because the trade-off is, as you said, when I am on YouTube, or Facebook, I don't want the platform to completely take advantage of my interest in cat videos. But I want to see cat videos too, I don't want them to educate. I chose YouTube over Coursera at that moment for a reason. I'm not looking for education. So, it is a trade-off. It is a hard balance and it's just really difficult to figure out which way to do this. But there are ways of getting better, at least.

Because it has so few inputs. All it knows is what you type in the box - and your history - a huge number of things it knows psychometrically about you, but other than what you want right now all it knows is what you type in the box. I'd like to visit the Asilomar Principles briefly because I've been fascinated by that for a while. In 2017, a few dozen AI researchers, philosophers, ethicists, got together in Asilomar, California and hammered out what I believe was the result of a lot of heated discussions: 23 principles for the ethical development of AI going forward, which read really well. And then apparently nothing happened with them. No one came out with saying, we have a stamp on our company that says Asilomar-compliant or anything. Just from your point of view, did that do any good? Did it do anything?

Empirically I don't know what will be the counterfactual if a certain event never happened. But I mean, I completely see your point. I mean, it's really hard to say that, as a result, now we have this great outcome, but I don't I don't see that either. I think so, principles are interesting topic. I did a project on this and then I kind of got stuck with this question with this, this principles approach, because they are, on the one hand principles are super useful, because they are good, rules of thumb. On the other hand, they conflict. They always conflict. So, I think that's one of the reasons that it's very difficult to say that now we abide by Asilomar principles, because what does it really mean to say that, let's say like looking at the principles right now, I think that you know, you have the safety and transparency and responsibility, but they will conflict and complex cases right. So, you cannot be there will be cases where safety, increasing the safety will result in decreasing the transparency, or many such trade-offs. So, it cannot be really such that we rely on principles, but we can take the principles as a starting point. So, I'm just going to go very briefly to a little bit of history of the principles, because I think this is relevant actually. Again, I connect AI ethics to medical ethics so much because medical ethics is a very well-established applied ethics field and there are so many ways that we can learn from medical ethics, but also we can learn from the mistakes of medical ethics. So, the principles approach was actually again in 1970s, came out from the big debate about research ethics, experimenting on humans, how can we do the medical experimentation in a way that is not horrible and inhumane? Because until the '70s, the cases are just fascinatingly horrible. So, what happens is that in the US they put together a working group; government etc., working Group and two philosophers who are engaging with this working group also on the side, write a book which then becomes a bit like a basically a very important book of the field of bioethics, the principles of biomedical ethics. And there they have principles, which gives us the respect of autonomy, respect for autonomy, non-maleficence, so, do no harm, and beneficence; do good. So these are the principles. And then they come up with like some applications; like informed consent is an application of autonomy, because by making sure that we get informed consent, we are allowing individuals to exercise their autonomy, it can make decisions about whether or not they are going to join this study and so on. But while I was in medical ethics, I was very annoyed by the principalism, because there is this approach among physicians that if they know principles, they're going to solve everything. And unfortunately, that never works, because any complex case has conflicts. Because these three principles, these three core values, capture the whole of moral and political philosophy, like literally the whole thing. And theories conflict. Kantian ethics does not agree with utilitarianism, Roseann theory of justice does not agree with egalitarianism. So, there are conflicts within our theories and the principalism just basically brings them all together and says "Here you go." But what I think Asilomar principles are trying to do - and then many other principles came out - for AI ethics principles came out is a starting point. If you know that you should think about safety, you are going to ask some questions that are relevant. And if you think that you should think about transparency, you're going to ask some questions that are relevant. And when they conflict, well then, leave the principles, go do the ethics work. And I think for that reason, in a very long answer to your question, there is no stamp, because ethics is a very difficult thing to give a stamp on. The process, I think, is what we should really focus on - with the help of these principles, of course - but the process of, "Have you checked these questions like, have you checked the transparency concerns? Have you checked the safety concerns? Have you taken mitigation measures? Did you put in place safeguards, if you recognize that there's a privacy issue, there's a transparency issue?" So, looking at the process, I think, is what the principles should need us do. And did they? So now we have easily over 100 sets of principles on AI ethics. Actually there's a page called "Dynamics of AI Principles" you can we have an interactive page, you can play around with the principles, see how they developed over time, on demand, and all of those things; compare the summaries. So, all of these principles, did they help? I think, in a way, yes, because at least they gave the non-philosophers some keywords. When we are talking about the bias, or the algorithmic bias, or the identified data and those things, I think those are useful keywords to keep in mind, but unfortunately, they take us only so far.

Very interesting. The Asilomar principles were, I think, more forward-looking than perhaps today's things that they didn't get into the sort of nitty-gritty detail of issues of bias and privacy that are being negotiated right now. So one thing I want to talk about, because everyone else is talking about it, is ethical questions raised by Blake Lemoine's assertion that LaMDA AI had become sentient. And I don't know that anyone aside from Blake Lemoine believes that, but let's use it as a springboard anyway. It reminds me of a hypothetical case that was raised some years ago by Martine Rothblatt, who litigated a mock trial of what looks like exactly the same case - an AI created by the hypothetical Exabit corporation that announced that it had become sentient and didn't win want to be turned off. And Exabit said, you belong to us and we can do that if we want and it went to court. Again: mock trial; didn't really happen, all the facts were imaginary, but the case was litigated and judged as though it were real, and she won. At some point in the future, that will become an issue where there's more debate than just one person claiming that some AI is sentient and everyone else saying you're wrong. And what sort of ethical guidelines or what sort of rails do we need in a company to help with - actually, I've got to mention this. This also reminds me of: I was at JPL - Jet Propulsion Lab - years ago, when the question of cold fusion came out, when it looked like researchers had discovered a way of making fusion happen with I think it was palladium, or platinum, metals, and it looked like it would be really easy to create a fusion reaction. And this memo went out to everyone at JPL saying - because it's a research lab - saying, if you're going to do this, then here are the guidelines, and if you start to observe this level of neutron activity, turn it off and call plant protection. And I sort of imagined a parallel sometime in the future, somewhere like Google that says, If you're creating conscious, artificial intelligence with free will, if it exhibits these capabilities, contact the ethics department. If you can project yourself forward to an era where that's plausible, what would you tell the ethics department to do?

Run away! It is extremely difficult question, extremely difficult, I mean, and I think there are multiple hard questions packed into this. So, the most obvious one of course, is how can you tell when something has a free will or something has consciousness, right? The consciousness is extremely difficult to define; even philosophically, consciousness is extremely difficult to define in Philosophy of Mind. So, we are already struggling with the concept, what makes us the way that we know that we are, all the thought procedures and so on? And then the next question is, of course - put aside consciousness but - what gives something a moral status? So, is it being sentient? What is the what is the thing that we are looking at? Is it just free will? Like, do we care only when the thing has (the "thing") has free will? Or do we care as soon as it says it hurts? You know, is this enough? So, the first question is how do we recognize how can we say that we are not being fooled by some word generator, versus something that is as authentic as the sentient beings that we know? And I'm saying this, purposely, so wait, because we don't know what is so authentic about us. You know, yes, we have not assembled ourselves or the other animals, but also don't know, so what is it that makes us special, or that makes us worthy of moral behavior, like we should be treated morally, why? Or what animals should be treated morally – why? And also going towards animals, how far do we go in treating animals morally? I mean, we know that we are failing on that front miserably. So, the second question there is then, once we recognize, what are the incentives around treating that thing morally, humanely, or just like ignoring that? So, that's one set of questions, right? But once, let's say we realize, okay, we have this thing has what it takes to be a moral agent or moral subject, actually, in this case. So, moral agent. Let's not go into this detail. But the second question will be, what is it to treat that thing morally because we know when I engage with a human, I know from my experience, what it means to hurt you, or violate your autonomy, or those type of things. I know the harms, I understand that from my own behavior. But it might not be the same, the obviously the pain might not be the same type of pain that we are thinking about; that that might not be relevant if you are a being that is connected to all the other AI systems around the world, maybe you don't even die when somebody turns you off here. What is that for them? What are the things that we consider bad? Are they bad for them? And what are things that we don't even understand? Maybe that's horrible for them. So, how do we even understand this scale of harm and moral violation in relation to AI systems? And I think the final thing to keep in mind is that if the AI becomes such that it is like us in a way, it has moral agency and morals, it is both able to act morally, and it should be subject to moral treatment, but it is different from us. Where does that put us in relation to the AI system? How should the AI treat us, now that is not just doing what we are telling it to do, but it is exhibiting that free will towards us? Should the AI treat us like we treat the animals? Is this something that we are comfortable with? Should the AI treat us like we treat our family? Or the question of this, we don't engage with non-human beings in the same way that we engage with human beings and we are not expecting the same behavior from non-human beings as we expect from humans. So, what does it mean in terms of our moral treatment? Yeah, so I guess my answer to the ethics team is that oh, my God, they are in big trouble.

Ask for more money. Well, and this centers on Google. And since so much attention is being given to them, I've got to believe that the sort of conversation we've just had has been repeated a thousand times there in the last few weeks, probably mostly informally. But since Google is also Ground Zero for a lot of this conversation, they had an encounter with their own ethical judgement when it came to the Department of Defense and the Maven contract. And their employees apparently revolted on that and said, we don't want you to do this; and Google apparently capitulated. And I wonder, what does that incident say about the power of employees to change the direction of a company through an ethical stand?

Yeah. that case is a good case to, again, show multiple aspects of AI ethics in action in a way. So, understanding that as being a part of those who are developing the AI system, or using the AI system, or being subject to the AI system, like, as the consumer as well as the population, we don't have to just take it as it comes. But we can raise our voice, we can make our opinions heard. If we systematically try to do this, if we don't just complain ourselves, but we take action and we systematically take some position on this, that is a great thing to know. Because I think in a way that this debate has to keep happening. It cannot be that whoever has the most money or whoever has the most resources, just continues to do as they like. It cannot be a wishful way of saying this. I hope it is not, it won't be the way. On the other hand, so you're right after the Maven event. So, this Google Maven project has an importance in AI ethics, because that is when Google developed, and after that grant, Google developed its principles, AI Principles, and published them. And that sort of inspired or incentivize others to do so. So, the conversations in a way got more multiplied, it became more available, it helped raising awareness in the industry as well in a way. On the other hand, the not so satisfying aspect of this case was that in the end the principles that they came up with did not really preclude a case like Maven. Because particularly in the case of Maven, it was such that like, if you, the Google principles basically managed to say that, no, we cannot develop the AI system that you want for the purposes that you want, which is to use it, to help to use it in order to help the military goals. But we technically can still develop it so that you can do say, search and rescue. The same technology their principles didn't stop them from doing that, they didn't go ahead. But the principles if the next Maven case comes the principles are not actually stopping the stopping that engagement, it just has to state a different objective. And I think that was sort of the problem with this walk out, and that way of debating the principles or debating AI ethics gets you only so far, which is important. So, I still think that was like a worthwhile move. But it has to continue in a way that is not outraged, that actually goes into the details of, how far do we want to go? I mean, it is a valid question to ask the Google employees as well. For example, what do you think about AI systems that are going to help search and rescue? How do you think we should draw the line where we know that technology is often dual use? All of these are questions that maybe after the initial reaction that says, hold on, don't go further. But then you have to sit down and make these like nitty gritty discussions. But like, don't go there. But like, exactly go where how, what is our position on this? What do you think is the right position? And really sort of laying it out, rather than leaving it vague. What Google did in the end was leaving it quite vague. So, did the employees win in a way in that case, but not really, in the bigger picture, I would say.

Is the problem that the more we develop AI, the more, or the easier it is, to make it dual-use? It's like educating a person who goes to a job fair, and they stop at one desk, and it's recruiting for a hospital and then the next one over is recruiting for the army, and that person could do either job.

I would say, I think this is a technical prediction that I don't know the answer, because you can also imagine that the more sophisticated AI systems become, their capability for dual use increases, but maybe the capability to design for certain goals and block certain uses could also become higher. So I don't know, the technical prediction part. I actually don't know. Logically, I think either is possible. But I don't know, what is the technical projection on?

Right? Oh, that's interesting, because it suggests a sort of - and this is really speculative, out there - but developing AI, with moral foundation built into it such that it would object to being used for certain purposes. We obviously don't know how to do that at the moment.

I think, yes, but I think a different way of saying it could be like developing AI with clear value judgments built into it, such that there's an internal stopping point for certain applications. So, I think, I'm talking about the more straightforward, the AI that we are more familiar with. Of course, as it becomes I mean, if we think of the AI becoming more like general intelligence, these type of blockages I'm guessing it's going to be harder to keep, so that when it you say becomes more relevant, like a moral foundation, just like you're raising a human, but maybe there's like a more simpler version of doing this by just like, can we have sort of like value judgments, as I say, built into the decision making that it doesn't go to certain places. I don't know.

Fascinating. Conversation could go on for so much longer than we have, I really appreciate it. What would you like to tell people about how to find out more about you, what you're doing how to follow you ,and get in touch?

So, they can go to AI Ethics Lab AIethicslab.com. So, that is where they will find all the information about what we are doing in AI ethics, the dynamics of AI principles, if they are curious about the principles, the puzzle solving, and ethics model, so that's where they can find all of that information. They can also look at the Institute for Experiential AI, Northeastern, and that is AI.Northeastern.Edu and that's where as I said, where I'm being responsible AI, work and research. So, both of them are my babies in a way that I'm very excited to work with, and to connect personally with me, they can find me on Twitter: @CCansu and more contact all my contact information is available on my page on https://aiethicslab.com/cansu.canca.

And they can find your TEDx talk as well.

"How to Solve AI's Puzzles."

Wonderful. Cansu Canca, thank you very much for coming on the show.

Thank you so much for having me.

That's the end of the interview. Really fascinating to hear how this work is conducted. As you can tell, I don't mind asking my guests the hard questions!

In today's news ripped from the headlines about AI, Britain's Department for Transport announced proposed changes to the Highway Code to allow users of self-driving cars to watch TV and movies behind the wheel. Even though – and this is the part I don't get – it will still be illegal for them to use a mobile phone from the driver's seat, they will be able to watch TV and movies if they are on built-in screens and the drivers are prepared to take back control of the cars if necessary. The changes, which Trudy Harrison, the transport minister, called a "major milestone in our safe introduction of self-driving vehicles," also say that the drivers will not be held responsible for a crash in this mode, and that insurance companies rather than individuals will be liable for claims. Now, there are as yet no vehicles or circumstances under which these proposed updates would be applicable as yet in Britain, but obviously they're thinking ahead. This really has to be interpreted narrowly before it becomes useful and not terrifying, though. In April 2021, the Department for Transport announced that it would allow handsfree driving in vehicles with lane-keeping technology on congested motorways, which is clearly an application where self-driving vehicles can have an edge over human drivers, because the environment is highly constrained, easy to train for, and is very boring for human drivers who are likely to make mistakes through being tired or distracted. Whereas if you can just set the AV to stay in the lane on the M25 while you take a nap and wake up half an hour later when you've moved four miles, that's a useful piece of automation that can be deployed right now.

Next week, my guest will be Justin Harrison, founder and CEO of YOV, who is using AI to make a conversational copy of his mother, and aims to do the same for other people's loved ones. Find out why and how he's doing that, next week, on *AI and You*.

Until then, remember: no matter how much computers learn how to do, it's how we come together as *humans* that matters.

http://aiandyou.net