AI and You

Transcript

Guest: Kush Varshney

Episode 82

First Aired: Monday, January 10, 2022

Hello, and welcome to episode 82! So, the popular narrative is full of questions about whether we can trust AI, propelled by *Terminator* comparisons and fears of Skynet taking over the world. We're not going there in this episode; that's a pretty incendiary space to occupy in any case and when we do talk about that, there's a lot of context to set and a lot of debunking to do. But how to trust AI is an issue today, right now, in a very real sense that computer scientists are engaged with; and we've got one to talk with on this episode.

Kush Varshney has a PhD from MIT and is a distinguished researcher IBM at IBM's Thomas J. Watson Research Center in New York, where he leads the machine learning group in the Foundations of Trustworthy AI department. He is the founding co-director of the IBM Science for Social Good initiative. He applies data science and predictive analytics to human capital management, healthcare, olfaction, computational creativity, public affairs, international development, and algorithmic fairness, which has led to awards for his research and papers. He is currently writing and self-publishing a book titled *Trustworthy Machine Learning*.

In this interview, we talk about really the whole ecosystem of trustworthiness, finding out where it goes in areas like privacy, anonymization, regulation, compliance, and oversight. Here we go.

Kush, welcome to the show.

It's my pleasure, Peter, thank you for having me.

So, tell me, why is trustworthiness of AI important to you?

To me, so I think the reason it's important for me is because it is starting to be used in a lot of things that I personally interact with. So last year, I bought a house, and I'm sure the mortgage lender must have used some machine learning as part of their process. And, when we're looking at candidates for hiring, and all sorts of things that are not part of our daily lives, it's there. So we really need those machine learning algorithms that are part of these processes to have all of the characteristics that we want out of other people that are trustworthy so that machines are as well.

And as you say there, you're talking about it being used in things like lending, and it's creeping into so many aspects of our lives. And is there enough oversight around it? Because it's clearly caused a concern on your part about the trustworthiness of it. Or how are we assuring the trustworthiness enough? And I'm wondering whether we are setting ourselves up for some kind of pushback due to a lack of regulation. What are your thoughts?

Yeah, so as we talked about just now, it's being used in all sorts of different application areas and different walks of life. And many of those areas already have regulations in place, so

whether it's housing or employment, or others. And the fact that machine learning is being used in them doesn't absolve the responsibility that's already there by those providers. So a lot of the laws that are already there, they've been there for the last 40 or 50 years. But what is new is, yes, that we have AI and machine learning involved. And I think, given that AI is a general-purpose technology, just like other things like fire, the wheel, things like that, I mean, you don't regulate the wheel, you regulate the use of the wheel, like on a car or windmill or other places. So I think the same thing is going to happen here. And when the laws are applicable, that's great, and we should seriously be looking into the use of AI there. When there's less risky, and less consequential application areas, I think it's okay to go without a really high level of regulation. But things are happening. I mean, the European Commission came out with the draft set of regulations for high-risk AI in April of this year, so things are headed that way for sure.

And if we compare this, say, to the airplane industry, which is highly regulated because millions of people travel by plane every year, and when I get on a plane, as a result of that regulation, I don't worry about an engine falling off. And in the rare occasions where an engine actually does fall off a plane, the plane is generally still safe as a result of the engineering that's gone into it. But it seems that AI, at the moment, is kind of near the Wright Brothers stage where there's a lot of barnstorming going on and the rest of us feel like we're passengers at a county fair, going for a ride in a biplane, and you get in and they go, "Oh, here's your parachute, just in case you need this." I don't know if that's fair at the moment, but I do know, we don't have anything that looks like an NTSB or an FAA for artificial intelligence. Do you see us heading in that direction? Do you think that would be a good idea?

Yeah, I really like the analogy that you brought up. So if we look at the Wright brothers, so their first flight was around 1903. And the first 50 years of aviation, up until about 1958, were all about just figuring out how to make planes fly. And in 1958, that's when the Boeing 707 came out. And since then, commercial jets haven't really changed very much - they fly at the same speed, have all of the same capabilities, and so forth. But what has changed since then is the safety aspects. So if we compare the safety of aircrafts today, compared to the 1970s, for example, the fatality rate per miles flown is something like 300 times lower. And so we can kind of think of aviation [like the] first 50 years, just understanding how to get it to work, second 50 years, which is now kind of at its end, working on the safety, efficiency, things like that. So, AI is kind of 50 years behind aviation in that sense. So we can say the 1956 Dartmouth conference is when AI kind of got started. So the first 50 years maybe ended around 2012 or so when deep learning won the ImageNet competition. And so again, first 50 years, just getting things to work, and now, we're right at the beginning of the second 50 years. And so, yes, things are not in the state where we could be, whatever, 40, 50 years down the line, but I think we're headed in that way. There's lots of researchers who have now started looking into safety and fairness, trustworthiness, and so forth. So to me, this is the right progression that we're on. And towards your point around the NTSB and so forth, yeah, when an engine does fall off, it gets recorded in these flight recorders and so forth, so-called black boxes. So I think we're headed in that direction as well. So one of our projects that one of your previous guests, Mike Hind, talked about was AI factsheets, and it's a way to instrument an AI development lifecycle to capture a

bunch of facts as they're being included. So lots of tests results, not just for accuracy, but fairness, robustness, explainability, etc. And they're all made available, and so forth. So, again, we're headed that way; I think we'll get there pretty soon.

I just want to draw a distinction for our listeners here in the kind of trustworthiness we're talking about because some of the earlier conversations that we've had on this podcast have been about existential risk, and we've had people on the show talking about how we might be able to make AI safer for developing general artificial intelligence so that it doesn't get away from us. People like Roman Yampolskiy have explored that space. And we're not in that area here, this is about the AI that's being used right now for everyday applications. And you have a book on this, *Trustworthy Machine Learning*. What are your goals for that book? Who do you want to read that?

Yeah, no, thanks for making the distinction with the AGI sort of stuff because, yeah, my focus, certainly right now is on those applications that are in the very present sort of state of things. So yeah, the book is talking about all these topics that I include in trustworthy AI. So things like fairness, robustness, explainability, uncertainty, quantification, transparency, value alignment, and so forth, as well as the inclusivity that we need in developing those systems. And it's a book that's kind of intended for practitioners, data scientists, and the like. And what I want to get across is mostly kind of the thought process, how to kind of think through things, slow down, don't take shortcuts as you're developing these machine learning systems, and kind of what are the things that you should be considering so that you're not singularly focused on accuracy as the only performance criteria that you should be worried about. So yeah, it's kind of using examples in each chapter, from a practical sort of use case that I've personally encountered, and kind of going through a different topic each time of what are the considerations, how to think about it, and so forth.

You mentioned some other books in yours that have catalogued failures in certain industries, notably Upton Sinclair's 1906 book, *The Jungle*, about the meat industry. And I wonder, do you see a moment of reckoning like that approaching with AI? Is that why you bring that up?

Yes. So I kind of in the preface, bring up Upton Sinclair and the processed food industry as a sort of example of another industry that was new at its time and had activists and investigators looking into the bad practices that might be there. And this muckraking tradition, I think, is very well and active right now in AI. So, as an example, just a couple of years ago, Joy Buolamwini and Timnit Gebru, they came out with this "Gender Shades" study, which showed how several commercial face attribute classification algorithms were providing differential performance across groups. And there's another book, Cathy O'Neil's book on *Weapons of Math Destruction*, which also brings to light several issues. And to me, both of them are definitely in this muckraking sort of tradition; Cathy O'Neil herself actually says it. And so, yeah, we're here. I think there's a lot of activists out there who are kind of showing all these issues, and I kind of bring a different perspective. So again, in that preface of the book, I can contrast Upton Sinclair with Henry Heinz, who started a large processed food company, which is obviously still active

today. And the Heinz Company did a lot from the inside, trying to do things in the right way. And I think that's the way we should be progressing as much as possible as AI developers.

And your book has got a lot of highly technical content about how to solve this, including bleeding-edge concepts like differential privacy. And so who should be reading this to generate the maximum benefit from it? Would it be data scientists at large corporations, researchers, or data ethicists? Who would that be?

Yeah. So I do have some technical topics, for sure. So there's even some things around detection theory, which kind of set forth what are the kind of baseline, the best accuracy you could achieve for machine learning, and so forth, and going into fairness, metrics, explainability, algorithms, differential privacy, and so forth. But it's not written in a way that is meant for extreme depth. It's more for highlighting the ideas so that people can follow up in greater depth if they want to. But it's really meant for data scientists who are practicing right now in industries that are either regulated or have high-stakes considerations. And so the data scientists or the project managers, product managers who they work with, mainly to get the thought process. "What are the things I need to worry about? How do I worry about them and give them hooks into and what they can look into in greater detail afterwards?"

And speaking about, as the book does, bias and privacy among other issues, those surfaced in a paper recently that was done by radiographers who were researching artificial intelligence interpretation of X-rays and found that machine learning could infer the race to a high degree of accuracy from X-rays, which is a problem for de-identifying or anonymizing datasets because they found that they could add no amount of noise to those images that would stop that categorization from happening. Are you familiar with that work?

I'm not familiar with it specifically but the way you've described it makes a lot of sense to me that that kind of study could be done.

Do any of the anonymization techniques that you're aware of hold a possibility of being able to anonymize data like that, whether or not you already know what they found out, that it correlates with race?

So the first thing I would say is that race itself is a problematic sort of thing because it's more of a social construction than being actually a physiological sort of construct. So yeah, of course, there are differences across populations, you can look at people and tell that they're not all the same. But if there is work that is able to do that, it's possible that it's great work. And it's also possible that it's picking up on some spurious sort of correlations. There's been other sort of related work that actually, if you look at it in detail, there's all these weird papers predicting criminality or sexual orientation or other things from images that have been debunked later because there were actually spurious correlations or confounding factors that led to the results. But it's certainly possible in this case that there could be something there. So in terms of anonymization techniques, the main thing I would say is, there's also institutional controls. Data doesn't always have to be exposed. So there's ways to encrypt data to keep it under lock and key,

introduce many sorts of other controls. And you can avoid a lot of problems with privacy if you just maintain various sort of security measures as well.

I'm thinking about the brittleness of a number of classification learning algorithms, particularly in image recognition, notorious cases of things like AI has got very good at recognizing wolves from dogs, but then it turned out that they were doing that because there was snow in the background of the wolves; ones that were recognizing sheep, but then only because there was landscape in the background; ones that were recognizing cancers, but then they were keying off the presence of a ruler in the image that was only in the cancers, and so forth. And these are all scary. They're funny when we find out about them. And they're funny because we realize that the AI is not thinking the way they that we are. But then what about the ones that we haven't found that must be happening? And because of this brittleness, is there anything that is being done to make this less brittle, to fail softer?

Yeah, no, that's a great way to put it, "failing softer." So there's actually plenty of work looking at exactly those sorts of examples. And basically, the thing that we want to achieve is to have these machine learning models focus on causal relationships, and not get confused by spurious sort of factors that might be present. So focusing on the patient's anatomy, focusing on the animal's actual shape, and things like that. So in order to do that, one piece of work that I've actually worked on myself is called Invariant Risk Minimization. And the general idea is that if you acquire data from lots of different environments, what's going to be common about the data in those different environments is the sort of stable or causal sort of relationships and all the different spurious sort of things that are going to be different in each of the environments. And so you can kind of have all of these different environments have local classifiers that compete with each other in a game. And then when they compete that way, then they end up with a global classifier that focuses on what's common across all of them, rather than focusing on individual experience sort of things. And we've shown this to be pretty successful so far.

What are the key applications of AI where trustworthiness is the biggest issue and also the one where we can make the most improvement?

I think what I said at the beginning. Whatever applications where we have actual usage these days, and where there's high consequence to the mistakes in a human sense.

Well, I mean, the biggest one in people's minds would then be self-driving cars. Is the state of your work ready to address that?

Yeah, that's a good question. So in a lot of static problems these days, I think we're making a lot of progress. Once you get into these very complex, dynamic environments where you're using a lot of reinforcement learning and other things, I think we're much farther away. So if we're looking at lending, medical diagnosis, hiring, stuff of that nature, correcting the criminal justice system as well, there's plenty of machine learning problems that come up. They're much more oriented towards allocation decisions of some sort, and I think we're getting close, if not already there. So I think those are the immediate sort of impact places.

I see you've done a paper, a presentation about reducing unfair discrimination in AI. And one of the problems with AI and bias at the moment is that the data can be completely accurate and still biased because it represents where we have been, not where we want to go. And so it would predict the next president of the United States would be a man, it would predict that a female applicant to a medical institution should be a nurse, not a doctor. And, as far as the existing data goes, that's representing it accurately, but again, it's not where we want to go. How do you assist AI with where we would like to be as opposed to where we have been?

Yeah, that's an excellent question. One way to think about that is looking at one of two papers that came out of Sorelle Friedler and Suresh Venkatasubramanian and Carlos Scheidegger. So it's about different worldviews and kind of how there's a construct space, in which it's the ideal world where you actually don't have any biases or anything that is objectionable. And it's through the process of measurement, through the features that we actually can have access to, that social biases come up. So an example that's a little bit less black and white, compared to the two that you gave, is, let's say you're looking at a college entrance exam and there's a reading comprehension section, where you have to answer a bunch of questions. If you have some cultural knowledge that related to the reading passage, you'll do better on the questions. And so people who actually are not from that social group who know what the topic is, are going to do worse on the exam. And so it's not a true reflection of their ability to succeed in college. So this sort of social bias exists, and there's fairness metrics like statistical parity difference or disparate impact ratio, which actually characterize the equality of the selection rate for different groups, and they don't require the model, which we can only train on the biased features. And so in that sense, if we're equalizing the selection rate, we're assuming there is social bias, and we can work towards mitigating those biases. One thing I can say more actually on this topic is the fact that there's often, in the popular sort of consciousness, this tradeoff that people talk about between fairness and accuracy. And we had a student intern a couple of years ago, Sanda Harabagiu who looked into whether this really is a true tradeoff or not, and if you look into the constructs based in the ideal world, where there are no biases, then there actually is no tradeoff between fairness and accuracy. The reason we see it in practice is because the accuracy itself is being measured in a biased way because the only thing we can measure it with respect to is the data and the features that we have. So this tradeoff that people observe empirically, would go away if we actually lived in that ideal world that construct space.

I see. To evaluate this properly, to adopt AI responsibly in an organization, what role in a large organization should be doing that? Is it one that they already have, like data scientists, assuming that they actually do have a data scientist, a chief data scientist, or is it one that doesn't commonly exist yet?

Yeah, I think there's a few different ways to do it. Some companies are starting to have AI ethics boards or other similar sorts of structures getting created, which is a different sort of oversight. Some are doing it through their existing processes. I think the two important things, regardless of how it's done, is that there should be a variety of lived experiences among that board because there's this thing called epistemic advantage. For people who have experienced marginalization, they're actually able to better see all the potential issues and sort of power dynamics and things

than people who have not. So having a diverse set of people on that board, to better identify what all the possible issues are, is one important point. Second point is that there should actually be some power given to that board. Because it's easy to have a board that's kind of saying things, but then the rest of the company is not really following what's being said or doesn't have any accountability to those recommendations. So I think as long as those two things are there, there could be a few different ways to do it.

What problems in this area do you hope we've solved in five years?

Yeah. I think the main thing I would say is that we would have solved the culture of developing data science for high stakes applications, that it is moving towards a culture where, from the very beginning, people are considering things like fairness and robustness and safety and things that are beyond predictive accuracy. Because if we get the culture right, then everything else will happen. I'm not concerned about the technical aspects or anything like that, but just the mindset that various users and developers of machine learning systems need to have.

Whose responsibility is that culture? Is that the developers of those systems? Is it the CEOs of the corporations that are adopting it?

I don't know. I think the responsibility should be everyone's but as we talked about before, if there's regulations coming down, that will definitely spur things along. Having values that are coming from the top down within the corporation are certainly helpful. But again, I don't think that we should absolve the responsibility from individuals either because if we look at other professions, there are ethical sort of codes of conduct that exist, even in Computer Science and Electrical Engineering and everything. If you're a member of the ACM or the IEEE, there is a code of conduct that you've agreed to. So we just need to make sure that those are kind of part of the thinking as we go forward.

You talk about code of conduct in the section on government. Well, those are not the words, but you've got some principles there that read like a manifesto, perhaps. Do you think maybe the AAAI should publish something like that?

It can't hurt. So, again, if there's professional societies that are kind of stating things that their profession should uphold, then I think that is a valid sort of responsibility of a professional society, so I wouldn't be opposed to it.

So Kush's book, *Trustworthy Machine Learning*, you can get a pre-released version of that off his website, trustworthymachinelearning.com. Kush, thanks for helping us understand these issues more; they are certainly complex and in a short amount of time, we have done only a small service to the complex issues and the depth with which you have tackled them in that book so encourage people who want to put this stuff into practice to go look at that. Anything you want to say if there are people listening who have career aspirations in machine learning that might entice them to come and work with you?

Yeah, thanks, Peter, for highlighting the book. I would say, just spend some time thinking. That's my main message, I would say. It's very easy to jump into working on a problem and kind

of not think about the broader considerations. So as long as you're doing that, I think you'll be in really good shape, and we'd be happy to collaborate on anything with those sorts of folks.

Fantastic, Kush Varshney. Thank you for coming on AI and You.

Yeah, it's my pleasure. Thanks, Peter.

That's the end of the interview. I think it illustrates how much work is being committed to AI right now that's designed to make it more reliable, safe, robust, and of course, trustworthy. I liked hearing how many dimensions that issue gets into and how they're being fleshed out today.

In today's news ripped from the headlines about AI, CNN reports how a restaurant in Texas called La Duni is responding to the hiring crisis by using robots to serve customers. The owner was doing twice as much business as before on a third of the staff, so he contacted American Robotech, who supplied very neat looking robots, like a lectern with an iPad on top that take orders, and take food to the tables. They will even sing Happy Birthday. They have a bit of personality, like they'll wink at you, and giggle, and… well, the thing everyone wants to know is, are they taking jobs away from humans. The owner laughed at that idea, because he said he couldn't find any people to take the jobs. Of course there's room for a lot of systemic reasoning about why that could be the case, and I'm not going into that now. But if you're in Dallas, maybe you'll be served by a robot when you go out to eat.

Next week, I'll be talking with René Morkos. René is the founder of ALICE Technologies, which applies AI to building construction. That's what he got his PhD in. I had to know how AI could change construction, and so will you, 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