Bruke Kifle: This is ACM ByteCast, a podcast series from the Association for Computing Machinery, the world's largest education and scientific computing society. We talk to researchers, practitioners, and innovators who are at the intersection of computing research and practice. They share their experiences, the lessons they've learned, and their own visions for the future of computing. I'm your host, Bruke Kifle.

Artificial intelligence has become an integral part of modern society, transforming industries from education, medicine, to law enforcement and finance. More recently, we're seeing the emergence of generative AI technologies like GPT and DALL-E, further enhancing the potential of AI and revolutionizing the way we interact with technology, the web and the world more broadly.

However, as AI becomes more ubiquitous, there's growing concern about the potential for misuse, for bias, unintended consequences and harm. Ensuring AI is trustworthy and responsible has become increasingly crucial with a focus on distributing its benefits fairly and equitably to society at large.

In this episode, we delve into the topic of responsible and trustworthy AI and how it can be achieved with Dr. Kush Varshney. Dr. Kush Varshney is a distinguished research scientist and manager at IBM Research in New York. He leads the machine learning group in the Trustworthy Machine Intelligence Department where he focuses on applying data science and predictive analytics to various fields, including healthcare, public affairs, algorithmic fairness, and international development.

He's also the founding co-director of the IBM Science for Social Good Initiative. Dr. Varshney has contributed to the development of several open source toolkits, such as AI Fairness 360, AI Explainability 360, and conducts widely recognized research on trustworthy machine learning.

In 2022, he independently published a book called Trustworthy Machine Learning. He received his bachelor's in electrical engineering and computer engineering from Cornell University and Master's and PhD degree in ECS from MIT.

Dr. Kush Varshney, welcome to ByteCast.

Kush Varshney: Yeah, thanks, Bruke. It's my pleasure to be here. Thanks for the invitation.

Bruke Kifle: Certainly. I'd like to start off with a question that I ask most folks. You've had a very interesting and remarkable career from your graduate work at MIT to your long contributions at IBM. Can you describe some of the key inflection points within your personal and professional career that have led you into the field of computing, and specifically calling out any experiences or projects that have really sparked your interest in responsible and trustworthy AI research?

Kush Varshney: So I wanted to start actually with a quote. It's a proverb from Sudan, and it says that we desire to bequeath two things to our children. The first one is roots and the other one is wings. And I think both have been important, I think the roots and the wings.

So in terms of roots, so I mean, my family over a long period of time has been very much interested in technology, but also kind of in how to make social impact in various ways. And so I think that's been a big starting point. And then on the wings, just kind of taking things into the future as best as possible. So let me talk through that a little bit more.

One of my great-grandfathers actually was the first person from India to study at MIT. This was back in 1905. And he studied glassmaking technology, and then he went back to India, used that knowledge to start a school and a factory for glassmaking that kind of illustrated to the people of India, that they can have a self-sustainable industry and use that as a way to fight for [inaudible 00:04:05], which was the independence movement in India from the British.

And since then, I mean various family members of mine have been kind of straddling this technology and social impact sort of space. And in terms of, I mean computing, it's always been something where it's kind of like the newer new technology that I've grown up with and have wanted to contribute to because anything that is up and coming I think is the best place to make impact.

Yeah, in college, as you said, I studied electrical and computer engineering at Cornell, and I was drawn more to the mathematical side of that field just because of, I think I was a little bit better at it and it made more sense to me. I mean, electrical engineering is a very broad topic. So yeah, mean I started doing more on the civil processing and that sort of side of things, and then went straight over to grad school at MIT and got into doing more machine learning over time as an outgrowth of some of the civil processing work.

And then when I was looking for jobs, I knew I wanted to do research and I was looking for mostly industrial research sort of positions. And this was a time back around late 2009, early 2010 when we hadn't yet had the machine learning explosion that we've had in the last 10 to 15 years. I mean, a lot of machine learning specific research groups and so forth like there are now and IBM made a lot of sense to me. So this was just before IBM actually came out with Watson, which won the Jeopardy thing and reignited a lot of the renewed interest in artificial intelligence. But the group that I joined led by Saška Mojsilović was actually doing something quite unique, something I had never even imagined would be possible. Now it's kind of old hat, but back then the group was using machine learning to make predictions about people in various contexts.

So thinking about how machine learning can be used to improve human capital management, so getting employees to be better aligned with what they want to do and predicting employees at risk of resigning so that we can offer them

incentives to stay, various things of that sort. And also, now looking at the use of machine learning in healthcare. So again, both were things that I had not even imagined were possible, and that was what drew me to the group and the people as well. And that eventually kind of led to where we are today.

So when we were working on those people problems with machine learning, it kind of brought up the fact that we need to do a lot of explainability and interpretability of the models because these are very consequential decisions and it needs to be clear why they're being made. And then also the fairness aspects of it.

Because I mean, these are again, consequential decisions. If certain people or groups are being systematically disadvantaged, it's not a good situation. So all of that kind of started happening. We kept pursuing things in that direction. And in parallel, a few years after I had started working at IBM, I went back to my roots and heard about this organization. It was called DataKind. Actually, right at the beginning they were called Data Without Borders, but then they quickly changed their name to DataKind, and it was an organization to connect practicing data scientists, which at that time was a new term by itself with nonprofits and social change organizations to do applied work.

And so I got to work on some really exciting projects with a few nonprofits. And then after doing a couple of those projects, Saška and I sat down and said, "Maybe we can try doing something like this internally at IBM research. We have all these smart people who are dying to make a social impact." So we started that program and have done a lot of projects with various nonprofits addressing poverty and hunger and health and education and inequalities of various sorts.

So all of that came together doing the research, doing the social good, and then that led to, I mean, the open source toolkits as a way to bring things from the lab to practitioners. And then things kind of evolved from there, eventually the book as well that you'd mentioned.

So yeah, things have kind of been a natural progression. And I mean, in terms of inflection points. I think there's some things that I would point out. I think it's a mix of things that you choose to do and things that you choose not to do, and people don't necessarily think about it that way. So yeah, I mean there were points at which I had opportunities to leave IBM and do something else, which I chose not to do because I felt that the work we were doing here was so important and could be so socially impactful.

So yeah, I mean I think part of it is that, and then choosing to take opportunities, start the social good programs, create the open source toolkits, write the book, these sort of things are things that you do choose to do. So yeah, I guess the long-winded answer to your question.

Bruke Kifle: That's awesome. It seems like I really like this common theme of social impact that underlies all the work that you've done and continue to do at IBM. And I think in terms of inflection points, it's certainly interesting. I think most folks default to thinking about what actions they have taken to drive them to where they are today, but also thinking about the actions that you chose not to take or the decisions you chose not to make, how that has ultimately led you to your current place is also quite interesting.

One thing before we get into some of the actual technical details of your work in trustworthy AI, you lead IBM's machine learning group in the Trustworthy Machine Learning Intelligence Department. Can you explain the concept of trustworthy machine learning or within the industry we hear this term of responsible AI. What is it and why is it important?

Kush Varshney: I think of responsible AI as kind of an overarching umbrella within which we have a few different things. So there's AI ethics, so that's about kind of thinking about the principles and policies. So what should be done, what would be good and bad, what are the right things to do? And using AI in societally consequential application domains.

Then there's trustworthy AI. So this is thinking through how do we take those principles and actually operationalize them? What are the theories and methods and tools that we need to go forward and actually do this?

And then there's AI governance as a third aspect. So these are more of the organizational aspects. So once you have the tools, what does it take to actually make them part of the workflows of different organizations? And so on the trustworthy machine learning or trustworthy AI side of things, the way I think about it is what are the attributes that you need from a person to be trustworthy? And they're kind of the same of what you need from a machine learning system.

So you can think about it. Let's say you're trying to hire a carpenter to work on your house or something. There's probably a few things that you want from that person for them to be trustworthy. And in the organizational management literature, they've come up with several of these attributes.

So the first one is that the other person should be competent at what they're doing, so they should be able to do what they say. Second is that they should be reliable so that that competence sticks around in various conditions, in various settings. Third is that they should be able to communicate back and forth with you so that you can understand them, they can understand you, and there's some level of intimacy. And fourth is that they should be working for goals beyond their own goals. So they should be selfless in some capacity.

And all of these exactly map to what we want from machine learning systems as well. So a trustworthy machine learning system is competent if it has good

predictive accuracy, it's doing what it's supposed to really well. And then in terms of the reliability, there's a few different things. We want these systems to be robust to distribution shifts because we know, I mean, COVID gives us a great example. If you have data from before COVID and now there's a distribution shift and you have a model trained on that data, it might not work so well because the world has changed. So we want that sort of robustness. We want robustness against attacks. So if there's a malicious actor who's trying to make the system do something that it's not supposed to, we want to make sure that that's not possible or it's difficult.

And then fairness comes in here as well. So we want to ensure that our AI systems work as equally as possible for different people, different groups, and in different situations. So those are on that second attribute.

And then if we come to the third attribute, which is the communication back and forth. So there's a few things where the machine is communicating to us as humans. So that includes interpretability and explainability so that we as people can understand how the model is making its predictions. There's uncertainty quantification. So we want the machine to be able to tell us its own limits, kind of be intellectually humble in a sense. So if it's not confident, it should be able to tell us that it's not confident. And then some sort of broad transparency as well. So throughout the entire development lifecycle of that system, where did the data come from, what are the intended uses, what are different processing steps that have been done, what tests have been conducted? And release all of that in some transparent documentation, like a fact sheet or a model card.

And then, there's the other direction as well. So for us as humans or society to communicate to the machine of what we want, and this is often known as value alignment. And we can maybe come back to that, look what that implies and how it's kind of evolving now that we're seeing these very powerful models coming up and so forth.

And just to close the loop, the fourth attribute of selflessness. So using AI for social good, for making positive social impact is certainly one thing that I would categorize in that bucket. And then also I would say empowering all people no matter what station in life that they're in to be able to use AI technologies for meeting their goals and their purposes.

So all of that I think combines together to make what I would call is trustworthy machine learning.

Bruke Kifle: Very interesting. One thing that comes to mind is at least on some of these aspects that you described with, for instance, reliability, there's a need to balance the technical aspects of trustworthy machine learning with social and ethical aspects.

For instance, one of my early introductions to the responsibly AI field was an experiment that I was introduced to at MIT, the Moral Machines experiment, which is basically conducting a set of studies on how humans would respond to different morally challenging situations. You're driving a car and you're arbitrating or choosing between sparing the life of one individual versus five. So the classical trolley car problem in philosophy and ethics.

While certain aspects of responsible AI or trustworthy AI that you mentioned like fairness have very clear mathematical formulations that we can actually pursue and optimize, certain things like ethics or morals are actually less objective or they may vary based on personal values or religious values or cultural values. So how do you balance these technical aspects with these social aspects? How do you incorporate different perspectives and values into the actual design and development process?

Kush Varshney: Yeah, no, that's a really great question. And let me first describe one project that we did a few years ago and then extrapolate from that. So like you're saying, different folks do have different value systems, and it's very important actually to be able to bring those in the natural way that they might express them.

So the demonstration that we created back in 2019 was looking at using the Pacman videogame as an example. We wanted to have this moral behavior of not eating the ghosts, if you're familiar with the game. So we didn't want to encode it explicitly because in reality, I mean like you just said, I mean there's often not a very clear cut mathematical way to bring sort of moral considerations into a system. So what we did was we used a technique called inverse reinforcement learning, which is able to take demonstrations of the behavior that you want and to induce policies from that.

So we were able to learn a policy from people playing the game without eating the ghosts of what it means to be kind of moral in that sense. But just that policy by itself wouldn't do the whole trick because we still wanted the system to use its own creativity and its strengths as an AI system to play really well as well.

So what we did was we actually had two different policies. We had this moral policy that was induced from an reinforcement learning, and then we had the normal way of doing forward reinforcement learning to play the game to win, to maximize the points. And we had this thing which we called the policy orchestrator. So it was using this technology called a contextual bandit, and it, the terminology bandit comes from actually from Las Vegas. So if you're familiar with slot machines, they used to be called One-Armed Bandits back in the day. And people, I mean, kind of expanded on that as math problems and called these larger systems multi-armed bandits. And the idea is that you're kind of pulling one machine's arm versus another and trying to figure out which machine's arm you should be pulling.

So in our case it's should we choose the moral policy or the point maximizing policy? And it was interesting we were able to do this. And in the dilemma sort of situations where the Pacman was cornered and the ghost was chasing it and it was near this power pellet, it would kind of utilize the moral policy and everywhere else it would use the point maximizing policy.

So the point that I wanted to make is that yes, I mean we do need to have different ways of bringing in different policies and then use the context, the use case, the specifics of what we're doing to inform which policy is the most relevant and what's the most important to be done at any point in time.

And there's this paper, but the first author is Abeba Birhane, and it's about the forgotten margins of AI ethics. And the point that they make in that paper is that exactly like you said, I mean there's different religions, different worldviews, different perspectives, and they're also contextual based on the use case. And even the Moral Machines project saw this. I mean, there was no universal set of morals. So I think they found three big clusters of countries. So I think Asia, then the English-speaking world of UK, Australia and US-Canada, and then more of the Southern European and South America sort of cluster. And they had very different ... not very different, but there were enough differences that they did cluster separately.

And that's kind of my view right now on what's happening with these large language models, so with ChatGPT or with Claude or with Bard or any of these models that are now coming out. They're starting to have these different safety apparatus and they've been instructed in ways to limit someone's sense of what's bad. But that someone is the engineers in that company.

And I think that's the missing piece that we need to work on, which is how do we instruct these language models to be able to take principles and policies and behaviors, whether they're coming from indigenous knowledge or corporate policies or laws or even psychiatry or other sort of places that inform how things should behave and take them in their natural format and then use those to instruct these language models on how to behave, be contextual about it so it empowers the actual stakeholders, the diverse lived experiences that they have and see to actually create the models that work best for them, for their deployments, and deal with conflict as well like in the Pacman example that I talked about.

Because you're going to have conflicts. Different people even who are talking about the same thing will have different worldviews. So I think it's important to be able to have those sort of technologies, to make sure that the large language models that are now going to be part of our world in a big way are actually deployed in ways that make sense for the deployers.

Bruke Kifle: And in terms of operationalizing, we'll get into the LLM side of it, but in terms of operationalizing these REI principles, fairness, transparency, explainability, what

is actually needed to operationalize them and maybe some of the toolkits that you've helped develop with AI Fairness 360, AI Explainability 360, which I've played around with play a role, but in production settings, how do you see REI principles actually being operationalized?

Kush Varshney: The tools are a starting point. They're clearly not the entire solution. Yeah, I mean, when we talk to various people who have worked with our toolkits, it's great to hear that they've found them useful, that they're able to incorporate them into their natural workflows as practicing data scientists. But there's a lot of education that is needed beyond just having the tools and a lot of organizational governance aspects as well, because the tools without the people and processes is kind of sitting on an island.

So it's important to have either an ethics board or a bottom up sort of approach in a organization that will help institute policies and make sure that as things are being developed, different models, different applications, there is a problem specification phase that involves diverse stakeholders because when people have different lived experiences, they're better able to recognize various sorts of harms, especially if they themselves have been oppressed in some way.

So bringing in a diverse panel, making sure that you spend a lot of time on the problem specification phase like what is our goal, why are we doing this, should we even do this, how would we measure success? If we spend a lot of time on that, then you're in a good spot because then a lot of people are skilled to carry out and meet requirements, but setting the requirements themselves is the challenge to ensure responsibility.

Bruke Kifle: So emphasizing this people process technology triad, right? Technology alone is not sufficient. In the context of the actual open source toolkits, two questions. One, how closely do you actually work with industry practitioners or end users of these tools to help design a tool that can actually address the needs of practitioners? I suspect a big motivation for opensourcing is actually to build community and bridge the gap between responsible AI research and actual AI development. But how closely are you coupling with industry practitioners and end users of these tools and the design and development?

Kush Varshney: So yeah, the motivation for open sourcing was exactly as you said, to bridge the gap between what is happening in both academic and industrial labs and the actual practice. When we first created AI Fairness 360 back in 2018, there was nothing like it around. And so we felt that it was important to have the tools collected in a standardized way matching the syntax that the data scientists tend to use, and then have that. And it was important for us not just to dump some code, but when we first released it, we put together several tutorials and interactive web demos, several glossaries and other reference materials and so forth, and created a Slack community as well where anyone can come and ask questions and have discussions. And so we've kept that up over the last five

years, and, yeah, we always get questions from various sort of folks on our Slack channel that we address.

We've run tutorials at several industry conferences as well as academic conferences getting people up to speed. One nice thing has been since, so we donated the toolkit to the Linux Foundation a few years ago, so it's openly governed, but obviously IBMers are still very heavily involved. And some of the IBM consulting groups are actually able to then use these toolkits along with some enhanced additions of capabilities that are written in the same way, but just haven't been open sourced with a lot of different industry partners. So they've implemented a lot of fairness and explainability and robustness things with companies in the financial services sector and the retail sector in all sorts of different places, healthcare and so forth. So we have a way to do that that expands the reach of just the small number of researchers that we're involved to begin with. So yeah, it's been great.

Bruke Kifle: I've had a chance to play around with both IBM Fairness 360, and Microsoft has similar tooling with Fairlearn, and I found that these toolkits are actually very effective and powerful tools. And most of the use cases that I've actually experimented with or seen are supervised learning tasks such as classification or regression.

How do we think about using these tools for emerging technologies like generative AI or LLMs for understanding these notions of fairness or explainability? Are these tools extensible to these kinds of AI technologies or do we have to rethink the way we build or develop these toolkits to work well for these new technologies?

Kush Varshney: That's an amazing question. It's something that I've been thinking about the last couple of months quite a bit. There's certain harms and risks that are the same that are addressed by Fairlearn or AIF 360 or other things that show up in classification and regression tasks. And if you have a large language model that's being used for a predictive task, then a lot of the same tools can still be used, especially if they're in the post-processing part of the lifecycle.

But then there's a lot of new risks that come up when the output is a generative content. So if it's sentences or paragraphs or code or images or whatever have you that are being generated completely new that don't have a fixed set of categories or some number line on which they're being output, then yeah, there's a lot of new risks that come about. So things like toxicity, hallucination, lack of factuality, things called prompt injection attacks, there's a bunch of new risks that come about. And all of those are not addressed by these existing toolkits and methodologies. So that really requires a new set of approaches to address these new harms that are coming up.

Bruke Kifle: So certainly lots of opportunity areas in this space, it seems.

Kush Varshney: Yeah, absolutely. And yeah, it's something that we're banging our heads on. Like what is even hallucination? How do you define it? And then if we can define it, what are ways we can start mitigating those sort of issues? Yeah, absolutely.

And on the toxicity side of things as well, I mean, when this LLM is behaving like a chatbot and interacting with you, there's so many bad behaviors that it can undertake. I mean, it goes well beyond the biases that we would count under fairness. So yeah, I mean, it could be narcissistic, it could be, I mean, bullying you, I mean doing all sorts of things. And so just understanding, especially from a psychology perspective, what does it mean for these systems to behave in that way and what can we do to mitigate it is very much an open area.

Bruke Kifle: ACM ByteCast is available on Apple Podcasts, Google Podcasts, Podbean, Spotify, Stitcher, and TuneIn. If you're enjoying this episode, please subscribe and leave us a review on your favorite platform.

And I think that's a good segue to a follow-up question that I was going to raise, which is in light of some of the rapid advancement of these LLM technologies, we've also seen growing concern around harms, around governance, around regulation. As a result, there's been an open letter that's circulating, which I'm sure you're well aware of from the Future of Life Institute, which is essentially calling for a six-month pause on the development of AI systems. And we've seen this signed by many prominent researchers, and the letter really cites some of the important risks to society and humanity at large by these human competitive AI systems.

What are your thoughts on this call for pause? Do you think it's necessary? Do you think it's feasible? Do you think it's desirable? And I think the overarching concern is what's the trade-off between innovation and safety more broadly? Do we stifle progress in this field and AI research more broadly? Curious to get your thoughts on this.

Kush Varshney: I guess you've already, I mean, learned that I like analogies and so forth. So let me give you another analogy. So if we look at commercial aviation, so between 1903 and 1958, so when the Wright brothers had their first flight and when the Boeing 707 was introduced, so this was a time, the first 50 or so years where it was all about just trying to get planes to fly, just get them to work. And then since the Boeing 707 was introduced, there has not been really a fundamental change in how airplanes work. So today's aircraft are extremely similar to the Boeing 707, but what has changed is that in the second 50 years, there was much, much more focus on safety and efficiency and automation. For example, fatality rate per miles flown today compared to the 1970s is something like 300 or 400 times less, and the number of miles flown is, I mean, much, much larger.

So is a shift over time where once you have something that is working, then the onus and the focus shifts to the safety aspects. And we're kind of completing that first 50-ish years of AI as well. So now we have these things. You could say

it was around 2012 when deep learning when the ImageNet was that the end of the first 50 years was it when how the transformer architecture came about in 2017? I don't know exactly, but we are at some point where AI is performing, it's being used in real world things. So now we are at that point where the focus must be on safety.

And to me, a complete moratorium is not something that helps push safety or alignment or trustworthy AI research. I think everything goes hand in hand. I do think that there should be more focus and more regulation as well. And when I say regulation, it's not just laws, but regulation through social norms, through the market, through various things that kind of make sure that that focus shifts. Because again, coming back to the aviation, I mean, if the industry had not shifted to more safety, then it wouldn't be where it is today and the creation of different regulatory bodies happened, but the market was asking for it as well. So I think this is going to be a natural progression where we, at least that is my hope, that we work towards more of the safety aspects as we go forward.

And just one more comment about the different views of what safety should be about. There's kind of a long-term view of what AI can cause in terms of existential risks to humanity, and then there's much shorter, clear and present sort of dangers that we see, I mean, every day. And to me, I think that every day things that all of us right now are encountering, the bullying from these systems or the fact that they lead to an equal allocation of things or that we don't know what's happening, that there's some incitement of violence or other things that these things are doing now is where our focus needs to be.

So yeah, I mean the letter is fine. I mean, it's great that people want to point out that there's risks. I like that. I encourage that. But I think the way to make progress is focusing on things right now that affect people now and just making sure that AI research is working towards safety.

Bruke Kifle: Yeah. I think it certainly sparked a good discussion and dialogue within the industry. And I certainly agree that while there are two classes of both short-term and long-term risks, there are very tangible risks that we're seeing in the immediate short-term. And so prioritizing some of those safety and responsible AI improvements and mitigations, I think is certainly of interest.

In terms of the path forward for governance and regulation, how do you see the role of different stakeholders, whether it be academic institutions, whether it be government, private sector, civil societies, professional societies, how do you see these different stakeholders coming together in that path forward for AI governance?

Kush Varshney: Yeah, I think everyone, all of those different stakeholders have a role to play. And one big group that you didn't mention was us, I mean, all of us as people, especially people who are maybe less powerful. And, yeah, everyone should be bringing forth their principles, their policies. And then systems should be

designed in a way that can bring in those kind of policy packs as input to guide and control the behavior of the systems as they go forward, deal with conflicts, as I was saying before.

Yeah, I mean, I think government regulation has its place. Self-regulation by industry has its place. Watchdog civil society organizations, I mean, civil society kind of exists because I mean, it is a kind of criticism to government and industry because if both government and industry were perfect by themselves, there would be no need for civil society in a sense. So yeah, I mean, I think everything is important and I think there's ways to bring all of it together into specifying the behavior of these systems.

Bruke Kifle: So as part of this effort, I think one of your major contributions is the publication of your recent book, Trustworthy Machine Learning. And I would like to quickly give you the opportunity to plug this publication, where it's available, but I would love to learn more about what inspired you to write this book, and I'm sure it's a big part of what we discussed, the path forward. How do we as individuals, as practitioners, as researchers, academics, democratize this information or this knowledge, but what inspired you to write this book? And one interesting thing is you actually chose to move forward with self-publishing. So what was the motivation with self-publishing and how was that experience?

Kush Varshney: In terms of writing a book, I mean, it's obviously a very large effort, so it has to be something where the motivation can't be just that, "Oh, I want to write a book, so I'll write a book." It has to be that you have something unique to say that you don't think that anyone else could say it, and that people should, I mean, have a need to hear. And I felt that way a few years ago. This was, I mean, right before COVID that I started writing the book and it was kind of ... I mean, I'd had about 10 years of experience up to that point, had a unique sort of way of approaching machine learning and using the projects that we did through social good. Also, the client sort of projects, the human capital management, et cetera.

So I mean, I kind of had this unique perspective of this is what the starting point is also interacting with different practitioners, being a practitioner partly myself. I mean, that was the motivation that there was a missing piece that if there's a practitioner who wants to do things in terms of responsible AI, how do they think about it? What are the conceptual aspects of it? Because to me, a lot of trustworthy machine learning isn't difficult to carry out if you start thinking in the right way to begin with. So that was the goal.

Yeah, the book is available as a free PDF at trustworthymachinelearning.com, and it's available on Amazon as a paperback. It's I think $6.85, which is the least possible price that I could set given that Amazon has to print it and they have some costs for that.

The reason I eventually self-published or independently published is because I wanted this knowledge to be out there, so people all over the world should have the ability to get the knowledge and to be able to put it into practice.

So even the first week it was released, I got an email from this student from the Ivory Coast, and I mean, he just was gushing that, oh, I mean this is so useful for me. And I've seen that again and again, that people are dying for knowledge. And if you kind of gatekeep and put a book up and it's like $80 or something like that, I mean, what's the point? Really what we want is people to have the knowledge to be able to use it and to make the world better for themselves and for their communities and for everyone.

That was the motivation. And I think it's been doing well. I'm not tracking how many people have the PDF, but even in terms of the book sales, it's reaching close to a thousand, which I think is good for this sort of publication. Yeah, I mean, it's been good for everyone, yeah, that's been able to take advantage of it.

Bruke Kifle: And I've personally been making my way through the book and I've so far have thoroughly enjoyed it. So you have one endorsement in me.

Speaking on the topic of the book, I know we've discussed a lot of the key trustworthy AI principles and concepts, but what are some of the key takeaways that you hope readers will gain or walk away with?

Kush Varshney: I mean, yeah, we've covered a lot of the content through this discussion, but I think the biggest thing is, yeah, I mean, start with the use case. What are you trying to do and why? And is it something that really should be done or not? And then progress forward to the machine learning aspects, because machine learning these days, I mean, there's great tools out there, but the question isn't how do I use the tool? It's more about what is the right thing to do? And having that as your starting point, I think is the biggest message.

And the sort of theme that kind of winds its way through the entire book is this kind of call for people not to take shortcuts, because I think that's where a lot of the issues crop up, where they're trying to do things very quickly without stopping to think, without stopping to think about what mitigations there might be because it's very easy to want to take shortcuts. But if you're somewhat wanting to be responsible, then that's the message.

Bruke Kifle: I'd love to quickly touch on some of your work with the IBM Science for Social Good, and you provided some pretty cool context with the motivation. I think you mentioned how you started off with DataKind or data ... was it Data Scientists without Border?

Kush Varshney: Data Without Borders was the original.

Bruke Kifle: Data Without Borders.

Kush Varshney: And then DataKind was the name.

Bruke Kifle: Yeah. So given your work with the Science for Social Good Initiative and your experience in leveraging technology to address these pressing societal challenges, how do you think about identifying problem areas to pursue, and what are some of the challenges that exist when you're developing technology solutions for new markets, new regions, new context? Developing a solution in the western world in the US is very different from developing a technology solution somewhere in rural India or rural Ethiopia where I'm from. So what are some of the challenges that exist when developing these solutions in new contexts?

Kush Varshney: So actually, I'm going to be presenting a paper at the ICLR workshop on practical machine learning for the developing world in a couple of weeks. And the discussion in that entire paper is this. So when you're thinking about the developing world and trying to develop AI sort of technologies, what are the considerations and how should you go about doing it? And in that, I kind of make an analogy again, so my favorite thing, to this concept of bottom of the pyramid innovation.

So this is the idea that if you're, whatever, creating a stove or I mean any sort of technology product for the developing world, then there's a lot of different requirements that come about. And a management professor [inaudible 00:45:19] who came up with this, I guess close to 20 years ago now lists down like 12 different characteristics. I'm not going to go through every single one, but the main things are that you need to start from the user's perspective. So start with the people who you're serving. Ask them what is it that they need. Try to understand that as deeply as possible.

So in the AI world, I mean people have started using this term participatory design. So that's certainly related. So it's a question of really understanding what is the true need, and then going from there, because what I can imagine sitting in my lab in Yorktown Heights New York is very different than being on the ground and experiencing what is truly needed.

So this bottom of the pyramid innovation sort of approach starts there. And then there's these other characteristics. The technology needs to be robust, inexpensive, appropriate, have good user interface that's matched to the people, I mean a bunch of different things, and have some reusability as well, so that they're kind of more of a platform sort of approach.

All of these things are what in some ways we've been working towards, and actually, I mean these language models, these chatbots are almost there in certain capacities because now it is true that anyone can really interact with them to solve their own problems.

I had an interesting experience in January where this organization we had worked with in the past, so it's the International Center for Advocates Against Discrimination. We had helped them develop some natural language processing technology, which marks different documents by which of the UN Sustainable Development Goals are discussed in a particular sentence. So these, for those that don't know, are a set of 17 goals that the member states of the United Nations ratified in 2015 as things to work towards by 2030. So things like no poverty, no hunger, specific indicators and so forth.

Anyways, we developed this NLP thing five years ago. Took us a whole summer. We fine-tuned it, there's a bunch of us working on it. And in January I sat down with the person from that organization, she had heard of ChatGPT by then, but she hadn't thought of the fact that, "Oh, I could solve my problem using ChatGPT." So we sat for two minutes and it just worked out of the box.

I mean that democratization is happening. The price point, the environmental concerns amortize over a lot of users and it is robust. We're getting there with low resource languages. Eventually, I'm sure that we'll be there, different interaction patterns and so forth.

The thing that is kind of missing though is the appropriateness in terms of the interaction with other things. So it's not the best right now in terms of the other tools that lower resource organizations might use. So if you have some Excel file or something like that and you want to use that to then interact with one of these chat systems, there's still a gap. And if having a little bit more appropriateness in terms of the integration with other stuff is the missing piece, but we're kind of getting there. So in my view, I mean these sort of technologies that we're seeing today are a form of democratization. They do have the robustness to work in hostile conditions and so forth.

Bruke Kifle: So beyond the actual maturity of the technology itself, in emerging markets there aren't the same financial incentives for tech development as there are in developed markets. The market may not be as large, the potential for paying for services or products may not be as high. So what do you see as a sustainable model for incentivizing, whether it be large corporations or startups or entrepreneurs to actually drive progress in this area and develop tech solutions for emerging markets?

Kush Varshney: Yeah. So the part that I forgot to mention about this in bottom of the pyramid innovation idea is that once you are developing technologies for that large base of low resource people or organizations, the solutions are often better and cheaper even for the top of the pyramid. So the incentive is, I mean, if you do something really good for that bottom, then it's also really good for the top.

In India, there's this eye hospital, Aravind Eye Hospital, which does extremely cheap cataract eye surgeries, and they've worked it out so I mean, they can do so many so quickly and the results are 10 times better than what you get in the

US in terms of the health outcomes and they're 1,000 times cheaper or something like that. So similarly with shampoo or cars, I mean all sorts of different things. So once you have that incentive that by focusing on the bottom that you'll also be better able to serve the top, then I think that's a really great motivation.

Bruke Kifle: This has been a very interesting conversation. I'd love to wrap up by discussing some future directions. I think as we've touched on, we're witnessing an AI arms race and we're seeing a lot of the rapid evolution and progress of generative AI, which to your point, is now becoming a platform where users can essentially explore different use cases and applications on top of these foundational models or technologies. So we're seeing conversational AI, we're seeing transformations in search and marketing and education.

What are some of the most exciting developments that you see for the future of AI and how do you see these developments impacting the business world in the coming years?

Kush Varshney: Yeah, I'm sure by the time the podcast is released, everything I say will already be done because the field is moving so fast, honestly. I mean, it's even someone in my position, it's so hard to keep up. Every hour of the day there's some new development.

Yeah, I mean, I think there's going to be multimodal these models, these foundation models to incorporate all the different modalities, image, text, tabular data, scientific data, I mean all sorts of stuff. So combined models for those. I think that's going to be out there very soon, I think to deal with. I mean, some of the risks that I talked about before with hallucination and so forth.

One thing that needs to happen is kind of separating out the memory, the facts, the information from the processing of the language, because right now they can get all mixed together and intermingled. So if we can keep those separately, then we'll avoid some of the issues that we see right now.

Yeah, I mean, I think just from a business perspective, just the incorporation of well-designed, well-rounded sort of systems that are appropriate for consequential applications because the state right now doesn't allow these things to really be used in operational sort of settings most of the time because there's kind of this hump that we have to get over to make them more trustworthy, more safe for those enterprise applications. And once we do, I think we'll see a flourishing of activity.

And one thing that I think that we will also have to think about very seriously is ensuring that the people have the dignity and agency remaining in their work lives with these technologies so that it's not, the technology is the driver and the human is just along for the ride, but there's some joint collaboration that's

designed into the way that these things interact with us so that it's utilizing the strengths of humans and AI, but also, I mean giving us the dignity of work.

Bruke Kifle: So AI as a copilot.

Kush Varshney: Yeah, copilot or some sort of advisor or something like that. Yep.

Bruke Kifle: Finally, what advice would you give to young students, researchers, practitioners, like myself in the field of AI, and how can they ensure that their work consistent with your personal careers theme contributes to a more just and equitable society?

Kush Varshney: I think be solid technically first of all. Just like with the trustworthiness, the first attribute is competence. So definitely focus on that and then start bringing in these other attributes. So just like the technology I mean has these attributes that we want from it, for ourselves, we want the same things. So we want to be leaders that are standing on a solid foundation and then bringing in the selflessness and the justice and all of those things that go beyond it.

And yeah, focus on a broad-based education that you're learning everything because I use everything I've learned in everything that I do. Those would be my recommendations. And, yeah, just believe in yourself, believe that your worldview is something that others will value.

Bruke Kifle: Well, I think those are some pretty good bytes for our audience. Dr. Kush Varshney, thank you so much for joining us. I think we are certainly at an inflection point in technology history, and I think with some of the amazing and foundational work that you're driving in trustworthy AI and responsible AI, I think the future is bright and optimistic. So thank you for all your work, and thank you for joining us on this episode.

Kush Varshney: Yeah, it was my pleasure, and I hope it'll be useful for the listeners as well.

Bruke Kifle: Thanks.

ACM ByteCast is a production of the Association for Computing Machinery Practitioner Board. To learn more about ACM and its activities, visit acm.org. For more information about this and other episodes, please visit our website at learning.acm.org/ B-Y-T-E-C-A-S-T. That's learning.acm.org/bytecast.