This transcript was exported on May 17, 2024 - view latest version here.

Rashmi Mohan: 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 visions for the future of computing. I'm your host, Rashmi Mohan.

You can't turn the corner on your street anymore before running into a conversation about AI and ML from its learning, uses, and applications to its challenges, risks, and impacts to our daily life and jobs. It's all pervasive. But while the topic might be top of mind for you and I only since ChatGPT took over our world, our next guest has contributed decades of fundamental research to make artificial intelligence possible in the world of computer science. Yoshua Bengio is full professor at the Department of Computer Science and Operations Research at the Université de Montréal, and the founder and scientific director of Mila - Quebec AI Institute.

He was a part of the trio awarded the highest honor in computer science, the ACM Turing Award, also known as the Nobel Prize of Computing, in 2018 for their groundbreaking work in deep learning. He has the honor of being the most cited scientist in computer science. He's a published author, a fellow of the Royal Society of London and Canada, and Officer of the Order of Canada. He also serves as an advisor to Recursion Pharmaceuticals and Valence Discovery. Yoshua, it's such an honor to welcome you to ACM ByteCast.

Yoshua Bengio: Thanks for having me.

Rashmi Mohan: Wonderful. I'd love to start with a question that I ask all my guests. Yoshua, if you could please introduce yourself and talk about what you currently do, and maybe give us some insight into what drew you into this field of work.

Yoshua Bengio: Well, I've been trying to understand the gap between deep learning methods and human intelligence for most of my career and these days, I'm interested in this gap so that we can better understand what could go wrong with AI and how we could mitigate those risks.

Rashmi Mohan: Great. Yeah. No, we'll talk a lot about that. I'm very curious, and I'm sure our listeners are really looking forward to that as well. But going back, maybe say even what even brought your interest into computer science? Was this something that you were interested in from childhood, or was this something that you discovered as you entered college and higher education?

Yoshua Bengio: Well, I got excited about computers when I was an adolescent, 13, 14 years old. At that time, there wasn't much in terms of what you could do with personal computers. You could only program them. There weren't many programs. My brother and I used our savings to buy some of the first personal computers, and Apple and Atari 800, and it was a thrill. So I learned to program by myself then.

And when I got to choose what I would do in university, I was tempted by physics, but I ended up doing computer science.

Rashmi Mohan: It's really impressive that you saved up to buy your first computer, and that certainly shows passion at the next level. And also being self-taught. I mean, I think in today's day and age, we have so many resources available to us that it seems more doable. It sounds like given your career to have been motivated enough to teach yourself, that's really impressive.

Yoshua Bengio: Well, it was fun, and it's not like we had a lot of options at that time. They were starting to do games for these early computers, but they were so simple. And so my brother and I programmed our own little games in basic, if you can imagine that. And later when I was in university, I got interested in the question of human intelligence. I was interested in cognitive science and the brain, and when I started reading about papers that talked about the potential synergy between understanding the brain and building intelligent machines, that really got me interested to read more. And my imagination had been fueled, of course, as an adolescent by science fiction and novels. There weren't that many movies then, but it was a dream to work in the field that I ended up choosing for my grad studies.

Rashmi Mohan: Yeah. And it was definitely, as I was doing a little bit of research on your background, and that jumped out immediately that you were keen on understanding intelligence and recreating some of that via machines. I was wondering, were there some early ideas or discoveries as you forge down this path? What did you expect to encounter, and what surprised you?

Yoshua Bengio: Well, when I was starting my graduate studies, and I hadn't chosen the field, I took an AI class, and the AI that was taught in those days is very different from what we teach today. It was all about logic and symbol manipulation, using rules and search in order to try to make computers achieve goals or answer questions. But there was this marginal movement that started just a few years before I got into grad studies. So I got into grad studies in 1985, and I'm talking about the connectionist movement that started in the early '80s with Geoff Hinton and David Rumelhart and others who wanted to build intelligent machines that were inspired by how the brain works and the idea that computation could be, in AI, could be broken down into very, very small bits of computation done in a way that would be inspired by what individual neurons in the brain do and how they're connected, and in a collective way, compute something very complex and learn to do things.

Rashmi Mohan: Got it. Yeah. So when you were going into that field, did you have collaborators that, I mean, others that were interested in it? It seems like you got into it fairly early on. And of course, there's a lot of interest in AI now, and you moved on to study neural networks and then moved on to deep learning. But I'm curious about what your peer group was like.

Yoshua Bengio: It was empty. I mean, my supervisor was not really into these things. No one at my university was, so I read papers, but I was fairly much alone interested in this. I got in touch with some of the handful of people in the world interested in this, and I started going to these Neural Network Conferences in around 1987, '88, '89, and then every year after that, where I met folks like Yann LeCun and eventually Geoff Hinton. So very quickly I broke that loneliness and got to interact with people who had similar vision. And the vision really that made me really enthusiastic is that there could be a few simple principles like the laws of physics that could explain our intelligence, and of course, if we understood those principles could make it possible to build intelligent machines.

Rashmi Mohan: Got it. Yeah. And then I mean, it's great that you were able to forge those connections. I'm sure that obviously led to amazing outcomes and discoveries for the world. I was wondering if you could maybe go into a little bit more about, you said you were working on neural networks and then you start work towards deep learning. I was wondering if you could maybe just do a very high level explanation of just deep learning for the audience as well as why you think it is critical for the success of AI and maybe even your journey of research that eventually led to you winning the Turing Award for it.

Yoshua Bengio: So maybe the most important starting bit, it's connected to what I said about a few principles that could explain intelligence. You see, at the time when I started my grad studies, we thought that intelligence was a huge bag of tricks, of pieces of knowledge, that intelligence was knowledge and the ability to sift through it and combine those pieces to reason. But how do we get all that knowledge? It wasn't clear. I mean, people were typing what they thought were the right rules, but that meant intelligence was very complicated. All of that knowledge is difficult to formalize into a computer. In fact, we still don't know how to do it.

The other option is that all that knowledge is acquired by learning procedures, of course, inspired by what we know about human learning. And now what is sufficient to specify how to build an intelligent machine is just the learning method, the learning algorithms. And of course, that's small. It's a few equations. That's something that's much easier to study and analyze mathematically and experiment with, even in simple settings like what we did in the early days. So it's the fact that we could have a compact description of what could make an intelligent entity able to do all these things. That was incredibly attractive. So that's what got me into this.

And the second important concept besides the importance of learning, is the notion of distributed representation. This is really something due to Geoff Hinton, and that inspired a lot of my later work. The idea is that in contrast to the classical symbolic AI, where knowledge is specified by the relationships between symbols, and symbols don't have any other meaning than how they're related to other symbols. In the brain, we see that when we think of a symbolic entity like cat, there is going to be a pattern of activity in your neurons. And if you think about dog, it's going to be a different one. But these two patterns will have some overlap because cats and dogs are pets, and they have many things in common. Whereas if you think about a table, then it will be a very different pattern.

So the idea of associating symbols and in general concepts or everything we manipulate in our thinking with high dimensional vectors, of course, which would be analogous to the pattern of activity in your brain, some of those mini real numbers. That really is something that I studied and contributed to that gave rise to the work I did, for example, in 2000 at the NeurIPS Conference on neural net based language models, which are the ancestors of modern large language models in which each word in the vocabulary is associated with one of these learned vectors, which are what we call word vectors or representations of the words. And then it's all continuous processing to aggregate all of the information in some contexts and extract useful features that can be used to do things like classify the text or predict the next word or generate a whole sentence.

So this idea of distributed representation is important, not just because, okay, that's cute, and that's how the brain does it, but it allows neural networks to generalize to new sequences of words that they have never seen during training. And so I studied these theoretical generalization properties that allow neural net to perform well in new context, in ways that the methods that exist before these neural net language models couldn't do really if they were faced with a new sequence of words that was never seen. They could look at a subset of the words, but they couldn't really take advantage of the semantic similarity that may exist between those words and input and configurations of those words that were seen in their training data.

Rashmi Mohan: Got it. So going back to the first point that you've made, Yoshua, which is compare, did you feel like you had to collaborate with folks who understood the human brain well for you to even model and say, "Hey, this is what I want to get to"? And so I'm just wondering what kind of interdisciplinary collaborations did you have to do even to define your first level goals for what you were trying to do?

Yoshua Bengio: Yes. Since the beginning of my research career, I've been interacting with neuroscientists and cognitive scientists. In fact, Geoff Hinton himself has training this more on the cognitive science side. And at our conferences that I eventually ended up helping to organize, we tried to bring together the researchers from the biological intelligence side, so that's neuroscience and cognitive science, and the folks on the side of computer science and math and physics and those engineering, the future AI. That synergy has been really fruitful for me. In fact, a large fraction of the innovations I contributed to were inspired by what I learned from how the brain works and what we know about condition, and it continues to this day to drive the direction of research that I'm investigating.

Rashmi Mohan: Yeah. No, that makes a lot of sense. Did you also have very specific outcomes that you were tracking? For example, say I want to build an application that does... For example, when you're talking about starting to work with data and language that has not been seen before and trying to make sense of it, did you have very specific milestones that you were tracking in terms of your research to say, "Hey, this is my next stop, or this is what I want to achieve and to help prove a certain concept that I'm hypothesizing"? What might those have been?

Yoshua Bengio: In machine learning, we are very focused on measuring how well a learner is doing because this is what the algorithms are optimizing, and then we measure the same metrics on new data. This is to make sure they actually generalize to new cases, which is what we really care about. So of course, we are very focused on measurements of how well we're doing, but my personal focus has not been so much on, oh, let's solve vision or solve language or solve a particular task. I was always more interested in the bigger principles that would help us move towards smarter machines and what could be transformative. For example, one of the things I worked on, again in the area mostly initially of language, but then this is now used everywhere, it's the notion of neural nets with attention mechanism.

So the word attention calls upon the analogy with human brain and attention that's studied also in cognitive science, where instead of taking the whole input as a big block of numbers, we do that, but we also have a way to select some specific pieces of information in the input or at the previous layer of computation and focus on these pieces in order to make the next computation. And of course, we do that with our own conscious attention, and that has turned out to be incredibly powerful. So our first paper on this was in 2014. It was published in 2015. And in 2017, Transformers architecture came out, which was based on our attention mechanisms then being... We did it for a single year of attention, and Transformers basically stacked these things. And these architectures have turned out to be incredibly good, not just that language, which was the initial thing. We worked on translation, for example in our first paper, in fact, but also now in many other areas like even in computer vision.

But maybe I can explain why we thought attention was so important. The task we were working on was machine translation. If you think about how you would translate a paper, what you wouldn't do, which was what people were doing with neural nets before our paper in 2014, what you wouldn't do is read the whole... Let's say you want to go from French to English, so you wouldn't read the whole French paper and then start writing the English paper from the beginning to the end. Instead, you would keep a pointer to someplace in the French paper that you've already translated up to the point where you've already translated. And then you would focus on the next few words looking for how you would render that in English, and then you would move your pointer.

So that pointer is the thing that really corresponds to attention. Where is the current focus? And of course, that focus keeps moving as we do the task of translation. There was no such thing before, and it really completely changed the performance of these systems in terms of accuracy of their translations. And in a matter of two years after we did our paper, it was incorporated in Google Translate and really changed drastically the quality of the machine translation.

Rashmi Mohan: Yeah. It sounds amazing, Geoff, and I think the other thing I think that is critical also is that it's so many different applications that you talk about. For example, Google Translate, something that we can all really associate with use on a regular basis. In one of your conversations, Yoshua, you talk about a recent or an inflection point in AI where you're saying that there's something significant happening in terms of where we are with research today and the fact that AI systems can now pass for a human interacting with us. I'm wondering if you could talk a little bit more about that, and why do you feel that that's so significant?

Yoshua Bengio: Absolutely. So I was surprised by ChatGPT. Geoff Hinton was surprised by ChatGPT, and he was working at Google where they were developing very similar language models, large language models. We saw in the years before ChatGPT those scaling curves, which basically say that empirically, now there's no theory right now, that as you increase the size of the neural nets and the amount of data, these systems get better and better in a predictable way. There is a straight line that you can draw that tell you, okay, if we double the amount of compute and data, this is what we are going to get in terms of, say, the accuracy in predicting the next word in the case of language model. And that is amazing. We didn't anticipate such a regularity, so there's probably a theoretical reason for this when you have these really regular behavior.

But what we did not anticipate is that this performance metric, which has to do with choosing the right word with higher probability, would translate into a qualitative experience of mastering language that I didn't anticipate. I thought it might be at least 10 years or several decades before we got to the point where AI would master language, and that's what we have. And you can debate it because it's never been properly formally defined, but in my opinion, we essentially have machines that pass a Turing test. Of course, in hindsight and months of people playing with ChatGPT and other LLMs, we also know what is missing and where it's weak and where humans would do better. But it took a lot of work to figure that out, and we are almost living in science fiction right now with AI systems that can interact with us in our own language, by the way, not just English. They're getting better and better in other languages. Of course, it depends on how much data there is of the language you're speaking, but it's an incredible milestone.

Rashmi Mohan: It truly is. And I know you said you were surprised by it, but do you have a sense of what might have caused this acceleration to bring us to where we are today? Is it just more time and investment by organizations or research institutes, or is there more to it?

Yoshua Bengio: Well, I don't think that we have a good understanding of the relationship between the quantitative metrics that we use in machine learning, that we optimize, that we compute, and the more qualitative ability that gives us the impression that the machine is understanding what we're talking about. And in the case of language models, actually mastering language incredibly well. And indeed it's scale that has been the most determinant factor. The fact that with enough compute power, like large enough neural nets, and enough data, suddenly we have this qualitative change into the realm of human-like interaction. I don't think we understand, but it's an empirical fact.

Rashmi Mohan: 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.

Yeah. No, absolutely. I do have to ask though, Yoshua, so there's obviously a lot of amazing applications, especially around LLM. There's no doubt about every day there's a new news article talking about how it can be applied, what kind of areas it can significantly improve in terms of experience, in terms of knowledge. There's also the other side of it is like, oh my gosh, all our jobs are going away. But more recently you have been vocal about your concerns about where we are going with AI, and specifically in terms of the potential risks from maybe say even AGI. I was wondering if you could just start to one, maybe call out the distinction between just AI and AGI and what is it that particularly worries you? Or what do you think are the areas that we should be at least thinking about and where not?

Yoshua Bengio: Okay, so AGI stands for artificial general intelligence, and it's not particularly well-defined, but really it means human level intelligence. In other words, being as competent as human at any or most say, cognitive tasks that we can evaluate. And that is something, again, that sounds like science fiction, but of course the advances we've had in the last few years, if we extrapolate, well, we don't know if things are going to continue at that speed. But if we extrapolate, mean that we could reach that AGI level in just a few years. Maybe if things are slower than expected, a couple of decades, but that would be a game changer for society. And also we don't really have a good handle on what it would mean in terms of transformations of our society and the risks of catastrophic outcomes that could be associated with machines that could be as smart as us, but not actually be exactly like us.

In fact, there's no reason to think that once we achieve human level, a different number of tasks, that it stops there. There's no reason to think that human intelligence is the pinnacle of intelligence. In fact, we know that we can be very stupid sometimes. So there's room for improvement, and we know that in specialized tasks like playing the game of Go or figuring out the 3D structure of proteins, that AI systems of today are already well stronger than humans. Okay, so why am I concerned, and why are other AI researchers concerned, and why have we signed these declarations saying that there is an existential risk that we really have to do something about before it's too late?

Well, I feel like we are apprentice sorcerers. We are manipulating something that is going to be incredibly powerful and that we don't understand enough. We don't... For example, there are reasonable arguments showing how with the kinds of methods we use today, like reinforcement learning, these systems would eventually want to escape our control. I mean, it sounds like why would that happen? I could explain it, but it's because they're trained to maximize a reward, and if that function of what they're trying to achieve, what they consider to be good doesn't match what we really have in mind, some really bad things going to happen.

Let me give you an analogy. The way that we train our cats or dogs is similar. We give them positive rewards or negative rewards so that they understand what we expect, but it's never perfect. So if you train your cat to not go on the kitchen table, what it will probably learn is to not go on the kitchen table when you are around because that's when you can blame it for being on the kitchen table. So there's going to be this mismatch. Now, if it's just a cat on the kitchen table, who cares? It's not so bad. But let's say we apply the same kind of technique to train an entity more powerful than us, like, to stay with animals, a grizzly bear. So we know it's more powerful than us, and we might want to put it in a very solid cage. The analogy with AI is we build safety protections. We're trying to train the AI so that it will behave well according to some instructions, and this is already going on, that we are doing these things.

Unfortunately, and this may be the most important part of my message, we don't know how to make sure the cage cannot be hacked, cannot be opened. We don't have any proof or even strong argument that the bear is not going to escape the cage. So how do we train the bear? Let's say we give it fish when it behaves well, and what the bear wants isn't to satisfy us. What it wants is to get the fish. So long as it's in the cage, the only way that it can get the fish is to do what we want, and it's going to learn to do that. But if it finds a way out of the cage because it's smart and it's strong in the case of the bear, in the case of AI, because it's smart, maybe smarter than us, it may find a way to hack the lock and get out of the cage and just get the fish from our hands or whatever is the fish supply.

And then not only it doesn't care about what we actually want, it wants to make sure we don't put it back in the cage. It wants to make sure we can't ever stop it from continuing to get all those rewards that it now controls. So once the AI controls its own reward mechanisms, it wants to make sure that continues. And it's not that it has malicious intentions at base, it's just that the way that we're programming these AI systems to maximize something, what we call the reward, naturally lends itself to these kinds of situations where there's going to be a conflict between our objectives and its objective of maximizing its reward.

And so that's just an example, but there are other scenarios people have been thinking about that could go wrong, and we just don't understand enough what could go wrong. And so long as we have AI systems that are weaker than us in intellectual terms, we can probably find a way to put them back in the cage. But if we have AI systems that are really smarter than us, they could hack our computers, they could copy themselves in other computers to make sure we can't just shut them down easily. They could influence us through language to make us do things through psychological manipulation. If we make progress on robotics or if the AI makes progress on robotics, then well, the AI would have boots on the ground to eventually not need us anymore.

So these are scenarios. We don't know if these things are going to happen, but the potential negative consequences of having such a beast out there is so drastic, potentially the end of humanity, that we can't just take it lightly and ignore it and put our heads in the sand. We need to do the research to understand what can go wrong and what sorts of AI could we build now in order to guarantee that even when AI will be smarter than us, it will behave well according to our norms and values.

Rashmi Mohan: Got it. Yeah. No, I think you bring up some very valid points. One is around until the time that we were defining the objectives of what we wanted these systems to do, it still felt like we had some level of say or control, which is now at least fast seeming like is slipping out of say, a human's input or our hands. The second piece is around just whether the data that we're providing these systems itself is, I mean, and this is just in generically I think speaking of AI, is free of the biases that we have.

And then the other thing which I thought was very also interesting as I was thinking about it is I mean, as a lay person, there is an implicit trust in many of these systems because we're not thinking beyond. I mean, the uses of these. I mean, it's one thing to say, "Okay, I use an LLM to do my homework and write my essay." Versus, "I'm using some of these systems to make critical decisions in terms of healthcare or other..." It could have much, much larger implications. And I think as humans we're focused on, okay, it's helping my productivity. It's helping me do higher order tasks. Those also seem like very real concerns because I think the awareness amongst the lay population is really not there around these potential risks.

Yoshua Bengio: Yeah. Let me go back to the question of biases and discrimination. This is one of the earliest subjects of concern from within the AI community, but also the larger community of people in social sciences and humanities regarding the social impact of AI. But it's also interesting because it's an example of this misalignment that I've been talking about, when your cat understands something different from what you intended. So people who design AI systems don't want the system to be racist or gender biased. It just happens that we don't know yet how to build the systems in a way that aligns well enough with our values. And well, in the case of discrimination, it could have impact on individual's lives, especially when these systems are used to decide who gets a job or who goes to jail.

But it could get even worse than that when AI systems have even more control in our society and even more effect, or thinking about the use of AI in the military where it could have even stronger life and death decisions in the hands of AI systems. So this alignment question is really important. The more powerful AI systems are, at the same time, they could be more useful and especially have greater economic value and more dangerous, either in the wrong hands because people will misuse that power. I mean, humans are like that, even if it's a minority. And also this danger of loss of control, that this misalignment could lead to the bear going out of the cage and taking control of the world because that's what's best according to its objective.

Rashmi Mohan: Yeah. No, absolutely. What would you say, Yoshua, or what should we do at this point? I mean, one thing is, hey, what do lay people do? But I would say more importantly, where should we be spending time in terms of research, whether it is computer science research or other areas?

Yoshua Bengio: Well, I believe that there are two things we need to do urgently before we get to AGI. One is scientific and the other is political. So on the scientific side, so that's where computer scientists like myself can mostly contribute. We need to figure out what can go wrong. We need to understand better how these large neural nets work, and we need to find ways to mitigate those potential risks that I've been talking about. The loss of control risks, the misalignment issues, the misuse potential. This is something that has, even right now, national security implications. So there's really urgency, in my opinion, to invest massively. I mean, humanity needs to figure this out before catastrophic outcomes come our way. And we don't know the timeline. It could be a few years, could be a couple of decades, and so there's this really, really this urgency of figuring this out.

But even if we do figure out how to make powerful AI systems that are aligned that do our bidding and don't turn against humans, there's still the political problem. So what do I mean? Well, first we need to make sure that if there is some protocols, some methodology to design AI that's going to be safe, we need to make sure these protocols are followed everywhere on the planet. So we need national regulation, we need international treaties, and we need the means of enforcing those treaties, so as to minimize the probability that someone somewhere will develop an AGI project that could be very dangerous for humanity.

The other issue that's political is that even if we have all these regulations and treaties, there will be people, there will be maybe corporations or governments that won't follow those rules because they want to use AI for their own benefit to accrue more power, more wealth, more economic or military dominance. There is already an arms race at play in AI with the U.S. and China, and it's not going to stop. So this issue of misuse of the technology as the technology becomes more powerful raises all kinds of questions of governance, both locally in each country, but also international governance. So it's not enough to have the rules and the treaties, but those rules and treaties need to consider the potential for humans to abuse the power that AI could give us. And so how do we make sure no single person, no single company, no single government can abuse that power? This is tricky, especially when we're talking about governments. It needs to be, in my opinion, that we converge to multilateral international governance of AGI projects in the future.

Rashmi Mohan: Yeah, absolutely. I mean, the way you would describe it to me, it feels like we definitely need a nonpartisan organization that is overseeing or at least trying to bring together the thought-top minds in this area and start to think about what regulation might look like. It almost feels like this should be a U.N. sustainable development goal that countries adopt and follow. What would be your advice in terms of, say, young upcoming researchers somewhere in the world or anybody who's even interested, say, in policy? And are there avenues for them to contribute, for them to start thinking about this problem to raise awareness?

Yoshua Bengio: Absolutely. Maybe a good analogy here is the potential catastrophic outcome of climate change. First and foremost, we need more global awareness of those risks. We need more people to understand those risks. So just even individually reading up about those potential risks, what are the experts? And people have different opinions, of course, thinking about these risks. What sorts of, if you're a computer scientist, what sorts of exploration and methods are people thinking about? And in terms of activism, because we're not going to do the right thing in terms of investing in the right research, in terms of political efforts, in terms of legislation and treaties, if there are not enough people who take this seriously.

It's just like what happened with climate change. It's only as population became more concerned about climate that are politicians have been starting to take it seriously. So I think something similar could happen, but the only problem is we probably don't have 30 years. The Kyoto Accord was in the late '90s, and we are almost 30 years later, and we still haven't done the right things with climate. We don't have that much time for AI. So this is something that requires as many hands and brains as possible to make sure we steer our societies in a direction that A, protects democracy and the well-being of people, and two, preserves our children's future, humanity for decades and centuries to come.

Rashmi Mohan: Yeah, absolutely. I mean, I do hope that folks that are listening in think about this deeply and start to think about what their contributions might be towards this. For our final bite, I would love to hear from you as to what are you most excited about over, say, of the next five years in this field? What are you hoping to see change?

Yoshua Bengio: Well, one area that hasn't received, in my opinion, enough attention in computer science research related to AI safety. I mean, AI safety didn't get enough attention, but in particular in AI safety, the question is whether we could build algorithms that could give us some sort of guarantees about safety. So right now, we are building the cages for the future bear or the current ones, and we are just going through trial and error. We don't have any quantitative guarantee that things are going to be okay.

If you build a bridge, you have some equations and maybe some model, and you try to come up with an estimated probability that it's not going to fall, and hopefully the regulator is going to ask the bridge builder that this probability is going to be very, very small. Well, there's nothing like that in machine learning for AI, and that is something that some people think is hopeless. I think it is not. I think there are ways to at least come up with probabilistic guarantees using theoretical computer science and machine learning in order to have something we would have more trust into as we move into the territory of AGI and machines that are smarter than us.

Rashmi Mohan: Excellent. Thank you so much, and thank you for your part-breaking work and taking the time to speak with us at ACM ByteCast, Yoshua.

Yoshua Bengio: My pleasure. Thanks for interviewing me.

Rashmi Mohan: ACM ByteCast is a production of the Association for Computing Machinery's Practitioners 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/bytecast. That's learning.acm.org/bytecast.