SDS PODCAST EPISODE 977: ATTENTION, WORLD MODELS AND THE FUTURE OF AI, WITH PROF. KYUNGHYUN CHO

Jon Krohn:	00:00:00	What's going to be the next big step function that blasts us forward in terms of AI capabilities? Welcome to episode number 977 of the SuperDataScience Podcast. I'm your host, Jon Krohn. Today I've got Professor Kyunghyun Cho, who is one of the world's leading AI luminaries. His papers have been cited 200,000 times, and he's behind countless game-changing AI innovations, including co-authoring the first paper on attention. He now leads the Global AI Frontier Lab at NYU, New York University, where he's a professor teaching data science and computer science students. He's one of the most brilliant minds that we could possibly ever get on the show. And so I've been so excited about this episode. I think you're going to love it and you're going to get a great picture into what's next, what's coming in AI. This episode of SuperDataScience is made possible by Anthropic, Cisco, Acceldata, and the Open Data Science Conference.
	00:00:50	Kyunghyun Cho, welcome to the SuperDataScience Podcast. It's an honor to have you on the show. How are you doing today?
Kyunghyun C.:	00:00:55	Well, doing well. Thanks for the invitation. I'm very excited to be here.
Jon Krohn:	00:00:59	Of course. Yeah, you're a big deal. I've been unusually excited about your episode trying to make sure that I have every I dotted, every T crossed and perfect for this episode. We met about a month ago at your Global AI Frontier Lab at NYU. You were launching, it was the launch of your advisory committee that day, I think.
Kyunghyun C.:	00:01:19	Yes, it was. So I'm co-directing this new lab called the Global AI Frontier Lab that was created by the funding from Korean government together with the New York University. And then this lab is there to provide a platform for the international collaboration, in particular between the Korean researchers and researchers in NYU, as well as in New York City. And I'm co-directing it

		together with Ian Lukun now. And then of course, we need to get a lot of, let's say advices as well as some feedback from the worldwide experts. And we decided to create this advisory council that consists of about six to seven people, including the William Falcon, who was a previous guest here at this data science podcast. And then there, we wanted to make sure that it's not only just about the advisors who are going to enjoy talking with the researchers at Global AI Frontier Lab, but also a broader public.
	00:02:07	So together with the Lightning AI and you as well, right? We hosted a happy hour. I think that was a big hit. The people loved it. People talked about it after that over and over.
Jon Krohn:	00:02:17	Yeah, that was a fun night. Met a lot of great people and reconnected with some people that I hadn't seen in years as well. So thank you for doing that. You also have amazing views from that MetroTech Center there in Brooklyn.
Kyunghyun C.:	00:02:26	Absolutely.
Jon Krohn:	00:02:27	Yeah, that was great. Speaking of Will, having been on the show, we did his episode, I think it came out about a month ago. I don't remember the episode number offhand, but he was actually, he was a PhD student of yours, right?
Kyunghyun C.:	00:02:38	Yes, he was. So he was a PhD student of me and Jan together. And then along the way, he realized that there are more important problems than just individual research projects, such as enabling those research projects. And then started to look into the software platform and the stack behind this deep learning research as well as the AI research in general. And then one day he actually came to my office and told me that I think I can make a company out of this. And he actually

wanted to do the company and do his PhD as well. And I had to tell him. So Will, okay, both of these jobs on its own are 200% jobs. So you can't really do both of them simultaneously. That's not good for either of them. So why not just go on a leave of absence for a couple of years, make the company and then make the company in a place where you don't have to spend 200% of your time and I'll be here, Jan will be here, NYU will be here.

00:03:36 You can always come back. I think that was like five years ago now.

Jon Krohn: 00:03:40 Yeah. And now it's wild to see the Lightning AI company that he co-founded based on the PyTorch Lightning framework that he developed in your lab.

Kyunghyun C.: 00:03:46 Absolutely.

Jon Krohn: 00:03:47 Now has over $500 million in ARR. It's an insane success story. And so I think he made the right choice.

Kyunghyun C.: 00:03:54 I think so too. I think so too. Yes.

Jon Krohn: 00:03:57 Yeah. And so speaking of PhD research, you did your PhD research in Finland, which explains your Finnish accent.

Kyunghyun C.: 00:04:04 Yeah. So I absolutely do have it. Yes. Depending on some of the words, I actually do. Oh, really?

Jon Krohn: 00:04:10 Did you learn to speak any Finish while you were

Kyunghyun C.: 00:04:12 ... So because I was a master's student first, it was actually required for us to take the Finish 1A and then finish 1B. And then I could have taken the Finish 2A and 2B as well. But then there was a point at which I realized during Finish 2A, I think, that I was spending way too much time on learning language when I was there to learn about machine learning and artificial intelligence. So it was a bitterswim moment, but I had to kind of give

up taking a learning finish more and then just focusing on machine learning research instead.

Jon Krohn: 00:04:43 Although even your machine learning research focused a lot on natural language.

Kyunghyun C.: 00:04:47 That's true. But that was after I left Finland.

Jon Krohn: 00:04:50 Although I

Kyunghyun C.: 00:04:50 Do attribute to this necessity that I felt to learn a new language that has absolutely no practical value. So there are about six million people in the entire world who speak Finnish. However, of course, living in Helsinki, I had to learn a bit of Finish. But I think that that actually gave me a perspective that there are many different languages in the world. And in fact, this linguistic barrier is the very first barrier that we need to actually overcome. Have we had a technology? Now we have it, right? Back then we didn't, in order to ensure that everyone gets access to the same level of the information as well as the resources that they can use.

Jon Krohn: 00:05:26 Yeah. And so specifically your research after your PhD on language, it was some of the most groundbreaking research that has happened in AI ever. And so specifically, you were working with Yashua Bengio, another name you mentioned Jan the Kun already earlier in the episode whom you're co-directing the Global AI Frontier Lab with. And back at your postdoc at University of Montreal, you were working with Yashua Benjio, another one of these kind of classic names in AI. And you co-authored with him a 2014 paper on neural machine translation and attention, which laid the groundwork for the attention is all you need paper, the transformer architecture, and basically all of the LLM, the frontier AI lab capabilities that we have today all around the world.

Kyunghyun C.: 00:06:10 Oh, okay. Well, thank you very much. I made a bit of exaggeration, but still, yes, I'm going to take it.

Jon Krohn: 00:06:17 Well, a little bit, but I can quantify it as well because that paper has been cited over 40,000 times, which is insane. I mean, if anybody who has 40,000 citations in their entire career is a big deal in that one paper has 40,000. So it's wild. You kind of gave us a little bit of a background there, how the Finnish language experience maybe inspired some of the interest in neural machine translation, but tell us a bit more about that paper and how it came about.

Kyunghyun C.: 00:06:43 Yeah. So let's see. So there are a lot of, let's say, lock involved in that coming up with the paper. I went to Montreal or I moved to Montreal at the beginning of August or the mid-August 2013. And then it was just open office space, so you could see any of the desks that was available. And then the Yash Evangelo's lab was so much smaller. Nowadays, I believe that the Miller has like 1500 or so people, including the PhD students, master students, postdocs and professors. But back then, I believe there were about 35 to 40 people only, counting everyone. So it was very easy to actually talk to everyone else and then trying to have a discussion, brainstorming sessions and whatnot. Then one day, Yasha stopped by at my desk and then started asking me what I want to work on. I told them, "Well, what do you have in your mind?" And then Yasha gave me four options.

00:07:38 Fourth one, I don't even remember, but I'll tell you why I don't remember that. The first option was to just continue to work on what I had been working on back then called the restricted or the deep both machines. I thought, well, I'm coming to Montreal, so maybe might as well I'm going to try something new. So I skipped that one. Second option was to work on something called the generative stochastic networks that were effectively the better versions of the deep and restrictive both permissions. I

thought, well, still sounds pretty much the same, I'm going to skip that one. Third option was machine translation. And I had to actually ask Yasha back a bit, but Yasha, I don't know anything about machine translation. I didn't have any training in natural language processing nor machine translation. And I seriously thought that the Ashore didn't either, nor anyone at the lab.

00:08:27 But then the phrase machine translation really got stuck in my head and I couldn't really shake it off and then I thought, "Okay, maybe it is time to work on machine translation." And I do recall that I did ask Yasha why machine translation at this point. And he's really a visionary, right? Yasha Benji, Yang Lakun, Jeff Hinton, Yogan Schumacher, those are the visionaries who've seen all those things, all the things that are happening now, 30, 40 years ago. And then Yasha told me that based on his experience as well as the conviction, machine translation seems to be the next problem that needs to be tackled. And he was so right, right? He was so right. Now to tackle machine translation or any kind of research problem, what do you need to do? You do extensive literature review. But as Jeff Hinton at one point said, you don't want to read too many papers because often reading a lot of papers gives you a lot of constraints under which you need to operate, or at least a mirage of the constraints that you feel like you have to operate under.

00:09:30 Without realizing that a lot of the research or the reportings from the past papers were actually created under the situation or the context in which those authors were operating. And of course, word changes, situations change, circumstances change, but those changes are not really written down anywhere. So I actually read some of the textbooks and then old papers, but then at the same time, I started to just think about how we can actually build machine translation system from scratch had I had

an opportunity to do so and then I actually had. And then that's how we started to think about how to build this purely neural net-based machine translation system. And at some point we started calling neural machine translation. Now that was going all well, except for the fact that we started with this kind of very simple recurrent neural network that's going to read one word at a time on the source sentence, and then eventually generate one word data time on the target sentence.

00:10:25 So there is a translation. It wasn't easy to train that model. This training, this recurrent neurons just never was easy. It's still difficult nowadays, and it was particularly more difficult. So effectively what we are tackling back then was the problem of underfeeding. So there is a data, and then we know that there is a signal in the data to learn how to do a translation, but somehow we cannot train this very powerful neural network that is able to, in principle, extract out all those signal. Now the issues include those vanish ingradient, lack of computer, the just slow computers and whatnot, but we had to somehow overcome those things. So eventually over time, we started to know how to train these recurrent neural networks by introducing all those shortcut connections or the gated connections that helped dramatically. But even then, we couldn't actually scale it up well.

00:11:18 Recurrents are notoriously difficult to scale due to the kind of sequential nature. It's very inherent. You cannot really avoid that issue. Thereby we got stuck and then at the same time team in Google that was led by Elia Suskaver came up with a very similar idea and then they executed it so much better. They were using 8GPUs. I know it's a tiny number nowadays, but eight GPUs in one machine. And then if I recall correctly, Ilya actually implemented the whole thing himself, the distributed GPU training. I couldn't even imagine how to do that. I wasn't a good software engineer nor had a kind of imagination that Ilia had. So we're using this one GPU here, 8GP

simultaneously, and their execution was so much better. Of course, it worked much better. And then we had to think about what is the right way to scale it up.

00:12:08 Now, there are always three different ways to scale it up, scale any kind of AI system up. One is increase the amount of data. Great.

00:12:17 But then of course, increasing the amount of the data is not what universities are supposed to do as far as I can tell. In particular, if there is a commercial potential, companies should do that. Second is that you increase the amount of compute you use either by making the models larger or just way longer patients, right? Or the third one is to come up with the algorithms such that the same set of the algorithm sort of models are going to benefit better from the same amount of the data. And then thereby we are either increasing the intercept that is a bias toward the up or increase the slope itself. So we had to actually rely a bit on the third aspect and then we thought that's the only aspect that we should actually do as university researchers, but it's not easy, right? The issue is that the collecting more data can be made systematic.

00:13:05 Making things faster and bigger can be done systematic. And in particular, you can put exact amount of the dollar signs next to how much you want to scale up. But trying to change the whole, let's say slope of the scaling law is not something that you can actually do systematically. It requires you to have this kind of aha moment, right? Until then it's like flat and then suddenly there's aha moment and then we actually see the jump. We were struggling, but then Yasha was very excited about machine translation and natural language processing in general. So what he did was that every time there was a master's or the PhD student interns who were visiting University of Montreal, he would just forward them to me because he really had this strong conviction that the

machine translation is going to be solved in this manner. I was like, okay, this is great.

00:13:56 I had a bunch of master student interns who are just trying out all those random things, reading papers, think about how to solve these problems better. And then one day, one of those interns was Dima Badenow. So he's a Badenow of the Badano L, I did a neuro machine translation paper with the attention. He came up with this brilliant idea one day in the morning. So on that day, I came to the office, Dema came to the office, and then I thought I had really an idea overnight. And then Demo was like, "Oh, I think I had some nice idea." I was like, "Okay, Dema, you go ahead." And then Dema described this attention, idea of the attention, the prototype of the attention on the paper, one step at a time. And then about halfway, this was one of those very few, if not the only ideas that I've seen where I could sense immediately that it was going to work.

00:14:49 It was going to work. Just made sense. It was so brilliant and idea. And on the other hand, my idea back then was so primitive to the point that I didn't even actually say out loud what that idea was. And then DEMA quickly implemented it because DEMA is a super engineer as well, software engineer as well. And then we noticed after three to four days that indeed it was working as DEMA anticipated and we could all see as we were listening to DEMA's description of the algorithm itself. So that was really interesting observation. And then that was also another point, in hindsight, where I think Yasha's kind of as a vision really shined because what was happening is that, of course, already it was a 2013, 14, we knew that if we make a very deep convolution on neural network, we could actually solve the problem of the object recognition in an image.

00:15:40 And then if we make, let's say, very deep again, neural network and then put the HMM on top of that, we knew

how to solve the tackle the problem of the speech recognition. But because now we are working with the natural language, we were now working with a very different set of, let's say, parameters of the constraints than before. What are those? The sequences can vary a lot in their length. So we had to come up with the algorithm that is able to work with this kind of variable length sequences. And then because of that, the issue of the vanishing gradient was made dramatically worse.

Jon Krohn: 00:16:13 Right. And I'll just quickly explain that for our listeners who aren't aware, who haven't been building neural networks themselves, is that as you layer ... So you get more ability to extract features and have more complex, more abstract features be represented as you add more and more layers of neurons into an artificial neural network. But the downside of this is that training really slows down because the further you get away from the loss function, from whatever you're optimizing with this algorithm, the signal becomes lower and lower and lower. And as you add more and more layers, as you get further and further away from that loss function. And so you end up with these vanishing gradients. These gradients of learning become very, very small in these layers that are far away from the whatever you're optimizing and yeah.

Kyunghyun C.: 00:17:02 Yep. So those issues were all there because we are working with machine translation. And then I still think about why Yashra brought up machine translation back then. Then perhaps he knew about all those issues that could be tackled how we worked on machine translation. So somehow DEMA together with me as well as Yashua, we were able to kind of crack this to a certain degree. Of course, there was a very oddly version there.

Jon Krohn: 00:17:28 So you're saying that the machine translation research, the Yashua had this hunch on this intuition that was really important. It helped you crack this vanishing ingradient problem with using attention, which allowed

you to also tackle these other areas like machine vision that were struggling from the same vanishing gradient problems.

Kyunghyun C.: 00:17:48 Absolutely, absolutely. So there was a key issue that had to be tackled, and somehow Yasha was able to suggest this particular problem that exhibits all those potential causes behind this issue. Then in order to solve this problem well, we had to solve all those issues as well by going into the root cause of it. That was a vanishing gradient. And then somehow it kind of worked. I don't know how he actually foresaw that that was going to happen. So that was a really interesting time. And then there was one really interesting moment, I don't think I've ever talked about it. So one of the interns back then, who became a PhD student is a Bart Van Marion War. So he was initially an intern. I think he eventually became a PhD student and then finished his PhD at University of Montreal. He's probably working at Google now.

00:18:37 So he's brilliant. He's one of the smartest people I've ever worked with. And then one day, Bart just designed the web app that's going to demo this neural machine translation system with attention. It was a really nice demo. You type the sentence and in French often, because we're in Quebec. And then it's going to generate the English translation using the underlying neuromachine translation system we already trained and it's going to extract out the attention pattern and then does amazingly beautiful visualization of how all those words are related to each other according to attention mechanics. It was amazing. It was beautiful. But then with the demo, demo ... So especially in computer science, demo is the biggest, most important thing in my opinion, because that gives us a way to play around with the systems that we build or the algorithms that we make, and then thereby we get more creative uses out of it, as well as the more intuition into how these algorithms work.

	00:19:34	So just for fun, back then, one issue was that we couldn't actually increase the size of the vocabulary. Nowadays, everyone talk about, oh, word segmentation or the bite pairing codings or whatnot, but because we didn't have any training in natural language processing, we were just using a white space based SAI segmentation. And
Jon Krohn:	00:19:51	You'd even specifically avoided the literature for the most part.
Kyunghyun C.:	00:19:54	Yeah, exactly. We did too many words, right? So the funny thing is that because of that, we had to use only the top, let's say 50,000 words only, and the rest of the words were mapped to this unknown token. So we used the capital UNC and then we call it like on tokens or whatnot.
	00:20:12	And then what that also means is that in the sentence you write, you can arbitrarily introduce untoken in the middle and then ask the model to translate this sentence with this unknown tokens or the unknown words. But when you let the model translate, you can force it not to produce unknown tokens. And then what it does is that by doing a translation, it kind of fills in what that unknown word should be. The model needs to figure that out. And then the example I used was that the unknown Korea is a friend of the United States and then unknown Korea is an enemy of United States. And then when those sentences are translated, the first one was translated to the South Korea, and the second one was translated to North Korea. And then that's a kind of time at which I saw that, oh, these models are really powerful, not only because they know how to translate, but because it actually knows about the knowledge that is embedded in the large amount of the data we have.
Jon Krohn:	00:21:17	And this reminds me very much of a century ago, there was this really famous philosopher, Wittgenschein, Lubig Vitkenshein, and he had this idea that a given word, so like this unknown token, it tends to be the average of the

words around it. And so you could see how those words like friend or enemy, it would skew kind of the average in some vector space towards Southern North in

Kyunghyun C.: 00:21:47 This case. Absolutely. Absolutely. And then these models were just capturing all those regularities in the large amount of data. I mean, relative to what kind of data we are using, small, but it wasn't that small. It was pretty large because we're using data that was released from the so- called workshop on machine translation. So WMT and they would release all the news articles as well as the all the parliamentary proceedings from the EU parliament because they need to be translated always into the multiple member languages, as well as the Canadian parliament because everything has to be translated into English and French. So it was a pretty large dataset that has a lot of information about geopolitical, let's say, relationships. And then this model was able to just get it automatically.

Jon Krohn: 00:22:32 That's amazing. Thank you for the story behind that. And you tell a story so well, but it's really enjoyable to listen to. Not too long after that NMT paper. So in 2018, you said that our brains are just biologically implemented computers, which I think is obvious that is what they do. And you described AI progress as a very, very small step toward true intelligence. Looking back now with the rise of these kinds of approaches like attention that you helped develop and these generative AI advances that we've seen in recent years, do you think we're just making small incremental steps or do you think things have moved a lot more quickly than you would've expected?

Kyunghyun C.: 00:23:12 It is definitely moving very fast and I'm thoroughly enjoying it. It's great. So the one thing that we are learning over and over is that when we try to solve any problem that resembles or looks like it requires any kind of intelligence or anything that looks like intelligence, it's always easier and then easier to look at it from the

information abstraction. So rather than thinking about, for instance, biology as the, here's the, let's say molecular dynamics happening that is telling us about how the atoms interact with each other, and then based on that one, we get the actual molecules, molecules interact with each other, we combine some of those molecules, we get our large molecules, and then very large molecules include those DNA based pairs and whatnot, then they are going to encode some of the genetic materials, and then they're going to produce the proteins and on, and we have a cell and so on.

00:24:10 Now, it's really difficult to tackle this kind of the molecular biology. If we try to really view it as a, here are the atoms, here are the molecules, here are the intractions of molecules, and here are the cells and whatnot. And then, of course, the interaction of the cells is us, right? It's too difficult because there are so many details that are probably the implementation details. However, there is a very clear information flow that actually happens. So there is information that is flowing via, let's say, genetic materials through the generations, and then those genetic information or information included in the genes will be actually transformed into the proteins. Then these proteins are carrying the information and then by combining orient to canceling with each other by having an interaction with the other proteins, the new information is created or the old information gets deleted, and then that happens at every layer.

00:25:01 Then suddenly everything becomes easy to tackle because we have been thinking about how to process information over and over throughout the past say centuries. So in that sense, is human brain, let's say, a computer, not in a traditional manner, but does it actually process information that we receive via the sensory organs and then send the information to our motor control systems? Absolutely, yes. By having this kind of level of the

abstraction, we can now understand much better what is happening and then what are needed in order to build this kind of system. And then what I think we have done is that they at least partially have figured out that right level of the abstraction when it comes to capturing all the correlations that exist in a massive amount of data in the most efficient way possible. And then thereby, more data we give, more correlations to be captured and then thereby gets better.

00:25:54 Of course, correlation is always just one side of the coin though.

Jon Krohn: 00:25:57 Yeah. We are unquestionably this biological machine that's processing information, lots of interesting ways that we try to adapt what we're doing in biological research into AI and make advances. And I remember a question that I had for you, we had to take a cut in recording because I was like, "Oh, I had this brilliant question to ask you. " So back, you were talking at the beginning of the episode about that happy hour and the event that we had around the advisory committee launch for your global AI frontier lab. You had a talk with Will Falcone there about where you see research going and what's really important. And one of the key areas that you discussed was sample efficiency. So the algorithms today, the transformer architecture that you helped pioneer, that it's a very sample inefficient learning mechanism the way we do today with whether it's the supervised learning, pre-training, or the post-training with reinforcement learning.

00:27:03 Either way, we end up needing huge numbers of examples, millions of examples in order to be able to learn, but in biological systems, in small children or humans of any age really, even a lot of different kinds of animals, you teach a dog a trick, some breeds of dogs are smarter than others and like to be trained more than others, but the ones that are designed for farming and

hunting, these kinds of things, those kinds of dogs, they do single shot learning. Yeah, exactly. Whereas for even some very simple conceptual problems, machine learning algorithms might need millions of examples. And so you talked already in this episode about how you can kind of, when you're trying to scale capabilities in AI, two of them are very easy to model because they follow cost curves and scaling laws. And so that's the amount of data that you have, the amount of compute that you have.

00:28:00 And so those kind of follow some kind of linear curve, but you talked about this kind of stepwise function of just having a new idea like an attention mechanism, for example. And so it's obviously a very tough question, but now it's almost a decade since Yashua Benjio came to your desk and said, "I think we need to focus on neural machine translation." You today now have this very broad perspective on where things are going in AI research. And so what kinds of problems do you think we can solve to try to identify this new big step function and be able to have machine learning algorithms start to approach the sample efficiency of learning that biological organisms have?

Kyunghyun C.: 00:28:43 Yeah, that's a great question. And this is something that I think a lot about. And in fact, any machine learning researchers should think a lot about sample efficiency, but I believe that there is a bit of a incorrect kind of perception of what sample efficiency is. So when we think about the sample efficiency, we often go to this mode where we say that, "Well, somebody gave us this small amount of data. Can I actually learn as well as a case where we had a large amount of data?" But that's actually often impossible. Now, this question only makes sense if you're working with the extremely bad algorithm. So my algorithm is so bad to the point that it cannot actually capture all the correlations that exist within the data, then I can think about, can I come up with a better

algorithm that's going to capture more of the correlation from the same amount of data?

00:29:40 But nowadays with these large scale transformers and amazing learning algorithms that people come up with, given any amount of the data, these models and algorithms will be able to capture most of the correlations that exist in the data. And then we can even measure how much we are getting close to the optimal, let's say, site by looking at the sole Called perplexity. Perplexity really tells about how confused this model is in predicting the next word. And then the level of the confusion is measured based on how much of the correlations the model were able to capture. And as it goes down and now, we know that most of the correlations are being captured. And then we know that those validation perplexity disease are extremely low. And they are really low to the point that most of the times the next word is determined fully. Now what that means is that it really has captured all the correlations that exist.

00:30:32 Now in this kind of scenario, we cannot really talk about sample efficiency when the data is given to us passively. We have to now start thinking about the problem that is more active. That is, if I had the same budget to collect the same amount of data, can I choose which data to collect so that I can actually learn the more important correlations more quickly? That's how we change the problem. And then that's the kind of sample efficiency that we need to think about. And then that actually connects to our conversation back then together with the million Falcon is that what we need to work on is not how to squeeze out more out of this data that was passively collected without the purpose in their mind, whoever that collected the data. But rather we need to think about how to actively collect data so that we squeeze out more from the universe or the word.

	00:31:26	And then that's where the aha moment comes in. Aha moment is effectively a very rare event. So if somebody just blindly collect the data, collect the samples, then chance of this aha moment happening is vanishingly small. But then the question now is more, can we actually make a machine learning algorithm or the AI algorithm that is able to intentionally make those aha moments or the rare events more frequent by actively looking for them? Then suddenly those aha moments are going to be more frequent and perhaps be even made systematic. So that's what I think we need to all work on. This kind of learning algorithms that are able to actively find and mine the data that is important rather than just consuming the data that is thrown at it without any purpose.
Jon Krohn:	00:32:15	Right. Does this start to relate to the idea of world models?
Kyunghyun C.:	00:32:20	There is the aspect to it. So now of course, collecting the data that is important cannot be done in vacuum because how do we know what is important? So it is really important to have these models that have some good sense of what the word is and how the world works. And then thereby contrasting that, that kind of background knowledge against what it runs into, thereby it can actually collect more focus as they are and the more useful data down the road. So in a sense, yes. And then that's the reason why many of these RL tuning these days, all these large scale language models work only after we actually train it somewhat with a massive amount of data, most of which have absolutely nothing to do with the actual problem for which we are tuning these algorithm models with reinforcement learning algorithms. So these are all connected there.
Jon Krohn:	00:33:15	Right. So these world models, and I'm certainly no expert at it, so I'm probably going to butcher this to some extent, but you talking there about how collecting data that are unrelated to the problem that you're solving, it provides

the model with some kind of general understanding of how the world works. So for example, there were papers a couple years ago that made a splash where you could say improve performance on a neural machine translation task by having a vision algorithm trained as part of the same model. And the idea there is that this gives, in the latent space of this algorithm, it starts to develop a more nuanced understanding of how the whole world works because vision gives a more complete picture, an understanding of language. And it applies the other way as well. So that if you are, say, you want a projectile, somebody kicking a football, somebody kicking a soccer ball in a generated video, that soccer ball should follow the laws of physics and keep moving in the same direction through the air.

00:34:18 It shouldn't change around or shouldn't suddenly appear in a frame suddenly, somewhere else in the shot. And you could see how potentially having a big understanding of all of the language that's ever been written on the internet gives a sense of these kinds of physical laws. And so that language understanding allows this video generation model to be able to perform better and understand real world physics.

Kyunghyun C.: 00:34:42 Yeah, absolutely. And then that's a really interesting example because until about let's say 2014 or so, most of research, although some people had been working on that already, such as Mag Mitchell, as was LP Morency and CMU, they were working on this kind of multimodal learning for many years and then there are many others. However, it was more of a niche topic. More people were working on using machine learning algorithms or developing machine learning algorithms for individual problems at a time or the individual modalities at a time. Now, what we have learned, of course, over the past, I say, decade or so, is that if we add in all those different data of the different types, these models actually can capture the correlations that exist across the data type

boundaries and then being able to benefit from having seemingly unrelated or the irrelevant let's say modalities as well.

00:35:35 And then what that also means is that then perhaps we can now think about building a machines that are not only going to look at a particular modality, but is able to really mine or to go out there and then capture data of different types as well. And then how we know how to do so is by seeing many of the already collected data. So yeah, this is all connected in a sense that in order to be active, which we need our algorithms to be when it comes to data collection, these models should have already been trained massively on a passive data. And of course, collecting all those data for an active manner, in fact, researching the data that has been collected and that are irrelevant to all the future tasks as well. But then they all become, in some sense, passively collected data for the future task.

00:36:22 And then we actually create this kind of loop. We started from somewhere. However, what happens is that we create more data that is very purposeful, and then the data goes back into this large pile of the unpurposed for data, but this actually forms the basis on top of which we can now collect more data for the future tasks as well. So can we actually close this loop and can we do it in a systematic way? I think that that's the key here.

Jon Krohn: 00:36:47 Really exciting. When you talk about actively mining data, I suppose this is like an agent searching the web for relevant information or in the not too distant future, perhaps a physical embodiment, like a robotic embodiment, being able to explore the physical space and maybe even interact with objects in its space.

Kyunghyun C.: 00:37:04 Absolutely. In fact, all of those included. In some sense, they are not that different from each other. Say you want to run some kind of actual experiments. You want to test

whether a particular protein is going to bind to yet another protein that we want to design a drug for. Now, one way you can imagine is that, well, I'm going to try all possible proteins and then see if any one of them bind to this particular target protein. Of course, all possible proteins doesn't make any sense. There are too many of them, right? We don't even know what kind of proteins exist that are stable and can be synthesized within the cells or whatnot. So we cannot simply say that we will try all possible proteins. So then what we do, we're going to use a very smart algorithm to pick only small number of the proteins that are time to test.

00:37:51 And then based on the feedback, we're going to continue to choose the next batch of the proteins so that we can actually find a good binder way earlier than trying out all those as all possible proteins. And then this is a sample efficiency we gain, right? And then it's actually the same with the search as well. And other idea, of course, we think of a search as a completely software based implementation or the system. At the end of the day, we use Google search. We are embodied, right? So we are a physical robot ourselves. And then the same thing. Internet has so many documents. One way to figure out which documents are relevant for my question is to read each and every one of them, and then trying to decide which subset matter. But again, there are too many of them. So what we want to do is that if we want to interact with the search engine to find or figure out how to find the small subset of time, and then after a couple of the rounds, I'll be sure that I have found the relevant documents.

00:38:46 So these things are all the same thing. Now, of course, when it comes to physical robots as well as the physical experimentations, there are a lot of issues, not necessarily because of the AI algorithms, but because physical words are much more fragile than software word.

Jon Krohn: 00:39:04 Yeah. You have to be careful. You can be tricky, for example, to pick a grapes without shattering them for a robot to be able to do. Exactly. But it's something that robots are starting to figure out more and more. It's an area that I'm personally excited about a lot now co-supervising this AI robotics research at the University of Auckland, and we'll see hopefully some exciting things come out of that. Definitely out of the research in general, exciting things will come out. I know Jan Lacune is bullish on world models. He's got his advanced machine intelligence startup that he has going. And even though this is a separate effort, you two co-authored a 2015 paper last year that won the best paper award at ICML in the workshop on physically plausible world models. This paper was called Planning with Latent Dynamics Models. And we'll have that in the show notes, of course, for listeners to check out.

00:39:55 But does that tie into the conversation that we were just having, that research?

Kyunghyun C.: 00:39:58 Yeah. And then it actually ties into all the things that we have talked about a bit, like the information processing and whatnot, is that the ... So what does it mean for us to know about the word? Because we talked about the word model, and then it turned out that alone is a very big question. Some people believe that if I can imagine what's going to happen in the future at a very high, let's say quality or the high fidelity, then I may be able to say that I understand about the word. Some people say that that's not true. We actually don't have a good picture of the imagination that is high fidelity. We have a very high level concepts that are extracted. We know how those concepts interact with each other, and then that's probably enough. But of course, who's to say which one is better?

00:40:47 But nice thing about this large scale machine learning or the AI era now is that in many cases we can test them. So we all have heard about and then maybe tested those

OpenAI SOR models or the Google's Genie models and all those video generation models. I think the runway has one. These are amazing models. And then some people who are building these models in their mind have this thought that by building this kind of video generation model or the action video generation model, maybe we can build a machine learning model or the AI system that understands the word. And then using that kind of imagination to plan out what's going to happen. But then on the other hand, some might think, and then it's very natural for us to question whether that is necessary. Say I want to travel to Paris. I'm not going to plan out every single step and then trying to imagine every step I actually take from here all the way to JFK and then walking toward the gate, walking to my seat, having a sit.

00:41:50 I don't really care about that. What I'm going to imagine is that they say I already made it to Paris. What am I going to do? What is going to the first restaurant that I'm going to go into and then trying to order some wine and whatnot. But I don't really need to imagine all those steps in between. So then one might say that actually what we need is a very high level picture of how the world works and then be able to jump back and forth. And then that's what we meant by the latent dynamics. And then that's the kind of foundation on top of which what Ian has been calling as a JAPA, that is a joint embedding predictive architecture. And then also this actually idea, of course, goes back decades. Many people in neuroscience have been thinking about it and control theory has been all about, can we find the abstraction or the small number of the knobs that actually matter in controlling any of the systems?

00:42:42 And then we work at that level. Then we're going to project it down eventually. So which one is right way to go about? It's unclear in my view, because the nice thing about predicting every step of the way makes computation extremely regular. What that means is that

it's very good with the current digital computers to implement and scale up. On the other hand, just intuitively saying, we don't do this kind of a say step by step, let's say imagination. So there is a kind of hope that maybe we really don't need to do that because we have a proof of concept here is that yes, we can skip all those steps. Which one will be right? It's unclear, but something will have to be done in this latent space rather than the pixel space.

Jon Krohn: 00:43:26 Right. Yeah. Sounds like that could be one of the big step function changes in how machines learn coming up.

Kyunghyun C.: 00:43:32 I think so, but I have been wrong many times myself.

Jon Krohn: 00:43:37 Yeah. When you're working at the frontier, I think you probably have to be wrong most of the time. That's true. So we now have been talking about that paper with Yonika and earlier in the episode, we talked about how you co-founded this Global AI Frontier Lab that is at NYU with the Metro Tech Center. And you can tell us a bit more about the Global AI Frontier Lab if you want to. You told us a bit already how it's co-funded by the Korean government, that kind of thing. But I want to give you the opportunity to chat about it a bit more specifically.

Kyunghyun C.: 00:44:03 Yeah. So I was born and raised in Korea. I did my undergrad in Korea at Kaist, and then I went to Finland, did my master's and PhD there, and then I went to Montreal, and then now I'm in New York City. And then throughout my career, I've seen a lot of different modes of research. So if you do research in Korea, you grind a lot. Korean's grind always a lot, grind a lot, but then the mode of research is to grind a lot, but often in a very isolated manner. You choose a problem, then you set a deadline, you grind it until you make it until the deadline. And then it's very isolated up there. In Finland, it's actually, there's almost no grinding there. It's like even as a PhD student, I was at the office by 8:30 in the morning,

and they would leave the office by 5:00 PM or the 5:30 PM, and then just a half an hour of the lunch, then work during that time, and then go home, enjoy the rest of you at the day.

00:45:00 So there was a good work-life balance. However, there as well, it's actually thereby it was also somewhat isolated after the research. You spend exact amount of time at the office, yourself, and then you do research at your own pace. Now, Montreal was somewhat different, and then New York, of course, very, very different. And then all my friends that I talked to at let's say Silicon Valley, very different. And then what is a major difference here is that the research as well as development are extremely collective affair. And then people are very good at working with others from all over the world. And I think that that's actually the superpower we have in North America, in particular in US, is that because everyone from everywhere in the world comes here, and then thereby almost automatically we encourage people to work together with each other, and then that actually makes them work with the people even abroad as well, very naturally.

00:45:52 And then I wanted to make sure that this difference is going to be narrowed down. I want the people outside US, outside North America to be as connected internationally as they are to the people here in US and Canada. So we started to talk with the Korean government. And of course, being in New York helps because the New York City is not only about technology, but it's all about everything. Everything from diplomacy, politics, economics to culture. So Korean government got interested in learning more about how NYU is working and then how being in New York City helps. And then I was, of course, brought in because I'm one of the very few Korean American professors at NYU as well. And then with this kind of conversation starting, let's a few years

back, we started to realize that there is a unique, let's say, type of the collaboration that can happen.

00:46:44 NYU is going to provide both the research topics as well as a research platform for the Korean researchers to come and enjoy and then learn about this different way or the different mode of doing research. And then Korea is going to provide a bit of funding as well as the brilliant students and the researchers who are in Korea. And then there is a win-win situation here. So we created a Global AI Frontier Lab. We have a massive space. You saw that a month ago in downtown Brooklyn with amazing view, but that space is created not only to accommodate the NYU researchers, but mostly to host the researchers such as PhD students and professors from Korea to stay here and then do some research together. And in doing so, we made sure that all these visits are going to be long enough because we want it to be the real experience and the real collaborative research.

00:47:42 So each and every visit is at least three months, sometimes up to 12 months at a time. And then these people are from seven institutions in Korea, not just one institution, but we really wanted to make sure that as many people as possible with a diverse set of background and topics or the expertise will enjoy this opportunity. So that's how we created it. I think we are actually demonstrating to the word that this kind of collaboration is possible and there is an amazing opportunity not only for the US institutions, but the institutions all over the world.

Jon Krohn: 00:48:16 Really cool. It's a nice initiative and I can't understate how beautiful the space is and the views that you have. And it sounds like, so maybe if we have listeners in Korea, they can actually, they can go and apply to be part of this program, right?

Kyunghyun C.: 00:48:30 Absolutely. Especially if they're part of those seven institutions in Korea, that is actually quite massive, let's say coverage, then they can always go talk to their supervisors or the program managers and then see if there's opportunity to come to NYU, Global AI Frontier Lab. And also for the students who may be listening from NYU as well, if their supervisors are involved in Global Frontier Lab, which is likely because there are about 11 PIs from NYU alone who are involved in this one, we can also facilitate them spending time for research collaboration in Korea

Jon Krohn: 00:49:03 As well. Ah, really cool. Nice. I think, I don't know for sure, but I do suspect that we probably have more listeners from NYU than we do. Potentially. Yes, that's

Kyunghyun C.: 00:49:14 True.

Jon Krohn: 00:49:14 We skew very heavily to a US audience and particularly a New York audience. So yeah, that'll be interesting to see what comes out of that. Speaking of the Global AI Frontier Lab and you're doing this research at NYU, you also teach at NYU. And so we've been speaking so far on the episode almost entirely about your research at the frontier, very exciting stuff. But I wanted to take a moment here to change gears and talk about the experience of teaching. So prior to us recording, you were talking about how you're teaching an undergrad level machine learning course this semester, and you're encouraging students to be using code generation tools. Tell us about this curriculum.

Kyunghyun C.: 00:49:52 Yes. So the course is titled Fundamentals of Machine Learning. So last time I taught this course, it was pre-ChatGPT. It was pre-pandemic. In fact, it was the spring 2020. So in the middle of the course, there was a disruption due to the COVID-19 related lockdowns. And then even the title of the course was different. It was called the Introduction to Machine Learning. But so a lot

of things have changed. And then I was asked to teach this course this semester because one of the professors who was going to join NYU last September and then was planning to teach this course, could not join us last September. He's going to join us this coming September due to the immigration related issues. So hopefully the administration actually listens to this and then you had to make everything a bit easier because we are losing a lot of, say, amazing talents that we should have gotten earlier.

00:50:46 But anyway, so department asked me to teach this course. I was like, sure, let's do this. But last time I taught, nothing was like this. Everything changes since then. So I've spent some time trying to think about what is the right way to teach these students or more like, what are the things that I have to make sure that the students experience before they graduate and then go on the job markets. So this is for the undergrad course, mostly for seniors and juniors, mostly seniors, some juniors as well. So I decided to just lean heavily, if not entirely on using coding agents or the vibe coding. What do I mean by that? I want to make sure that these students make all the mistakes they want and that they can using these coding agents while they are students because that's our job as an educational institution is to ensure that the students are going to learn things, but also able to exercise what they are learning within this boundary of the school.

00:51:50 And then if they make mistakes, we'll be able to correct their mistakes, thereby they're not going to make mistakes on the job. That's the worst case scenario, right? So I decided that, okay, so anyway coding agents are here, they're not going to go away. In fact, they are going to be everywhere. So I decided to teach them how to use it as much as I can. And I somehow made a one mistake assuming that these students already are using these coding agents. And in some sense, there was a fair

assumption in my opinion, because these are computer science majors and they're juniors and seniors. So in my mind, they're using these coding agents every day, but that wasn't the case. In fact, about 80% of the students out of the 200 did not even have any one of these coding ID or coding agent ID or the coding agent installed on their laptops.

Jon Krohn: 00:52:41 Wow.

Kyunghyun C.: 00:52:42 So we spent couple of weeks while teaching them the basics with the TAs, how to install these to begin with.

Jon Krohn: 00:52:48 Wow. And then

Kyunghyun C.: 00:52:49 Trying to provide them instruction on how to get the free access to many of these coding agents because they're students, thanks to Google and Microsoft for providing the free access, by the way. And Swapik, there's no free access even for students. I'm a bit sad about it.

Jon Krohn: 00:53:03 Something to work on

Kyunghyun C.: 00:53:04 For them. Yeah, I think so. Yes. So we taught them how to install all those things. And then of course, even then there are some troubles because a lot of these tools are designed for the majority software engineers. What that means is that they work amazingly well on Linux machines. They work amazingly well on Macs. They work reasonably okay on the very reason when those laptops, but many of these students actually don't have any one of those, but some laptop with the old versions of the Windows

Jon Krohn: 00:53:34 Right now. Really? I

Kyunghyun C.: 00:53:35 Know.

Jon Krohn: 00:53:36 Modern computer science students.

Kyunghyun C.: 00:53:37 Absolutely. Yeah.

Jon Krohn: 00:53:38 It turned out that was the

Kyunghyun C.: 00:53:39 Case.

Jon Krohn: 00:53:40 I think they would all be on Unix systems.

Kyunghyun C.: 00:53:41 That's what I thought as well, but apparently that wasn't the case, and then that's where the Lightning AI really helped. So I told them, just at least go to Lightning.ai, create a studio, there's a free credit, 15 credits every month, and also CPU machine one machine, all free, so please set it up. So thereby, finally, we got all 200 students set up to use these coding agents. And then what I now do is that every week, on Monday, I spend an hour or so teaching them about one algorithm or one or two algorithms under one umbrella. And then after the lecture, I come back to my office and I'm going to start vibe coding one full web app that uses that algorithm using a new data that I choose online. And then I'm going to send the link to the data and then show them the screenshot on our course webpage.

00:54:34 And then on Wednesday, I go in and I spend an hour going through the entire transaction log that I have used to work with the coding agents and then show them the code. And then I ask them to open their laptop after about 30 to 40 minutes and then start implementing it. Now initially students were very confused. Actually, some of them did not even bring their laptops to the lecture. So I told them to please bring your laptops. And then now, starting from about last week and this week, I do see that the students are actually implementing it very rapidly, taking a screenshot, share their screenshot on the course webpage so that everyone can see what they are building. And then it's very amazing in a sense that you can see how much better they get every week because they get used to it and because it's fun and because it's not just

filling in this tiny box out of the gigantic software that they need to do now, but they can actually build everything from scratch.

00:55:33 So it's been about a third to a halfway into the semester. So I might actually call you up after two to three months and then tell you how disastrous whole thing was. But I think we are making some new effort at how to teach computer science majors in particular when it comes to these kind of applied subjects such as machine learning and AI.

Jon Krohn: 00:55:55 I love it. And there's a really encouraging story here, I bet for a lot of listeners, including for myself, which is that when you see what people are doing, people post on social media, how much impact they're getting by using these kinds of code generation tools. You hear today about these ideas of a 100x engineer, these kinds of things, individuals like Peter Steinberger who are supposedly committing millions of lines of code every month. And when you hear about those kinds of things, it's easy to feel very left behind. And so it's encouraging to hear that students, young students who are probably going to be some of the first adopters at one of the top institutions in the world to be learning machine learning and computer science for 80% of them to not be using these kinds of code gen tools is really encouraging for the rest of us.

Kyunghyun C.: 00:56:41 Exactly. And also, and then thereby we need to fix this situation. We want everyone to have access to this amazing tool that they're going to benefit from, that we want to somewhat train them and prepare them. So it's just a very small first step, but I think over time, higher education as well as a K to 12 education will change forever. Then we'll have to change it quickly enough so that not only a small number of people, those 100X engineers benefit from this kind of coding agents and then AI technology, but 99% who are not the 100X engineers also benefit from this tool.

Jon Krohn: 00:57:18 For sure. And we actually, we do have an upcoming episode. It's going to be our first ... We're almost at a thousand episodes of this show, and we're going to have our first episode ever with a K to 12 school principal.

00:57:30 And she's at an institution in Boston where they're being very progressive about the adoption of AI tools. And so we're going to have a really interesting episode about how a younger education will be transformed by this as well. So look out for that in a few weeks, listeners. I don't know the episode number yet. We haven't recorded it yet, but we are scheduled to record it already. I want to go back to one last piece of research before we start to wrap up the episode, because I think this is another really important area that you had an impact on. I also want to make sure that I got this right. We were in a lot of this episode talking about what you were doing kind of in 2014 with the Asho Benjio, with the University of Montreal, and then I got talking about your more recent work, and I feel like I might've had a brain fart, I'm not sure, but I was talking about this best paper award that you won at ICML, and I think I might've said 2015, but it was 2025.

00:58:19 It was last year.

Kyunghyun C.: 00:58:20 Time flies.

Jon Krohn: 00:58:21 Yeah, exactly. And so when you see that in the show notes, it's actually that 2025 paper. But going back closer to 2015, you had a 2019 paper on reranking and multi-stage ranking. And this laid the groundwork for an approach that has become ubiquitous today, retrieval augmented generation, RAG, which allows us to search over a large number of documents quickly, pull out the relevant information, and then generate a response. Tell us about reranking and multi-stage ranking that led to RAG.

Kyunghyun C.: 00:58:57 Right. This also has a kind of interesting story behind it. So one of my very first, if not the first PhD students is named Rodrigo Noguera. He was already a PhD student at NYU when I joined NYU, and then he wasn't entirely sure what he wanted to work on. He dabbled a bit on visualization research. He dabbled a bit on the computer vision research, but then he wanted to actually talk with me about the natural language processing as well as the machine translation research. We had chat, and then over the first year or so, it was more of a brainstorming, a lot of brainstorming, a lot of literature review, but then at some point it clicked for Rodrigo. Rodrigo was actually very sure in 2016 that we are going to have an AI scientist or the AI researcher. Then this autonomous researcher will be able to do the literature review automatically for us, and then based on the collected information, we'll be able to answer any of the questions.

00:59:59 And then that's the reason Reason why the first project that we worked on together was what we call as a web navigation or web nav. We used a Wikipedia in order to create this kind of, let's say, network of the webpages. And then we train the reinforcement learning based NLP agent to read one, let's say, a webpage at a time, read the question, trying to decide which hyperlink to follow, and then just do it over time until the point at which he finds the page that has the answer to the question. And then he wanted to push it forward, but then that's when he realized that this kind of say direct navigation is just not scalable. There's too many webpages. And then trying to find the relevant webpage from this web of webpages, just not a good idea. We don't do that either because it's so unscalable.

01:00:47 So what he decided to do is in order to get to the ultimate goal, he decided that I have to work on retrieval, information retrieval. And he looked into the information retrieval really, really carefully. Did a lot of the literature review and then realized that there are just tons of things

that can be done better with latest technologies and the technologies that he was going to develop over the next few years. So first he decided working on using reinforcement learning in order to reformulate the query automatically to maximize the record rate of the relevant document. And then after that, he realized that, well, changing the query alone is not enough. We need to be able to tell whether a retrieved document is really good. And then there are many different techniques to do so, but Rodrigo was, "Well, we have an amazing neural network and it was a time of birth, the mask language model.

01:01:41 We probably should use it. These models were trained on massive amount of text." So he knows a lot about the contents and the semantics of this text. So he actually took the bird, fine-tuned this model to do the reranking and it just blew everything away. It was amazingly good. And it was very efficient as well. It's a surprising thing. A lot of people associates the neuron that at least back then with the accurate but inefficient. And many of the traditional method or the sparse that say a vector method as the inaccurate but super efficient. But in fact, when you implement it correctly and then use the right hardware, that's not necessarily the case. This neural nets are extremely extremely efficient as well. So from there on, he's been pushing this direction of retrieval. Then he went to do his postdoc in Universal Waterloo in Canada together with Jimmy Lin, who is one of the world class researcher in information retrieval.

01:02:37 And then now he's actually back in Brazil continuing this direction, trying to build this kind of autonomous researcher that knows how to use the retrieval engine and be able to answer these questions. He's now a professor and also a founder in Brazil.

Jon Krohn: 01:02:51 Yeah. There's lots of exciting things to be done there. No question. There are lots of parts of any workflow,

including a research workflow that can be accelerated with AI. And one of these things, like these kinds of ideas of being able to search over a very large number of documents, glean insights, you could have machines which can have in there, they can have representations from every field. And so you can kind of cross pollinate ideas from different areas of research that any individual human couldn't ever imagine to be able to do.

Kyunghyun C.: 01:03:24 Absolutely. It's so fascinating. In fact, one of the reasons why I think it's fascinating is exactly what you just pointed out is that not a single person, not a single individual can actually have expertise in more than one areas. Now what that means is that we cannot really find these correlations or the connections across these area or the topic boundaries. And that's a huge limitation we have. That's why we try to make people to work together, talk to each other, and then you have to build up this kind of network, but it takes a lot of effort. And even then there is a limitation due to time being, let's say one dimensional and then space being limited. We are all 3D being said whatnot. But then with these large scale models that are trained on all the data coming from all those heterogeneous sources, and also being able to look at multiple things at the same time, gives these models a superpower that we just never possessed before.

01:04:25 So very exciting time.

Jon Krohn: 01:04:27 Yeah. And these superpowers seem to be progressing very rapidly. Codegen models in December, there were lots of issues when you tried to use them for real world applications, but by February, now we're in March recording, but even in February, there was such a big jump from December to February in terms of what we could do with these codegen models. So it's obvious when you're kind of at that frontier trying to use models to do these tasks that would take a human multiple hours, maybe even up to 10, 15 hours to be able to do. Now

we're at a point where machines can do that with a kind of 50% accuracy over these 12 hour human machine learning tasks. With that kind of exponential progress happening, do you have some sense in your mind? Do you spend much time thinking about where we're going some years from now or decades from now?

01:05:20 How different will the world be? How different will teaching be and research be and maybe even all of our society?

Kyunghyun C.: 01:05:27 Oh yeah. Maddie, I do think about it, but I have no clue. So generally I'm on the side of ... I don't know. When I listen to many of these kind of podcasts or the interviews of my fellow scientists in machine learning and AI or this vision like the Yashan and Yan and so on, they all were raised reading amazing these sci-fi books and whatnot,

01:05:51 But I wasn't the case. I actually still don't read sci-fi. I mean, sometimes if they're really, really good and come with a highest recommendation, I might try to do that, but not much. So my kind of imagination is slightly down to earth. And then what I think is that things always look like they're going to change a lot, but then usually that's only because we are looking forward, but whenever we look back, so in hindsight, those changes won't feel like it was dramatic. So I don't think we are going to see 10 years later saying that, well, now compared to 10 years ago, it's a completely new world and then how we operate or how we live is completely different. We're going to say that yes, it has changed it, but somehow I didn't feel that it was actually going through that kind of change.

01:06:42 So I don't know. I feel like that's how it's going to be like, but my really imagination when it comes to this kind of scientific fantasies, very limited.

Jon Krohn: 01:06:51 Yeah, I understand. I think you and I are on the same page on this in that it seems like I don't have a great

imagination for how quickly these things could progress either. When I was finishing my PhD in 2012, if you had asked me if we could, in our lifetime, have the kinds of code generation capabilities that we have today, I would have said, no way. How could we possibly pull something like that off? Because I was so deep in the weeds of some Beijing model where you're setting all the priors yourself and this kind of felt like the state of the art back then. People working on restricted Boltzmann machines.

Kyunghyun C.: 01:07:23 Precisely, precisely. And now you probably use coding agents daily and then you're very natural at it as well, right?

Jon Krohn: 01:07:32 Yeah, certainly I love it. As somebody who is never a very good engineer. This

Kyunghyun C.: 01:07:36 Helps even more dramatic dramatically, right? Yeah, exactly. Yeah, I love it.

Jon Krohn: 01:07:41 Nice. Well, that brings us to the end of our technical questions for the episode. Thank you for an amazing, super interesting episode. I loved it. Before I let you go, I always ask my guest for a book recommendation. It sounds like it's probably not going to be a sci-fi book.

Kyunghyun C.: 01:07:55 No, it's not going to be sci-fi books. Can I recommend two books?

Jon Krohn: 01:08:00 Yes, I'll allow it. Ah,

Kyunghyun C.: 01:08:00 Thank you. Thank you. So the one book that I read when I was in high school and then was so shocked and then I really loved it. I loved reading it. So I read the whole thing overnight was the hundred years of solitude. So definitely a huge recommendation in particular for students. I think that reading that book was the time at which I realized that the writing can actually have its own power,

	01:08:29	Power of keeping people just not being able to leave and then just read, read, read, and then observe what was actually the message that the author was sending. So yeah, that book I highly recommend to anyone, especially students. And then the second book that I want to recommend is much more modern book. It's The Atomic Human by Neil Lawrence. So it's a pretty recent book. I think it was published last year or the two years ago by Neil Lawrence, who is a professor at the University of Cambridge and has had an amazing, let's say, career as the top class, world class machine learning researcher, the father of the Gaussian process later variable model, GPLVM. And then interestingly, he used to work at Oil Rig as an engineer of the coast of Scotland or something like that, right? So his kind of view of AI or the intelligence is extremely grounded and then grounded on engineering principles, as well as the various other things that are not sci-fi.
	01:09:31	And then reading that gave me really the clear sense of what we are actually talking about because when we think about the AI, and then because AI used to be a sci-fi concept, there are too many floppy UI departs that make it really difficult for us to grasp the core. But this book has nothing like that. It's just a very clear, to the point, kind of as a description about what he thinks is happening and then what we should care about and then what we should worry about when it comes to AI. So I highly recommend it to everyone. My students, I recommended them this book over and over. Not sure if they're reading it though.
Jon Krohn:	01:10:09	Love it. Atomic human. I hadn't heard of that before. It sounds like a great book. Thank you. It's
Kyunghyun C.:	01:10:12	Really
Jon Krohn:	01:10:13	Nice. Yes. And then final question, Kyungyang, how can people follow you after this episode? What's the best ... Do

you post on social media or should we follow the Global AI Frontier Lab? Tell us.

Kyunghyun C.: 01:10:23 I mean, I'm on Twitter or X a bit too much. So yes, everyone can follow me. My handle is KCHONYC. Also, Global AI Frontier Lab has the handle on X as well as BluSky and LinkedIn. And in particular, because we are a more professional organization, we post a lot of the announcements on LinkedIn. And in fact, we have a biweekly seminar series that we started to open it up to the public starting from this semester. And it happens every other week, Monday evening. So they can always register. Anyone can register and then come, listen to amazing research talk, and also have some nice meal as well. We provide a dinner together. What else? They can always email me. I try to reply as quickly as possible, but often fail to do so. But only reason is because I'm not using CloudCode to reply to my emails automatically.

Jon Krohn: 01:11:15 Not yet.

Kyunghyun C.: 01:11:16 Yeah, not yet. Not yet. But yes, they can always email me. In particular, I try to reply to the emails from the college students as well as the grad students almost always.

Jon Krohn: 01:11:26 Wow. Really appreciate you making the time to do that. Must be a lot of info to handle there. Fantastic. Kyongcho. This is an amazing episode, such a great experience to sit here with you for an hour and listen to your thoughts. Maybe we can catch up again in a few years and hear how things are coming along.

Kyunghyun C.: 01:11:43 Yeah, absolutely. Absolutely. It was great talking with you, Jon. At the end, thanks for coming last time, and then hope to see you more often at the Global AI Frontier Lab.

Jon Krohn: 01:11:52 I hope so as well.

Kyunghyun C.: 01:11:52 Yes.

Jon Krohn:	01:11:55	Tremendous episode today with Professor Kyongyung Cho in it. He covered how the attention mechanism was conceived by intern, Dima Budnau in a single morning. And Kyangang said it was one of the only ideas in his career where he could immediately sense it was going to work. It was implemented and validated in three to four days and now has completely changed the world. He talked about how it's not about squeezing more from a fixed dataset because modern models already capture nearly all the correlations in a given dataset. He says the real challenge today is active data collection, choosing which data to gather so that rare, important signals like aha moments occur more frequently. He talked about how on the question of world models, it's an open debate as to whether we need high fidelity step-by-step imagination like video generation models or whether a high level latent representation that lets us skip ahead is sufficient.
	01:12:45	And he talked about how when he started teaching undergrads this semester with coding agents, he discovered that 80% of his 200 computer science students had never installed a coding agent. As always, you can get all the show notes, including the transcript for this episode, the video recording, any materials mentioned on the show, the URLs for Kyongung chose social media profiles as well as my own at superdatascience.com/977. Thanks, of course, to everyone on the SuperDataScience podcast team, podcast manager, Sonja Brajovic, media editor, Mario Pombo, our partnerships team Natalie Ziajski, our researcher, Serg Masís writer, Dr. Zara Karschay, and our founder Kirill Eremenko. Thanks to all of them for producing another super episode for us today for enabling that super team to create this free podcast for you. We're deeply grateful to our sponsors. You can support the show by checking out our sponsor's links, which are in the show notes.

01:13:35 And if you would ever like to sponsor the show yourself, you can get the details on how to do that at jonkrohn.com/podcast. Otherwise, help us out by sharing this fantastic episode with people that would like to hear it, review it on your favorite podcasting app or on YouTube. Subscribe if you're not already a subscriber, but most importantly, just keep on tuning in. I'm so grateful to have you listening, and I hope I can continue to make episodes you love for years and years to come till next time. Keep on rocking it out there, and I'm looking forward to enjoying another round of the SuperDataScience Podcast with you very soon.