SDS PODCAST EPISODE 725: NEUROSCIENCE + MACHINE LEARNING, WITH GOOGLE DEEPMIND'S DR. KIM STACHENFELD

Show Notes: http://www.superdatascience.com/725

Jon Krohn: 00:00:00 This is episode number 725 with Dr. Kim Stachenfeld, Research Scientist at Google DeepMind and Affiliate Professor at Columbia University. Today's episode is brought to you by Gurobi, the decision intelligence leader, and by ODSC, the Open Data Science Conference.

00:00:21 Welcome to the Super Data Science Podcast, the most listened-to podcast in the data science industry. Each week we bring you inspiring people and ideas to help you build a successful career in data science. I'm your host, Jon Krohn, thanks for joining me today. And now let's make the complex simple.

00:00:52 Welcome back to the Super Data Science Podcast. Today's episode with Dr. Kim Stachenfeld is one of my favorite conversations I've ever had on air or off. She's an exceptionally gifted, super fun explainer of complex topics. Kim is a Research Scientist at Google DeepMind, the world's leading AI research group. She's also an Affiliate Professor of Theoretical Neuroscience at Columbia University. Her research interests include deep learning, reinforcement learning, representation learning, graph neural networks, and a brain structure called the hippocampus.

00:01:23 Today's episode should be fascinating for anyone. In it, Kim details her research on computer-based simulations of how the human brain simulates the real world. Simulation of a simulation, yes. She also talks about what today's most advanced AI systems like large language models can do and what they can't. She talks about how language serves as an efficient compression mechanism for both humans and for machines. How a leading neuroscience theory called the dopamine reward-prediction error hypothesis relates to reinforcement learning in machines. She talks about the special role of our brain's hippocampus in memory formation, the best things we personally can do to improve our cognitive

abilities, and what it might take to realize artificial general intelligence. All right, are you ready for this extraordinary episode? Let's go.

00:02:17 Kim, welcome to the Super Data Science Podcast, it's lovely to have you here in person for this episode. It's always nice when I can wrangle a New Yorker to film in person, I think it's a lot more fun.

Kim Stachenfeld: 00:02:29 Yeah, absolutely.

Jon Krohn: 00:02:30 Stare at a screen all day, it's nice to have some in-person interaction.

Kim Stachenfeld: 00:02:35 Yeah.

Jon Krohn: 00:02:36 So yeah, what's that like in terms of, so you're split between both Columbia University and I guess it's the Google offices is kind of the main campus in the Meatpacking District, right?

Kim Stachenfeld: 00:02:48 Yeah. Mm-hmm.

Jon Krohn: 00:02:49 So how often do you go in? And how often do you go into one or the other?

Kim Stachenfeld: 00:02:54 Yeah, so I'm at DeepMind, which is based in the Google office in New York, that's where DeepMind's New York presence is. And I go there four days a week, and then-

Jon Krohn: 00:03:03 Oh, really? In person four days a week?

Kim Stachenfeld: 00:03:06 Pretty much, yeah. I think Friday is sometimes a bit hit or miss, but Tuesday, Wednesday, Thursday, it's just nice, I like spending time with people. It's nice to have the social element for research. And then Mondays I go up to Columbia, that's my Columbia day. Meet with students, go to the lab meeting, meet with the other theory professors, so that's really fun.

Jon Krohn:	00:03:24	Nice. Yeah, I'm so jealous of that, I really miss going into the office regularly.
Kim Stachenfeld:	00:03:30	Mm-hmm.
Jon Krohn:	00:03:31	Yeah, there's a kind of a dynamism and just a kind of being aware of other people's lives that I don't...
Kim Stachenfeld:	00:03:38	Yeah.
Jon Krohn:	00:03:38	No one wants to stay on a Zoom call longer to talk about their weekend, it's like you're like, "Just get me out of here."
Kim Stachenfeld:	00:03:43	Yeah. Oh, there's just something so energizing about it, it's kind of weird how zapping it is, how it just takes your energy away to socialize on Zoom. Whereas, in person it feels like it really feeds you, your scientific enthusiasm and just the benefit of company.
Jon Krohn:	00:03:58	The flip side is getting uninterrupted work in sometimes.
Kim Stachenfeld:	00:04:01	Yeah.
Jon Krohn:	00:04:02	Cool, so we met because it was this random sequence of events that happened to me on the internet. Kind of the weird things that happen I guess when you don't go into an office is I was looking up an image for a previous guest whom I think you know her name, Raluca-Ada Popa.
Kim Stachenfeld:	00:04:21	Yeah.
Jon Krohn:	00:04:23	And this great photo came up of her on an MIT site, on an MIT page. And the person immediately below was Noam Brown, who's also a great guest that we've had on the show. And so it was a list of amazing young people in AI at MIT. And I looked through that you popped out, I think also because of your neuroscience background, which I have as well. So I looked you up on YouTube and I was

blown away, the quality of your speaking. We're going to include some links to some of Kim's talks in the show notes, and I encourage you to check them out because it doesn't get better than your ability to deliver confidently, humorously. I absolutely loved the content, so...

Kim Stachenfeld: 00:05:16 Thank you. Yeah, that's really nice. Like any sane person, I get nervous for public speaking, but I also kind of love it when I get to talk about neuroscience uninterrupted for a good patch of time-

Jon Krohn: 00:05:27 Nice.

Kim Stachenfeld: 00:05:28 ... so I'm glad that came through.

Jon Krohn: 00:05:29 Yeah. And then we've got the show for you here because our audience loves getting deep in the technical weeds. So we've got an exciting blend of neuroscience, of machine learning and the interaction term neuroscience by machine learning. So let's start off with talking about simulated intelligence and how that might generalize. So you recently spoke at a lecture series at Columbia University. And again, so here's the first video of yours that we'll be including in the show notes. And so it appears online with the intriguing title, Can Machines Learn Like Humans? In the lecture, a word that appears often is simulation, which is also often featured in your research papers. And we've got then a short list of those to include in the show notes as well. So what's the significance of simulation in the context of human intelligence?

Kim Stachenfeld: 00:06:20 Yeah, so actually, in preparing for this interview, I looked up simulation, what actually is the definition? It's something we kind of talk about a lot. The definition was imitation of a system or process. It's not very helpful necessarily, I think the kind of intuitive or classic example I think of is when somebody tells you, "Think

before you speak." What they mean is, imagine what you are going to say and imagine how it'll go over. Will people be delighted with that? Will people be offended? Is the consequence what you intended? And this process of playing it out in your head, constructing, using your mental model to construct the situation and then see how it goes, that's basically simulation.

00:07:01 It seems like it's an enormous facet of human intelligence, I think there's a large body of research on this. And I think it's also just very intuitive and familiar as part of the experience of being human, that you try stuff out in your head, you think things through, you see how they go. It's a big part of human reasoning. It's also a big part of how we augment our intelligence, a big technical method that we use a lot for large scale scientific or engineering questions as we build simulators. A stat I learned from-

Jon Krohn: 00:07:30 In our brains?

Kim Stachenfeld: 00:07:31 No, we build them on computers as well. Yeah, thanks, that's a helpful clarification. Yeah, so we build them in our mind, we use mental models to simulate things. We also build simulators of different physical systems, we use them for lots of stuff. I learned this stat that 8 out of 10 of the world's largest supercomputers are being used for simulating different complex physical processes. It's just a huge part of how we do science and engineering.

Jon Krohn: 00:07:56 For sure, we actually recently had an episode focused on this, so Professor Margot Gerritsen in episode 719. That episode is focused on physical simulation, and so she specifically is interested in fluid dynamics. Which also turns out, I hadn't really thought about this, but lots of things have kind of fluid dynamic properties, they don't have to be literally liquids.

Kim Stachenfeld: 00:08:23 Yeah.

Jon Krohn: 00:08:24 So like air flows or even the way that the earth can be moving, predicting volcanoes, that kind of thing.

Kim Stachenfeld: 00:08:32 Yeah.

Jon Krohn: 00:08:34 So they have tons of these kinds of physical simulations. And yeah, I think a lot of these supercomputers get tied up in forecasting weather, for example.

Kim Stachenfeld: 00:08:41 Yeah, that's a huge one, very hard prediction problem.

Jon Krohn: 00:08:44 [inaudible 00:08:44].

Kim Stachenfeld: 00:08:44 Yeah, I think we'll talk about fluid simulation a bit because that's something I've worked on too.

Jon Krohn: 00:08:49 Oh, yeah.

Kim Stachenfeld: 00:08:49 And it's really a classic example of why you would want simulation for physics. We know the rules, they're pretty simple, we know the equations, but actually comprehending their implications is computationally challenging. That's kind of the role that simulation I think serves in general. You can kind of set up a situation, but playing it out, seeing what the consequences are require some sort of mental or computational effort.

Jon Krohn: 00:09:12 Mm-hmm. And so physical processes like weather, that seems to me like quite a different thing from modeling human intelligence or human behavior because with a physical process we have well-defined equations, things like gravity. We're like, "Okay, we'll put in the gravity equation and that's going to have this impact, and we've got friction and whatever." You have all these different equations that kind of work together, and the computer through a whole bunch of crunching is able to be able to simulate a piece of the ocean or a weather system or

whatever, forecast climate over the coming decades. But with intelligence, where do you even start with that? What equation do we have?

Kim Stachenfeld: 00:10:01 Yeah, so this is true for a lot of physical systems we'd want to model to. If you want to apply the same sort of simulation techniques to biological systems for which we don't have great mathematical models, you start doing something that looks a lot more like the brain does. Which is trying to observe data and then build a predictive model, rather than taking some mathematical equation and then obeying it or executing it or solving for it. You kind of have a model of weather too. If I see clouds forming, I suspect it's going to rain soon, I have a predictive model. I don't really know how a voxel of airflow will affect one next to it, but I do know clouds means rain. So there's a level of abstraction.

Jon Krohn: 00:10:47 A voxel quickly there, I think that's from brain imaging that that really comes from, but it's the idea of a pixel in three dimensions.

Kim Stachenfeld: 00:10:51 Yeah, a volumetric pixel, I should say. I think it's kind of a cute word, but it is jargon. Yeah.

Jon Krohn: 00:10:59 So yeah, I guess, I don't know, give us a sense of what you're trying to simulate with these experiments, how you run them, where the state of the art is.

Kim Stachenfeld: 00:11:12 Yeah. So the basic approach that we're interested in is trying to see if we can augment or replace more classical simulators with learned simulators. And this is an approach that folks on my team have been working on for a while. There's a number of collaborators I have at DeepMind, Pete Battaglia, Alvaro Sanchez-Gonzalez, Toby Pfaff, a bunch of colleagues who've been working on this kind of stuff since even before I joined the team. The philosophy is that a lot of things we can model well, we

have good mathematical equations that describe how the system will go. And we can try and solve for this analytically or just use these rules to predict what will happen. But these are limited in a couple of ways. Sometimes they're just really computationally expensive to run. You can run them in principle, but you need tons of money, tons of time, tons of compute in order to do it. And you can use machine learning to make approximations to do almost as good, but not quite as good with a lot less cost. So this is one application we've worked on a lot.

00:12:19 Another application, which I think is in some ways even more exciting, is that there's just some things we don't have the mathematical equations for at all, we don't have anything particularly close to the equations. Neuroscience is filled with examples like this. We have very poor mathematical models for really describing the dynamics of the system at scale in a way that relates to intelligence. So observing the data and then being able to build a predictive model from what you observe, that just seems like such a powerful approach for being able to get a handle on some of these different kinds of processes.

Jon Krohn: 00:12:53 Gurobi Optimization recently joined us to discuss how you can drive decision-making, giving you the confidence to harness provably optimal decisions. Trusted by 80% of the world's leading enterprises, Gurobi's cutting-edge optimization solver, lightweight APIs, and flexible deployment simplify the data-to-decision journey. Gurobi offers a wealth of resources for data scientists: webinars like a recent one on using Gurobi in Databricks, it provides hands-on training, notebook examples, and an extensive online course. Visit gurobi.com/sds for these resources AND exclusive access to a competition illustrating optimization's value, with prizes for top performers. That's G-U-R-O-B-I.com/sds.

	00:12:53	Cool, yeah and that makes a lot of sense. I've read recently, particularly in the context of the same kind of climate prediction, that there are groups, I think even, if I remember correctly, even Nvidia is wadding into this kind of thing themselves.
Kim Stachenfeld:	00:13:52	Yeah.
Jon Krohn:	00:13:52	To have these learned models of weather so that you can do it at a thousandth of the compute with maybe only a small percentage difference in model accuracy.
Kim Stachenfeld:	00:14:02	Yeah, Nvidia is definitely in this game, partly they build awesome hardware that's useful for building the computers that can simulate these kinds of things, whether you're using a machine learning method or not. And then they also have some pretty cool machine learning methods for graphic simulation and trying to capture a system in a high degree of detail.
Jon Krohn:	00:14:20	Cool, very cool. All right, so could we use these kinds of simulations? The better we make these simulations... First, if there's some way that you can kind of concretely describe some of the experiments that you've done, some of the things that you're trying to simulate with a learned intelligence simulation that might be helpful to kind of understand. But then as a kind of direction to go in with that, does having better simulations of human intelligence help us perhaps approach artificial general intelligence, AGI?
Kim Stachenfeld:	00:15:04	So, okay, I'll answer the second part of the question first.
Jon Krohn:	00:15:08	All right.
Kim Stachenfeld:	00:15:08	Does having good simulations of human simulation help? So one thing that I think is kind of useful-

Jon Krohn: 00:15:17 Good simulations of human simulation.

Kim Stachenfeld: 00:15:19 Simulations [inaudible 00:15:19] human simulations, yes.

Jon Krohn: 00:15:19 So this is key thing here is that specifically your simulations are often of our mental simulations.

Kim Stachenfeld: 00:15:24 Yes.

Jon Krohn: 00:15:25 So the thinking before you speak is the kind of modeling that you're modeling.

Kim Stachenfeld: 00:15:31 Yes, simulating a simulation of a simulation. Yeah, that's exactly it. So I think basically in general, understanding human cognition or just the brain's cognitive mechanisms for solving things potentially just has a lot to add. Right now the state-of-the-art language models, the really big language models that do fantastic jobs at stuff, ChatGPT, things in that family, they imitate human behavior right now. They don't imitate the aspects of human thought that are not captured in behavior, but they imitate human behavior. And the fact that we have gigantic data sets with very rich human behavior has just been fantastic for getting these models off the ground. You have a warm start on language processes that you can then adapt to lots of different kinds of language-based tasks. There's a lot in language data that isn't reflective fully of what's going on in the brain. You can imagine that's what's happening between the ears and the mouth could also have really useful elements of the cognitive process that aren't present in the text form.

00:16:40 So I think in general, just understanding what's going on inside of the brain has a lot of potential information. If you thought before you spoke, what did you decide not to say? That probably has a lot of information about social structures and your goals and your thinking process and things like that. If you paused for a long time before

speaking, what were you thinking about in that time? That's very interesting information, information that I think would have a lot of potential. The limitation of course, is just the data, as you know, it's pretty non-trivial to record what the brain's doing and model it. As it relates to simulation in particular, there's a lot of extra benefit because that relates so much to our ability to imagine new outcomes, reason about a world that's different from the one that we have. If we want to say, "What's going to happen if I do this differently? What would have happened in this data if something had been different?" These are processes that humans reason about with simulation.

Jon Krohn: 00:17:39 It drives us all mad, doesn't it?

Kim Stachenfeld: 00:17:40 Yeah.

Jon Krohn: 00:17:41 It's this constant... Even we kind of got into it a little bit before we started recording. I actually didn't even really finish talking about this, but when we were talking about what you do for a living, and I was starting to describe how jealous I am of what you do. And how there's things about what... It's so amazing to me that you get to be at DeepMind, arguably, but I think in my opinion, the best AI lab to be working in in the world, while simultaneously being Columbia faculty. Doing neuroscience in intersection with machine learning, this is I'm so jealous-

Kim Stachenfeld: 00:18:24 It's pretty sweet.

Jon Krohn: 00:18:24 ... of what you're doing. And I'm like, "Where did I mess up along the line?" So I'm simulating just in the run-up to this interview, I'm doing so many simulations in my head of, "Well, okay, I messed up there, I messed up there, I messed up there." In terms of, okay I'm like, "Some things I got right." Where I was like... This is what I started explaining to you before we started recording is during my

PhD, while other PhD students were getting very specialized in doing recording electrodes in a ferret. I was like, "That's not a super transferrable skill, spending five years becoming really good at putting recording electrodes in a ferret. There's not a lot of places where that is going to come in handy."

Kim Stachenfeld: 00:19:00 It's important, but it is specialized.

Jon Krohn: 00:19:02 Yeah, super specialized. And yeah, you get amazing results from it, and I can totally get why somebody would be super passionate about that one specific thing. But I was specifically looking at transferrable things and I was like, "Okay, teaching myself better programming skills, machine learning, these are going to be useful kind of whether I stay in academia or not." But yeah, one of the really big things for me is that... So my PhD was called neuroscience, but I was working with machine learning labs. And so we had this multi-year collaboration with people at the University of Edinburgh, which is amazing for AI research, has been for decades. And so I can't remember the exact, I think it's 2010, I have a NeurIPS paper, working on this thing with people from Edinburgh, but I've never been to NeurIPS. And I'm like, "What was I thinking? Why was I going to Mouse Genomics Conferences in my PhD and not NeurIPS?"

Kim Stachenfeld: 00:19:59 Mice have some pretty sweet genes.

Jon Krohn: 00:20:06 They do.

Kim Stachenfeld: 00:20:06 Yeah. No, I think in terms of... As you were speaking, I realized I might be proposing some really anxious machines or we need to introduce an extra dose of neuroticism. I want a machine second guessing it's every move, that kind of thing would... It can obviously have a regime where it makes us less happy, but I don't know, it's a big part of how we optimize ourselves.

Jon Krohn:	00:20:28	Yeah, I think if we're going to simulate intelligence in machines, I want them to be just as miserable as all the rest of us-
Kim Stachenfeld:	00:20:34	It seems only fair.
Jon Krohn:	00:20:35	... constantly going over their mistakes. That would be great.
Kim Stachenfeld:	00:20:38	Yeah, you definitely don't get a sense from interacting with them that they're doing a lot of second guessing, right? That they're really deeply apologetic when they get it wrong.
Jon Krohn:	00:20:49	So these simulations that you do, what's the output? Or what's the input?
Kim Stachenfeld:	00:20:54	Right. Yeah, so we've worked on a couple of different domains. And the machine learning side of my research where we're working on learned simulation, on learning physical dynamics, we've focused on a couple of different types of physical systems. One example is fluid dynamics, we collaborated with some researchers in New York at Flatiron Institute, astrophysicists who are interested in fluid dynamics, because it turns out a lot of stuff in space is made out of fluids. So when galaxies form, that's a fluid dynamic event. And they're particularly concerned with trying to make these simulations run more efficiently because as they explained it to my non-fluid dynamics brain, space is big and galaxies are pretty hot. So you need a really high resolution and large simulation to get everything. Basically, the kind of system we used for this was we would use what simulators they currently have to make a simulation of some fluid dynamical process.
	00:21:52	Whether it was the mixing that would occur at the boundary of a galaxy or some other kind of more classic fluid dynamic system. We would simulate it at high

resolution, these would be pretty expensive to get. And then we would make it lower resolution and try and train a machine learning model to predict what's going to happen without access to all these details. This was kind of a specific project in some ways, but it's really exemplary of the role machine learning can serve in trying to make simulations more efficient. In physics, if you want to run something more efficiently, you're basically just simulating a different physical process, you're simulating what it would be like if you just had a much coarser system interacting. And if you want to try and say, "Okay, I have this coarse system, but I want to know the correction. I want to know how it would be a little different and subtle and hard to mathematically describe ways if it was higher resolution." There's patterns to that, but they're hard for humans to articulate, they're hard to express in math. So you can use machine learning to try and pick up those subtle patterns and make the same predictions, but at lower resolution.

Jon Krohn: 00:22:59 Mm-hmm. But the thing that led me immediately to my first question, this galaxy stuff is super interesting for sure. But I don't know to what extent you were joking or whether there is something that you're actually looking into here. But when I was talking about making sure that these simulations of our mental simulation are brooding all the time, masticating over and over, just chewing on these past events and wondering why they didn't go to NeurIPS all those years ago too. You said they don't seem angsty.

Kim Stachenfeld: 00:23:38 Yeah.

Jon Krohn: 00:23:38 So was that just a joke or are you interacting with some simulations of mental simulation where you are in some ways getting a sense of what they're [inaudible 00:23:52]?

Kim Stachenfeld: 00:23:52 Yeah, so that's an interesting question. I was partly just joking, but I do think that there's some substance to it. They're rolling it out once and they're using... A large language models gathering, all of the computational resources it has and playing it out once.

Jon Krohn: 00:24:06 Okay, so it is LLM. So you have LLM simulating?

Kim Stachenfeld: 00:24:08 No, sorry, for the fluid dynamics, we're not using LLMs.

Jon Krohn: 00:24:11 No, not, for fluid dynamics, of course, yeah.

Kim Stachenfeld: 00:24:13 Yeah, so I think maybe another... These physical systems we're doing are not really ruminating either because we're not asking them to solve a problem with the simulation, we're just saying, "Simulate."

Jon Krohn: 00:24:26 Right.

Kim Stachenfeld: 00:24:27 And we leave it to the physicists to ruminate and to set up different initial conditions and play the model again and again and gather statistics. That sort of the rumination is initialized by the researchers using these as a tool. Another project that I worked on was maybe more... it was more task oriented. It used simulations to try and design solutions to different physical problems. So it was basically, we have a fluid dynamic simulator, it's trained with a learned model. So it basically learned fluid dynamics just by looking at them. And we want to see, "Okay, here's a fluid challenge, a bunch of water's going to fall out of the sky, can you catch it in something? Or can you move it over there? Or can you navigate a system of pipes to put it somewhere?" And there, the process we use involves running the simulator again and again and again and refining it according to what happened.

Jon Krohn: 00:25:21 Right.

Kim Stachenfeld: 00:25:23 Yeah.

Jon Krohn: 00:25:23 Okay. Okay, so you are simulating a mental model of a physical process.

Kim Stachenfeld: 00:25:30 Yeah.

Jon Krohn: 00:25:30 Okay. Okay, so yeah, there isn't any, at least at this time, simulations of make a plan for grocery shopping. You're hosting a dinner party for six people, come up with a plan for a successful dinner party so that this cute person at the dinner party wants to date you or something.

Kim Stachenfeld: 00:25:54 Yeah, there's work on this kind of stuff. So I haven't worked much with language models, I think I like using them as examples because they're intuitive to people and also just vastly familiar and kind of exemplary of why AI is exciting right now. But the stuff I've done has been all with physics simulation or navigation.

Jon Krohn: 00:26:13 Yeah, so your stuff is... Okay, so that's a concrete example. It's there's a bunch of water falling from the sky, come up with some kind of solution that catches the water, and it has to think it through. So this reminds me of kind of... There's stories of prisoners of war who will get through the decades that they were interned by imagining that they were golfing their local golf course that they used to do all the time.

Kim Stachenfeld: 00:26:43 Yeah.

Jon Krohn: 00:26:44 And then when they get out of prison, they're better at golfing than ever, or I don't know. I don't know if that's actually a fact or more urban legend, but-

Kim Stachenfeld: 00:26:52 Yeah, could be. There's definitely well-documented studies on visualization. And I think longer reaction times is often associated with more performance. The

hypothesis being that there's some simulation or some visualization or process that takes time, that must be dynamical because it takes some time. Wherein people reason about things and then improve even on really short time scales.

Jon Krohn: 00:27:18 Yeah.

Kim Stachenfeld: 00:27:19 It does seem like it's part of the process, there's some work too more in the applied psychology world of positive visualization makes you perform better. I didn't walk over to your apartment today thinking about all the ways I could blow it. I don't think that would be a healthy...

Jon Krohn: 00:27:34 Oh, we still have plenty of time.

Kim Stachenfeld: 00:27:35 Yeah, we'll see.

Jon Krohn: 00:27:38 It's going so far so good, but yeah, we're going to try to mess it up at some point. Yeah, because it's interesting. So there's that kind of simulation of the positive psychology of simulating just the things are going to go well and you visualize success. And then there's also this kind of... And so that's probably something to do with, and we are going to talk about the hippocampus and that kind of stuff later, neuroscience stuff. But yeah, that's probably something to do with just kind of framing your perspective on some event, potentially a stressful event. But then there's also interesting... this idea of the golf swing.

Kim Stachenfeld: 00:28:19 Mm-hmm.

Jon Krohn: 00:28:22 I suspect that probably has something to do with, so we have the cerebellum, which I think literally means tiny brain at the back of your brain, and is responsible for motor coordination. And it has tons and tons, the density

of connections there is much higher than in other parts of the brain.

Kim Stachenfeld: 00:28:43 Yeah.

Jon Krohn: 00:28:45 And this seems to be... So something like regularly practicing something like a golf swing or playing the piano, you will... it seems like it's associated with developing a lot of connections in this cerebellum, and it kind of coordinates that fine-grained motor activity. And so without being an expert in this at all, it seems like these kinds of simulations that we run in our own head allow something like that, those cerebellum connections in the case of the golf swing or the piano playing to form, even in the absence of the physical work.

Kim Stachenfeld: 00:29:25 Yeah, that's cool. Yeah, I also, not a cerebellum expert, so a little bit outside of my wheelhouse, but the cerebellum, the kind of classic cerebellum example is this experiment people do with prism goggles. So you put prism goggles on and [inaudible 00:29:40]... You probably did this also at some point [inaudible 00:29:43].

Jon Krohn: 00:29:42 I've never done it, but I've seen videos of people.

Kim Stachenfeld: 00:29:44 It's fun. Yeah, they also give people... Well, okay, I'll explain the experiment, but basically you put these goggles on that shift your vision to the side a little bit. They just take all the incoming light and shift it a little bit so that your vision is not aligned with what's happening in the world. And then they have you try and throw a ball at the wall and hit a particular target. And of course at first you miss because your vision is off and you're aiming maybe a little to the left. If you keep these goggles on, keep practicing, eventually your throw corrects. And this has been linked to cerebellar learning, that this very rapid correction, this adjustment to align your motor behavior with your observations. This is something that cerebellum

classically does. And there's a lot of connectivity between parts of cerebellum and parts of motor cortex. The motor learning makes a lot of sense.

00:30:31 But there's also really similar, there's different lobes of cerebellum that have different connectivity to the rest of cortex, the rest of your brain. And it seems like similar circuitry is present in another part that connects to the more cognitive parts of your brain, the parts that are less about motor and more thinking. I don't really know what these do, but it's kind of hypothesized maybe they also serve some kind of error correction, but on your cognitive processes and that seems, yeah, I think so, I don't totally know. Seems reasonable to speculate about though.

Jon Krohn: 00:31:05 Be where our data-centric future comes to life at ODSC West 2023 from October 30th to November 2nd, join thousands of experts and professionals in person or virtually as they all converge and learn the latest in deep learning, large language models, natural language processing, generative AI and other topics driving our dynamic field. Network with fellow AI pros, invest in yourself in their wide range of training, talks and workshops, and unleash your potential at the leading machine learning conference. Open Data Science Conferences are often the highlight of my year. I always have an incredible time. We've filmed many Super Data Science episodes there and now you can use the code SUPER at checkout and you'll get an additional 15% off your pass at odsc.com.

00:31:48 So yeah, I'll be sure to include a link to at least one of these prism videos, is pretty fun. And there's also, I'll try to look up, not try to look up, I am writing a note to remember to look up, this is like a random tangent, but I really like, there's an artist Ben Folds, and he also had Ben Folds Five.

Kim Stachenfeld: 00:32:11 The musical artist?

Jon Krohn: 00:32:12 The musical artist. Yeah, and he has a song where he uses one of these prism stories. He uses it as the verse of one of his songs.

Kim Stachenfeld: 00:32:21 Oh, cool.

Jon Krohn: 00:32:21 Where it explains basically what you explained about the experiments.

Kim Stachenfeld: 00:32:24 Oh, neat.

Jon Krohn: 00:32:24 In a verse of one of his songs.

Kim Stachenfeld: 00:32:26 Oh, that's cool.

Jon Krohn: 00:32:27 And it's something like he's relating it to life more broadly somehow. I forget how.

Kim Stachenfeld: 00:32:31 You should have him on the show, play that guitar.

Jon Krohn: 00:32:34 Yeah, I mean that would be an incredible guest. If anyone out there can get me Ben Folds on the show, we could at least do a Five-Minute Friday with him as the guest. No, seriously, if he wants to Tuesday slot, we'll do it. So yeah, so I've had a whole bunch of thoughts. Really quickly one, which is one that I didn't prepare for, but I've just had my own, I've been running some simulations in my head on things to say, and one of the simulations that popped up was, do you happen to know someone, he's based at DeepMind, I believe in London, although he did postdoc work here in New York at NYU. His name is Neil Rabinowitz?

Kim Stachenfeld: 00:33:17 Yes, he's awesome. I love Neil Rabinowitz. He is just like, yeah. How do you know Neil? Did you...?

Jon Krohn: 00:33:24 Neil and I were, we did our PhDs at Oxford at the same time.

Kim Stachenfeld: 00:33:27 No kidding? Yeah, he's awesome. He is so smart and so poetic.

Jon Krohn: 00:33:34 He was doing recording from ferrets with electrodes, I'm pretty sure. When I give, that was kind of my random example of some very specific thing to be learning. But I'm pretty sure Neil was doing exactly that.

Kim Stachenfeld: 00:33:46 He's branched out. I guess he maybe had a similar, I don't know, he is doing machine learning AI stuff now, he's at DeepMind.

Jon Krohn: 00:33:53 Well, he was doing, so in our master's in neuroscience year, we had to do research projects. You'd spend a term doing a shorter research just a few months in different, and so you could do, it was at least two of these rotations into different labs. And I remember when he was analyzing his results, he used an artificial neural network to analyze his results. And at that time, I had never come across someone having done that before. So he's always, I think he's had machine learning applications and kind of cutting edge. Because at that time to do, I don't even know what you would program that in to do it in, it would've been 2007.

Kim Stachenfeld: 00:34:34 Oh wow.

Jon Krohn: 00:34:35 So yeah, I don't know.

Kim Stachenfeld: 00:34:37 Honestly, I don't know either. I was in high school.

Jon Krohn: 00:34:44 But yeah, I recently saw some fascinating, I don't know how I came across it. I stumbled across that Neil had been doing... And so it relates to the simulation stuff because I remember a few years ago I came across that he

had done a poster at a conference where he was trying to understand simulations like intent. So I thought maybe-

Kim Stachenfeld: 00:35:10 Is this his theory of mind stuff?

Jon Krohn: 00:35:11 I was hoping for less of a, "Hmm," and more of a, "Mmm," and continuation. It was theory of mind stuff. So it was simulations of theory of mind and it was like one machine learning algorithm is watching another one learn and the first one is trying to guess what the second one might do.

Kim Stachenfeld: 00:35:30 Yeah, so I don't remember. It was a little while ago. I don't remember that work super well. But yeah, it was about trying to operationalize theory of mind for machines. This old idea in cognitive science that we have a simulation of each other's minds and how we're going to think about things, is often kind of also extended to that, maybe I have the same model of myself that I do of other people. And this is deeply related to consciousness and how we reason about our own minds and other minds in the same space. I mean it's been very influential in social psychology and thinking about how people reason about each other. I think I was going with the language model example earlier, like having a model of how your user will respond and how you'll respond, all kind of bundled up in the same system, has some elements of it, but a lot less explicit as a model.

Jon Krohn: 00:36:27 I don't know, there's fascinating stuff there and maybe we can just put a pin in that is something that, now that we've had this conversation, I've written a note down to try to get Neil on the show.

Kim Stachenfeld: 00:36:34 Yeah, Neil would be great.

Jon Krohn: 00:36:36 Yeah, it would be an incredible episode. So stay tuned for that, an episode on simulations of theory of mind. And that was several years ago, so his research might have

developed beyond that now. But yeah, it'd be great to have him on. Okay, so back to what we were talking about. I was trying to get, so now I think I have a better understanding of the kinds of simulations that you're doing. So you're doing mental models of physical processes and I think, you said you were going to answer this question first, but I don't feel like we've talked about it too much yet because I've probably taken you off on too many tangents. But do these kinds of simulations, do you think that they are helpful for realizing artificial general intelligence? Do you think they're kind of a step on the way?

Kim Stachenfeld: 00:37:23 Yeah, I definitely do. I think that basically, I guess there's different kinds of simulation. I think one, there's the models that we have doing language reasoning, for instance, language models and they are predictive models. They're making predictions about what's going to happen next. That prediction is just synonymous with the actual thing that's going to happen. It is not like there's multiple different things that it's reasoning about. I think one of the things, there's a couple different use cases I think, for more explicit simulation and they come up often when you want to try to use these models for a particular function. So in the design example where we're using physical simulations to design something, if you don't get it exactly right on your first try, you want to be able to iterate and improve and try multiple different configurations to see if something else is better.

00:38:17 A particular use case of them, I think, especially the way thinking about the way humans construct mental models is that we can combine things in different ways and try them out. So if you want to think how do we go beyond what we've directly experienced, being able to create new combinations of things, new compositions of things and try them out, see what happens seems really important.

Jon Krohn: 00:38:39 There is even actually, there's a bit of an analog to that. Even with the way that LLMs make their next prediction, because you can have different kinds of search over the possibilities. So there's things like beam search or contrastive search that allow you to run this simulation several times and then pick the best one or pick, mix and match even a little bit to get the best language output. And obviously that increases the computational complexity, but it's like, it's having these kinds of search mechanisms with LLMs, it's kind of asking it, okay, five times, think through in your head what the best thing to say is and then pick the best one or mix and match from the five things that you've thought of.

Kim Stachenfeld: 00:39:30 You can think of it almost as just like ensembling these models, I have a bunch of them, I'm going to try them out, I'm going to see which one is best. Ensembling, that's just a technique where you have multiple instances of the same model under slightly different random deviations and try them out. So I think I'm kind of pausing a little bit on this question just because a sense in which these models already are a kind of simulation, if I wanted to simulate, I could ask ChatGPT, what would happen if I submitted this essay? What would the teacher say? What would the grade be? It would say something, and that's kind of a simulation. But it's a little bit different than the way we use simulation and all of the different abilities that it can afford in terms of reasoning, more abstractly and reasoning about things that are outside our direct experience, outside things we've seen before.

Jon Krohn: 00:40:19 So something else, and I realize I keep going back to large language models, but I guess it's something that's kind interesting.

Kim Stachenfeld: 00:40:27 Bit of an attractor right now.

Jon Krohn: 00:40:28 Yeah, exactly. We actually recently, at the time of recording at least, Kirill Eremenko, who founded this podcast and who was host of the show for the first four years and he still owns most of the show and he sent me a Slack message, he'd come back from a month away on holiday and he listed the last 10 episode numbers and he was like, "LLMs, LLMs, LLMs. He was like, eight of the 10 were focused on LLMs." And I was like, "I don't really do that on purpose." But it just seemed like, it's like we get amazing guests on and it's what they want to talk about. And then for the episodes that I'm doing solo, I often feel like it's the thing that's changing most quickly and the listeners need to be aware of the most. And he didn't think that was totally wrong. But I also then said, I was like, "And look at who we have planned as the next guests. We won't be talking about LLMs." And you were an example of someone I was like, "It won't be an LLM episode."

Kim Stachenfeld: 00:41:31 Yeah, well I mean, it's omnipresent. I mean it's relatable to people. So it's an easy example to draw on to for that reason. But then also the performance of large language models really has just changed my perception about what kinds of things can be possible. I mean a big topic I've thought about has been like how do you get something new? Machine learning, learning is fundamentally about patterns in your past, patterns in your experience. It works on things that are familiar, almost by definition. But we all kind of have this intuition like, oh, you can learn to do something new or you can use things you've learned and still do something new, generate something creative. And there's ways in which these models do that and there's ways in which they don't. And thinking about that kind of decomposition has just been really, it's really I think solidified a lot of the ways that I think about this now. It's been very illuminative.

Jon Krohn: 00:42:26 As well, do you have anything more that you'd like to add on that?

Kim Stachenfeld: 00:42:30 Yeah, I think a big thing people talk about with language models is in context learning, that you can have some kind of pattern or template or thing you've seen before but adapt it to a new context. So like a recommendation letter for a new person, a biography for a new person. There was a while where everybody was saying, "Write a biography for Kimberly Stachenfeld," and sending it around like, "Ha ha, that's pretty funny." And got some stuff right and some stuff wrong. So it can do something that's novel, that's applying its patterns to a new context, something that isn't already-

Jon Krohn: 00:43:02 Write a job description in the style of a '40s gangster.

Kim Stachenfeld: 00:43:04 Totally. And that kind of mix and match, that's like a form of composing novel things. What kind of works about it is it's seen such a gigantic rich pile of data. It's seen so many different patterns. It's seen them applied in so many different ways that it can generate new combinations of them. The kind of thing that it won't necessarily do, just the kind of thing that's not really a strength of the method, is that it won't go beyond the complexity of anything it's ever seen before. It's not going to write, I mean fundamentally it can't write a story any longer than it's seen before. It's bound by its context length and the length of stories it was trained on. It's not going to take a, if it's seen, if you've told it, what happens, I don't know. I guess I think of dominoes as an example because it relates to physical simulation.

00:43:53 It won't do the equivalent of having seen a row of 10 dominoes and then telling you that effectively the same thing will happen for a row of a 100 dominoes. It won't extrapolate or build on the parts it's seen. So it can mix and match things. It can make novel things. It can

definitely do, I would consider that abstract, compositional, in some ways the elements of creative. But it doesn't necessarily tell you how you would go from seeing some kind of simple rules or some limited data set and construct something more sophisticated than anything you've seen before.

Jon Krohn: 00:44:28 Mathematics forms the core of data science and machine learning. And now with my Mathematical Foundations of Machine Learning course, you can get a firm grasp of that math, particularly the essential linear algebra and calculus. You can get all the lectures for free on my YouTube channel. But if you don't mind paying a typically small amount for the Udemy version, you get everything from YouTube plus fully worked solutions to exercises and an official course completion certificate. As countless guests on the show have emphasized, to be the best data scientist you can be, you've got to know the underlying math. So check out the links to my Mathematical Foundations and Machine Learning course in the show notes or at jonkrohn.com/udemy. That's jonkrohn.com/U-D-E-M-Y.

00:45:13 That's super fascinating and so I don't know if you have any thoughts on what we can be doing to bridge that. Something that I used to argue in a pre GPT-4 world, which as you say, GPT-4 has vastly changed my perspective of what could happen in our lifetimes with AI. It also leads to a lot of soul-searching for me around being human and what value we can provide. It's like already you see these glimpses in that today that it just does, there are so many questions now, that I know it can answer better than almost anyone on earth could. Because while yes, there are some constraints on what it can do, like you just described its ability to hold so much information and be able to blend that together however you like, is unreal.

00:46:23 And it's like, for so many questions you could be like, well, to answer that question I need to, there might only be a few experts in the world that would understand something to that level of detail. Obviously it would be hard to find them, much easier to ask just GPT-4. But then if you're like, okay, well I'd like to blend two different research ideas, and actually this is something that I recently talked about on the Last Week in AI podcast. I was a guest host on that show and I was reviewing, I was talking about an Economist article that was describing how we can be using today and even more so in the future, LLMs like GPT-4, to be scouring research and suggesting where there might be opportunities like cross-disciplinary opportunities because it can be so expert in so many different things. And so it could even do things like it could suggest to you like, "Hey, you might want to consider working with this other researcher at this other lab because they have this other expertise and you guys could do this thing together and it could potentially lead to these discoveries."

Kim Stachenfeld: 00:47:30 Yeah, I mean no single human has been trained on the entire internet, which is probably fine. I mean I think it definitely has seen more information of a particular type, at least more information that is text-based than any one human. And it will have pretty powerful interdisciplinary abilities because of that. I mean, one thing it won't do, it might not propose the idea in the first place of using ChatGPT for research or something.

00:48:05 Still the decision, I really do kind of think of it as a simulator in some sense. If you have some idea you want to flesh out, like what are maybe interactions between material science and neuroscience that could be useful? It'll generate a couple ideas. Some of them will be potentially interesting, some of them will be probably nonsense. And you can use that in concert with your own intuition about what are problems worth doing, what's

important to do, your own judgment about what's grounded and factual and what's kind of just maybe statistical nonsense. That kind of integration with human performance seems really, really powerful. But there's still, in thinking about what role do humans serve, are we getting automated by this? It does some stuff that humans currently do, but I don't see it totally supplanting human cognition.

Jon Krohn: 00:49:01 Yeah, I-

Kim Stachenfeld: 00:49:02 It's like a calculator for words. Like a really good calculator.

Jon Krohn: 00:49:04 I agree. I agree. I agree and I agree for now, but the point that I was trying to make is not that GPT-4 does everything yet, obviously it isn't this, and there's a lot of ways that we can be defining AGI, but in terms of AGI just being an algorithm, a single model that can do all the various kinds of thinking and tasks that a human can. Obviously GPT-4 is nowhere near that today. But what I meant by what I was saying earlier is that it is such a huge step change from GPT-3.5 that and who knows, maybe scaling, will run its course and it'll turn out that with scaling there is some barrier that we run into. But it seems like we still have some orders of magnitude potentially, of scaling to go. Plus some clever ideas on maybe how we can achieve some of the same. Like everything right now is relying on a transformer architecture. Which is, I mean maybe just kind of a random choice. And it could be the case that there's some way more computationally efficient way of having an attention mechanism that is even more effective over long stretches of language.

00:50:17 So yeah, it's just this sudden, this trajectoring of number of parameters or complexity of approach, GPT-3.5 capabilities. Six months later, GPT-4 capabilities, more

tokens, more parameters. It just seems like we're going this trajectory where I'm like, the number of cognitive tasks that got usurped that only humans could do in that one step. Yeah, I don't know. It just seems like we're moving in a really interesting direction and I am not giving you much of a chance to speak. And I know you have-

Kim Stachenfeld: 00:51:00 No, that's okay.

Jon Krohn: 00:51:00 I know you have a really, really interesting thing to say, but there's a specific context that I kind of want to frame whatever you're about to say is in, which is kind of relating back to your neuroscience stuff. So pre GPT-4, I recorded two podcast episodes, episodes number 588 and 590, so these came out in July of last year, July 2022. And I called them AGI is Not Nigh Part 1 and AGI is Not Nigh Part 2. And a big part of, I think it was in the second episode in particular, I made the case that the way that we are modeling intelligence is so simple. Like we are using, for example, the transformer architecture and just scaling it up. And at that time, in July of 2022, it seemed to me like there wasn't enough nuance, enough sophistication because the human brain has things like the cerebellum that we talked about earlier, the hippocampus, which you particularly have a lot of research background in. There are so many different kinds of brain structures that to me, seem like altogether like these different kinds of intellectual processing need to be combined together. We can't just be like, okay, let's take one thing to the transformer and scale it up. And so I don't know. Finally, I'll let you speak.

Kim Stachenfeld: 00:52:25 I guess there's a note about transformers. I mean transformers are really cool. And one thing about them, a lot of the things about them are ways that they just are good for problems where you have a huge scale of data. They scale up well, they train efficiently. They seem

capable of learning diverse problems in a giant patch of data. But then there's also things about the way that the structures over which they operate, which are very conceptually compelling. And one thing is that the form of data they take in is very general. Traditionally sequence models, models that are trained on prediction problems like language, they take in a sequence of data. You have time step one, time step two, time step three, time step four, and that's a bunch of vectors in a row. If you have a model that operates over image data, it will take in an image, a two-dimensional picture where each pixel has some values associated with it.

00:53:31 The kind of data that transformers take in is a set of tokens. These tokens could be elements in a sequence. They could be pixels in an image. They could just be something more general purpose than that too. They could be particles of a fluid that you want to consider their interactions. They could be objects and you want to consider about how they're going to bump into each other and relate to each other. The fact that they operate over this very general data structure and can process different kinds of relational structures, whether it's sequential or image-based or more relational, is a really deep property that might actually make them apt for lots of different kinds of processes.

00:54:14 So I think that there is something kind of general purpose about this. And in terms of their role in sequence modeling, they, in a lot of ways, started modeling sequences not as sequences. They started making it really easy to learn interactions between words in a sentence that are really far away from each other. Whereas in sequence models, it's a lot easier to reason about relations between things that are close together in a sequence, where most of the linguistic structure is. But if you want to, for instance, remember the name that was said at the beginning of a sentence, that kind of relation

can be really, the fact that you can reason more flexibly about these, can be quite useful. Just like as architectures, they are pretty interesting and especially, yeah, I don't know, I think as they relate to a problem with as much structure and variety as language modeling, I think it's interesting that these have kind of searched the front of the pack.

Jon Krohn: 00:55:01 Yeah. So one of the interesting things about all that you're talking about here with transformers and then being so flexible, and part of why I now feel like my AGI is Not Nigh thing and human brain structures argument that I was making a little over a year ago. I think that what I've realized is that it can turn out that the way that we allow machines to train, for example, by training them on all of the internet, which obviously a human brain, maybe not obviously, but it seems the human brain can never do that. It seems like no amount, there's too much, the brain doesn't live long enough to possibly be able to read all that and then retain it.

00:55:45 And so even if GPT-4 is not able to capture the full breadth of human intellectual capabilities simultaneously, it is doing something, in the same way that the calculator example is really good one because the calculator can do all kinds of arithmetic much faster than a human brain could. And so similarly, this tool, it can be capturing intelligence in a new way and doing it in a different way. Maybe it doesn't make any sense at all to be thinking, "Oh, we're going to need a hippocampus part and a cerebellum part if we're going to have AGI." That might not be the case at all. Maybe we can have a machine that can do all the kinds of things that our brain can do and more, maybe just following some kind of simple thing like scaling up a transformer.

Kim Stachenfeld: 00:56:40 And I actually, I remembered, I think didn't make this connection explicit, but the reason that the generality of

the data structures that transformers operate over occurred to me, is just thinking about their usefulness for multimodality. This ability of the brain to process different types of information and process it in different ways. So got different types like audition and vision.

Jon Krohn: 00:57:02 Yes. This is my literal next thing that I also wanted to talk about. This is perfect.

Kim Stachenfeld: 00:57:07 Yeah. So I mean you've got different modalities like vision and audition. Then you've got different kinds of processes, like cognitive processes like hippocampus, for instance we think of being really good for episodic memory and then other cortical areas for more semantic memory. So memory for your own experience versus memory for general purpose knowledge you have. The fact that transformers can operate over general data structures makes them good for multimodality and also potentially good for different kinds of cognitive processes that can be expressed in relational terms where you want to reason about how the different entities that you're thinking about relate to each other. Right now, it's all kind of a big bag of computation, and this is in some sense, quite powerful because it lets you seize on every possible statistical correlation. On the other hand, you might not want that. You don't want to, if for instance, the information that you've trained on has changed. You don't want to have to relearn how to produce language or something. If I want to incorporate updates from the most recent news cycle, I don't want to be retraining a part of my system that knows how to produce language or something like that.

00:58:18 This is something called the continual learning problem. It's been a problem in neuroscience for a really long time. How does the brain keep updating itself, keep acquiring new knowledge, new abilities without just overriding every other thing that it's learned before? It doesn't maybe seem

intuitively obvious why that would be a challenge. Why would learning new things necessarily compete with what you've already had before? But when you actually start trying to implement this in machine learning systems, it's really hard to update without overriding or racing or recontextualizing everything you've learned before.

Jon Krohn: 00:58:50 Catastrophic forgetting.

Kim Stachenfeld: 00:58:52 Catastrophic forgetting, a beautiful phrase. Yeah.

Jon Krohn: 00:58:54 It's so dramatic. Why isn't it just forgetting? It's catastrophic.

Kim Stachenfeld: 00:59:00 No, I actually never totally understood why they didn't just call it forgetting. But I think a lot of people are drawn to the drama of the term.

Jon Krohn: 00:59:09 Not only did it forget how to do this task, but it also caused an earthquake.

Kim Stachenfeld: 00:59:12 Yeah, it was so bad that we just kind of lost a chunk of history or something. Yeah, I think catastrophic forgetting is this problem specifically of overriding information that you've learned previously when you try to learn something new.

Jon Krohn: 00:59:28 Yeah. So that brings up, so this leads perfectly to the next thing that I want to talk about because with this catastrophic thing, catastrophic forgetting, catastrophic thing, or the continual learning problem, this ties into the idea of negative transfer. So this idea that, and it seemed up until recently, so for example, it's a few years ago now, the Gato model, it was an approach, maybe it was kind of around the time that GPT-2 came out, kind of era and Gato was the idea, and I can't even remember now, you might, but I can't remember what kind of architecture they were using and scaling up for that.

Kim Stachenfeld: 01:00:14 Transformers.

Jon Krohn: 01:00:15 It was Transformers, but it wasn't exactly the same as the GPT kind of setup somehow.

Kim Stachenfeld: 01:00:23 Yeah. I'm not sure exactly how.

Jon Krohn: 01:00:23 Yeah, I can't remember exactly. But with Gato, they observed at least something to do with the architecture or the training regime or something. But Gato was designed to be able to handle a very broad number of tasks, and they observed a negative transfer where, as they tried to add in more tasks, it would perform worse. But what we're seeing now in the GPT series architectures recently is positive transfer. We're often taking more examples, more kinds of tasks, gives the algorithm, for lack of a better word, a better mental model of the world. So for example, and I wish I could off the top of my head recall exactly this research, but I remember from a few months ago a research report coming out around a language model improving once that had been trained on a visual task. So this ties into the point you were making about transformers being useful for so many different kinds of data types. And so the algorithm is able to represent information and encode across these different modalities so that if you, in your training data, you have lots of spatial examples of layouts of rooms and it provides better context, just as you might imagine it would for a human ... Don't take my ... It's not literally the same mechanism.

01:02:00 But to give a human analogy, if you looked at a bunch of drawings of the layouts of a building and then somebody asked you a language question about how would you get from the front porch to the back porch or whatever. How would you get from the front porch to the living room? And if you'd studied these kinds of drawings, you'd be able to do that mapping from the spatial reasoning to

linguistics. You'd be able to express it verbally. So yeah, I don't know. So this positive transfer, which seems to be happening more recently with the GPT architectures, I don't know, I think it's potentially an interesting breakthrough. There's probably some limits on how much-

Kim Stachenfeld: 01:02:45 Yeah. I mean, I think basically more data is a blessing and a curse. You have to have ... You can potentially do more. It has more information in it but it's also more burdensome to handle. The example that came to mind just in this conversation is if I bought more clothes, would I necessarily be better dressed? Maybe to a point I would have better outfits to choose from. But at a certain point I would be a crazy hoarder who could barely claw my way through my sweaters to get to my dresses or something.

Jon Krohn: 01:03:22 Ooh. Yes. I love that.

Kim Stachenfeld: 01:03:23 Once you have more-

Jon Krohn: 01:03:24 That's a good analogy.

Kim Stachenfeld: 01:03:24 If you have more information, you have to figure out how to organize it properly so you're not getting the compression artifacts of maybe using ... putting too much information in the same spot. The forced analogy here would be like if I had crammed too many outfits into one closet, I could not [inaudible 01:03:41]-

Jon Krohn: 01:03:41 Yeah. Just every day, you always use whichever one just happens to be on top. Because the closet is so stuffed.

Kim Stachenfeld: 01:03:46 Or maybe, and the analogy breaks here, but I can't ... There's no equivalent of an average of three scarves, none of which is really perfectly appropriate for the occasion. You can't average clothes. But that everything blurs

together is not really what you want. So more data is not always a benefit unless you really can organize it and know what to do with it and know how to find it when you need it. This is something that's magnificent about the brain and studying the hippocampus. That's the brain area that I focus on. It's involved in memory and trying ... And it seems like it plays an especial role in helping us organize new experiences, which seems like such an important function for how we cope with the gigantic dataset that is our lives.

Jon Krohn: 01:04:31 Yeah. Yeah. I mean, so this has been ... I had topics planned and we-

Kim Stachenfeld: 01:04:36 Yeah, we can veer back.

Jon Krohn: 01:04:37 -literally we've gone from the first question and we have ... So regular listeners will know that we have a brilliant researcher named Serg Masis, who probably 95% of the time when I ask a great question on the show, it's probably his question. And so he comes up with these amazing topics and he digs so far into your research and we asked the first question and then we've been going off forever since. So we did have stuff in here about the hippocampus, for example. So yeah, let's talk about that more. Let's talk about the hippocampus, why that's important. So I remember from my days as a neuroscience PhD student that, so the hippocampus, yes, critical for memory formation. And also interestingly, there seems to be some kind of spatial thing going on there where, for example, famously ... This probably isn't the case anymore because of GPS rotting everyone's brains, but it was the case some time ago, probably up until a decade ago, that to be a black cab driver in London, you had to pass ... I can't remember what it's called.

Kim Stachenfeld: 01:05:53 An incredibly extensive taxi driving exam. You have to memorize every street in London. Streets in London also are not easy to memorize. There's almost a rule against right angles and so nothing makes any sense and there's just tons of street names.

Jon Krohn: 01:06:08 Exactly. So it's super, super crazy. And so they have ... They anatomically, in brain scans, in not functional MRI scans, which ... So with functional magnetic resonance imaging, you get a sense of what parts of the brain are being active at different times. And so that's typically what we use to get a sense of brain activity and what's important. But in this case, just an anatomical scan, a static scan of these cab driver's brains showed that their hippocampus was bigger than average.

Kim Stachenfeld: 01:06:40 Yeah. So hippocampus has ... the two things it's studied for the most are its role in episodic memory, our memory for experience, and its role in spatial navigation. And there's a tight tie between these. There's really many ways that our model of space and memory might interact with each other. One is just that if I want to navigate around a city, I need to remember where stuff in the city is. That's just basically a memory problem. Another aspect of how memory and spatial representations interact is that as I'm laying down new memories, their spatial context might be really important. The fact that I experienced something at one time in one location, that location is really important for understanding ... for organizing my memories, for organizing my experiences, for knowing what situations that will be relevant to again. Next time I'm in that room, maybe that's when I want to remember those things. Or if I'm thinking about that room, other things that happen in that room, those are memories you might want to stitch together.

01:07:41 So hippocampus has been most studied in these contexts. The taxi driver example, super wonderful, super famous

experiment. Hippocampus, one thing that's really unique about this area is you have a significant amount of adult neurogenesis there. What that means, the fancy ... Yeah, this is neuroscience speak for new neurons get born. In most parts of the brain, you don't get new neurons. If you recover from a stroke, you've made new connections, but you didn't make any new neurons. But hippocampus, you can make new neurons. So that could be the reason that hippocampus ... maybe hippocampus basically swells up with new neurons when you study for the taxi exam. Maybe just people with big hippocampus are more likely to pass the test. It's a little hard to de-confound, but yeah, it grows. It's how you acquire new information and it grows through your life. It's pretty cool.

Jon Krohn: 01:08:27 And neurogenesis is really fascinating thing because for decades it was assumed, because in most of the brain, it turns out to still, as far as we know today, in most of the brain, like you're saying, a stroke patient, you don't see just because there was loss of brain that some new brain forms. There's just a few places where we seem to have new neurons born, new brain cells born. The hippocampus is one of them. The noses, the other one.

Kim Stachenfeld: 01:08:52 Oh yeah, the olfactory bulb. Yeah.

Jon Krohn: 01:08:53 Olfactory bulb.

Kim Stachenfeld: 01:08:55 Yeah, it's weird. Yeah. The olfactory bulb actually has a ton of circuitry in common with the hippocampus, which is so strange.

Jon Krohn: 01:09:02 Really?

Kim Stachenfeld: 01:09:03 Yeah, so I mean, one thing that's really unusual about smells compared to other sensations is just its geometry is pretty unique. If I am looking at something in my visual field, maybe it's over, I guess you guys can't hear if you're

not watching the video, but maybe it's over here, it's at some XY coordinate in your visual field. And it can move continuously through your visual field. If I'm listening to a pitch, it can rise continuously or fall like [inaudible 01:09:31]. Smells are much more discreet. The way our ... It's a particular molecule, it's different from other molecules. It activates a particular receptor. It's very specific.

Jon Krohn: 01:09:40 It's not on a continuum.

Kim Stachenfeld: 01:09:41 Yeah, yeah, exactly. It's not on a continuum. It's very specific.

Jon Krohn: 01:09:44 Categorical classification algorithm needed for [inaudible 01:09:47].

Kim Stachenfeld: 01:09:47 Exactly. And that seems to be related to how we represent memories in hippocampus too, that the role of hippocampus as a memory system is to represent what's unique and specific about a particular experience. And so this very sparse, non-overlapping organization seems to be really related to its ability to keep that aspect of experience, what's unique and special and different from other things and not necessarily bleeding in with all of the other similar experiences you've had.

Jon Krohn: 01:10:20 I had not known that. That is really fascinating.

Kim Stachenfeld: 01:10:23 Yeah, it's cool.

Jon Krohn: 01:10:24 I don't want to take up too much of the podcast episode [inaudible 01:10:27] time because the audience would probably much rather be hearing new things from you. But I did ... So going back to ... I was describing earlier in this episode in our first year of, so with Neil Rabinowitz, when we were at Oxford together, before you went off and did your PhD on your specific project for many years, you

did a one-year master's. So it's what they call in the UK a one plus three program or whatever. One year master's, three year PhD. Maybe in my case, the PhD dragged on a little longer than three, which happens.

Kim Stachenfeld: 01:10:56 I think that's more the norm than the exception in my experience.

Jon Krohn: 01:11:01 But in that one year master's, we had to do, in addition to the big research projects, we had papers on more discrete topics and I did one on neurogenesis. And I thought it would be really fun to frame it as a ... I framed this, what was supposed to be serious academic work, as a self-help article of, "How to grow more brain cells." And I was like, there's exercise. Exercise was one of the biggest ones for-

Kim Stachenfeld: 01:11:30 Yes, exercise is huge for neurogenesis. Yeah, I think exercise ... there's a lot of stuff on exercise and dopamine. I think for whatever reason, even though people often don't report finding exercise rewarding, it seems to activate a lot of the reward circuits, the circuits involved in motivation and continuing to experience. And those circuits in general also promote a lot of neurogenesis. Neurons that are born around the time you also experienced some dopamine, they're more likely to stick around.

Jon Krohn: 01:12:00 Yeah, yeah. There's a really small tangent off of this. Have you ever heard of type one fun versus type two fun?

Kim Stachenfeld: 01:12:05 Yes. Yeah, it was a big theme in grad school. I don't know if that's where you heard about it too.

Jon Krohn: 01:12:12 Yeah, so I was recently on the Ken's Nearest Neighbors podcast talking about ... I was a guest on his show, Ken Jee, he's a big YouTuber. He has 250,000 subscribers on YouTube and he's a data scientist ... Ken's obviously ... The show's called Ken's Nearest Neighbors, so it would

have to be machine learning. And he hadn't heard of type one or type two fun, so I was like, well, maybe it's not very universal. I don't know ... So I'm delighted that ... So yeah, so it's interesting. Exercise is classic type two fun, where as you're doing it, it rarely is enjoyable for its own sake. Whereas alcohol, drugs, sex, these are type one fun. They're just intrinsic. You just dopamine and serotonin explosions.

Kim Stachenfeld: 01:12:56 It's not complicated to enjoy them. I think another term for it is extrinsic reward as opposed to intrinsic reward. Intrinsic reward, intrinsic motivation, satisfying a sense of curiosity. You don't get paid for being curious necessarily but you could learn something that's going to be useful down the line. It's an investment, whereas extrinsic reward, that's right away, that's type one fun.

Jon Krohn: 01:13:20 Yeah, I guess it's interesting that maybe somehow, just tying on that point there about the hippo- Not the ... [inaudible 01:13:28]-

Kim Stachenfeld: 01:13:28 Dopamine neurogenesis.

Jon Krohn: 01:13:29 Dopamine neurogenesis. Yeah.

Kim Stachenfeld: 01:13:29 Exercise. Yeah.

Jon Krohn: 01:13:30 And exercise. Yeah, that even though it's not rewarding at the time, it seems to tie into that type two fun idea. You're getting a type two reward.

Kim Stachenfeld: 01:13:38 Yeah. Yeah. I mean, I've wondered this too. I think there's also just motor behaviors in general. Moving around seems to involve a lot of dopamine and so does rewarding stuff and I think this is still a pretty active area of research is understanding why those things both converge on the same system. If it's just about learning to do more behaviors, maybe when you exercise you end up

learning a little bit and those same circuits are activated. I don't know. It's a whole ... It could be its own podcast.

Jon Krohn: 01:14:08 Tying back to the not going to the office thing, that's impoverishing my moving around reward. It's wild. Probably for my dog too. It's like he's just off camera [inaudible 01:14:20] recording and I feel so bad for him that it's like we're in this apartment most days, all day. He gets a brief walk and I'm like, this has got to be so bad for both of us and you're reaffirming that now with the neuroscience.

Kim Stachenfeld: 01:14:32 I just looked at him and he's passed out with his tongue out. I think he just woke up, so he might be all right with his dopamine sparsity.

Jon Krohn: 01:14:41 It comes and goes. It's nice. It is nice to be able to have that midday nap that was always awkward in the office but I would do anyway. So yeah, so we've talked about simulated intelligence, we've talked about physical simulation. So let's get into reinforcement learning. So your PhD at Princeton, your dissertation was about learning neural representations that support efficient reinforcement learning. So maybe you could give us a quick introduction to reinforcement learning in general for our audience members who aren't aware of it. Although I've had ... For people who want a deep dive, I've done entire episodes just explaining what reinforcement learning is. I think it's episode ... I don't know, I'll put it in the show notes. Oh, here it is. It's episode number 510 was specifically on how reinforcement learning works. But yeah, you could give an explanation for our audience and then tie together what it meant in the broader sense for your PhD.

Kim Stachenfeld: 01:15:48 Yeah, I can give a quick summary. I mean, so basically reinforcement learning is learning from trial and error and just seeing what the outcomes are and how good it is

and then repeating things that led to reward. The classic example is if you're training a dog, then you can't tell the dog what you want. The dog doesn't really have any intrinsic desire to sit or stay,. If anything, quite the opposite. But if you give the dog a treat whenever it sits, and after hearing the word sit, you'll gradually modify the dog's behavior in order to sit when it hears that sound, just in order to maximize the probability of a treat. So this style of you get a treat if you do a certain behavior, you're more likely to repeat that behavior, that's essentially reinforcement learning.

Jon Krohn: 01:16:36 That was a great way of explaining it.

Kim Stachenfeld: 01:16:37 Thanks. Yeah, I think as somebody who's pretty food motivated, I think that example really relates to ... I really empathize with that. I think one thing, the counter to that, I think that intuition makes reinforcement learning seem really warm and loving and friendly. Like, "Oh, you did such a good job. You get a reward, a gold star, a piece of chocolate, a little dog bone or whatever." And there's an aspect to reinforcement learning that I think is a bit more brutal too, especially when you think about it in the context of human learning. If I was trying to train you to learn biology, if I was a biology teacher and I had a duty to do that, there's a bunch of different ways I could do that.

01:17:17 I could give you lots of textbooks and lots of material, and then you could try and train yourself to identify patterns or predict the next word in a biology textbook or some kind of more pattern based process. You could try to just take a bunch of biology tests and try answers out and then see what the actual answers were and modify your understanding of biology to maximize it. These are both unsupervised or supervised learning. The reinforcement learning version of this would be if I had you take a biology test, and I didn't tell you what you got right or

wrong, I just told you what your score was at the end. And I didn't tell you what the right answers were, I was just like, "You got 35."

01:17:56 And you were like, "Is that out of a hundred? Is that out of a thousand? Is that out of 35? Was that good or bad?" "Well, 35." Later you take another test, "You get 37." "Why? What helped?" "I don't know. You're just getting point totals." And you have to reverse engineer a pattern of behavior from this extremely sparse feedback. It's almost like, for very large scale problems, it's nice because it sets up this very general learning problem that's quite representative of the challenge of how do you autonomously reason about the world. But it also is very sparse in the information that it gives you and the structure that it has. It can feel almost a little bit passive-aggressive, just like, "35, what do I do with that?"

Jon Krohn: 01:18:35 Yeah, that was a great analogy. So yeah, that's the most interesting explanation of reinforcement I think I've ever heard. And it ties really well into my intro on you as we started our conversation about you just having incredible talks. So yeah, so how is this, in neuroscience dissertation, for example, is a question. So how was studying, learning neural representations that support efficient reinforcement learning, how is that in neuroscience PhD?

Kim Stachenfeld: 01:19:09 Yeah, so the reinforcement learning problem, the role that that serves in neuroscience is it's a model of how humans and animals learn from reward. They learn to do something that increases the amount of reward they get, whether that's food or just some kind of abstract sense of survival. And as my second example is intended to illustrate, actually doing this in an efficient way and in an attractable way is really challenging. If you're just stumbling around in the world trying to identify patterns in what leads to reward, you will be stumbling for a really

long time and you might stumble into some really bad situations that are not ideal for your survival, like off a cliff, into traffic. You don't want to learn everything in a way that's strictly bound to reward. So what we did with the subject of my PhD dissertation was trying to understand how we organize experience in order to identify useful candidates for reward. So how do you build a model of your environment, a representation of your experiences and how they relate to each other?

01:20:19 And then use that as some scaffolding, so you're coming to the reinforcement learning problem with a warm start. In the context of navigation, I don't want to be just stumbling around a maze for a while, not learning anything until the first time I find a tasty treat or a bite of cheese or something. I want to be learning about the structure of this environment and how I can go down different channels and how I can get back to the start if I want to. That way when I do finally experience some reward, I can link that to all the other places I've been and learned about. So this relates to the hippocampus, an area that seems to have a lot to do with our representations of the environment, what we do with new memories that we experience. So the topic of my dissertation was what kind of representation of these experiences that we have is going to be the best organization or represent the most concise statistics for a downstream reinforcement learning agent?

Jon Krohn: 01:21:14 Very interesting. So these ... You're a theoretical scientist and so this ties into this idea of ... These insights that you were making then, they also were insights that are applicable to both machine learning and biological organisms learning it sounds like.

Kim Stachenfeld: 01:21:42 Yeah, yeah. So in this particular case, the model we were using, the whole literature on representation learning that we were appealing to was a literature that came out of

machine learning. From this problem of how should an artificial agent represent its environment to make downstream learning processes more efficient? The particular model that we used in this was something called the successor representation. A paper from Peter Dayan in the nineties first introduced this idea that you can represent information about what's going to happen in the future. And if you compactly summarize predictions about what's going to happen in the future, this has a lot of information that's also likely to be relevant to predicting how much reward you can get in the future. I think the simple intuition is if I approximately know what's going to happen and where I'm going to go, and then later I find out one possible place I might go is rewarding, then I know how much reward I'll get at that location. It sets you up to do a whole suite of prediction problems, one of which is predicting reward.

Jon Krohn: 01:22:44 Right. And so you just said location again there and that ties into ... So a lot of your most cited research is about cognitive maps in the hippocampus and entorhinal cortex, which we haven't talked about. I don't know if you want to talk about entorhinal cortex.

Kim Stachenfeld: 01:22:58 Entorhinal cortex. Well, we'll see. We'll see if it comes up organically.

Jon Krohn: 01:23:03 So yeah, so it seems like we're getting this clear picture here, or I'm starting to get this clear picture, that ... So my next question for you, thanks to Serg, is what makes the hippocampus a map? And so I have this vague memory of you being able to say ... I can't remember exactly how it was measured but you could have the sense that when a rat learns a maze that actually the two-dimensional shape of that maze is represented roughly two dimensionally in the same layout in their hippocampus.

Kim Stachenfeld: 01:23:42 Yeah. So this is pretty ... Many decades ago in the forties, this cognitive psychologist, Tolman, introduced this idea of a cognitive map, like an early way of discussing mental models in a way. And he used it to describe the reasoning behaviors that he observed in rodents and in cats when they were placed in a maze or a box and then had to figure out their way out. And in contrast to the dominant view at the time that animals really just learned associations between observations and rewards, he described this process where it looked like they were building some sort of cognitive map that could be reasoned about and iterated over and different paths through it could be explored. So this was a dominant metaphor for thinking about mental models in the brain that are map-like. And then later on, John O'Keefe was the first person to report place cells. He and the Mosers, who also did some experiments with grid cells, got the Nobel Prize for this a few years ago.

Jon Krohn: 01:24:50 Grid cells. Yeah.

Kim Stachenfeld: 01:24:51 Yeah. Yeah. I mean, just a huge body of work has blossomed around these discoveries. But it basically seems like this abstract idea of a cognitive map had a really literal correlate. And that if you record from the brain of a rodent as it's running around, a rat or a mouse, if you record from its hippocampus, you find these cells that they call place cells that fire for particular locations. So if you have a little rat running around this table, whenever it's in one corner, you'll have some place cells that care about that corner will fire and you different cells will care about different locations in space.

Jon Krohn: 01:25:27 And those are in the hippocampus?

Kim Stachenfeld: 01:25:30 Those are in the hippocampus. Yeah. And collectively the entire population of neurons will all code different locations and comprise this whole map of space. So this

is where this idea that hippocampus is a cognitive map comes from. And there's a lot of other stuff that hippocampus encodes too. One really interesting study, Dmitriy Aronov did this study during his postdoc, is that if you have tasks that aren't spatial, in this case it was an auditory task where the animals heard a rising tone and they had to release ... It was basically like classic psychology. Rat presses a lever, then releases it at some point. If it has to release the lever at a target pitch instead of go to a target goal location, you see the same kind of cells, but they care about pitch rather than space. So it seems potentially like this is a more general memory area than just being a map. But it certainly has a lot of spatial receptivity. It encodes a lot of dimensions of space.

Jon Krohn: 01:26:31 Yeah. It sounds like ... It's something that I hadn't really thought of before that has come up many times in this conversation, though maybe not quite so directly, is what I'm about to ask, is that it seems clear that our spatial understanding of the world seems to relate to so many other kinds of memories. It seems like so much of ... Yeah, I mean, several times in the conversation now you've brought up how memory formation in general it seems like often has this spatial component.

Kim Stachenfeld: 01:27:07 Yeah, yeah. It's really fascinating. I think a big part of ... A big thing that we wanted to do in the modeling projects that I worked on in my PhD was try to have a mathematical model that didn't assume spatial structure as a given. So we formulated it in the classic RL way, which is basically instead of specific locations in space, you have states. And those states could be locations in space, they could be more generally the state of having eaten that day or not, or the state of a particular tone you're listening to, some other aspect of experience. And then have ... You can reason about relations between locations and space. You could reason about relations between different states you're experiencing too. And the

reason for this is that memory and space seem to relate in a pretty complex way. And a lot of the literature on hippocampus is on general memory formation. A lot of it is specifically on representations of space. And we wanted to try and use a modeling framework that was a bit agnostic, that could apply equally well to spatial situations, but also other kinds of memory structures you might be reasoning about, other sorts of associative systems that you might want to be navigating.

Jon Krohn: 01:28:26 Nice. Yeah, super interesting. And so this ties into ... So a week ago, at the time of recording, I posted on LinkedIn that I would be interviewing you on the show. There were tons of reactions and we had a question from Raju Basumatary who's based in Toronto, and Raju said, "I had a question for you which was based on neuroscience research and training machine learning models. What advice do you have for everyday folks to build their faculties?" And it sounds like we have a practical answer here. It sounds like taking advantage of this spatial relationship in learning, there might be something about, if you go to a new place to learn something new, it might be easier to remember that new thing because you can associate it with that space in your mind. Does that seem like a reasonable-

Kim Stachenfeld: 01:29:22 Yeah, absolutely. I mean, so I think I should probably direct this, in answering this question, appeal more to the specific topics I study. I think that the first things that come to mind is probably just exercise and getting a lot of sleep are the single best things you can do for your brain. But yeah, I think a key ... So in Alzheimer's, for instance, or in developing as the brain ages and loses some of its cognitive faculties, one of the main signatures of this is that you get a worse coordination between hippocampus and prefrontal cortex. The hippocampus isn't talking to the rest of the brain quite as well. And some of the things that have been shown to delay this are exercise and sleep

is supposed to help a lot. And other things are just continuing to cognitively stretch your brain in new ways, like exploring novel environments, probably something in that category. Crosswords is something people talk about. Continue search-

Jon Krohn: 01:30:21 Yeah, new skills.

Kim Stachenfeld: 01:30:22 Yeah, exactly. Continue searching your memory, continue doing new things, continue to mix and match ideas and experiences in new ways. I think also just trying to do things, this almost sounds too positive to be true, but I think doing things that you enjoy or find salient or stimulating in some way, those are the kinds of things that wash hippocampus with dopamine. Get a bunch of new neurons to stick around. That salient input is really useful for having a healthy hippocampus. In fact, in depression hippocampal volume tends to go down. The hypothesis at least, being that you just have fewer joyful or otherwise salient things breaking through and exciting hippocampus.

Jon Krohn: 01:31:11 Whoa, I didn't know that.

Kim Stachenfeld: 01:31:15 A sad way to think about it.

Jon Krohn: 01:31:18 Yeah, so you got those London cab drivers with their giant hippocampi and the sad, depressed people with their shriveled little ones.

Kim Stachenfeld: 01:31:25 Just in a sheer state of bliss.

Jon Krohn: 01:31:27 Don't ask them for directions, if you have a depressed friend, don't ask them for directions. That is not neuroscience advice that you heard on the show. Don't treat on that, that's not financial advice. Okay, so back to the script here a bit. Thank you for answering that audience question. Yeah, so we talked about dopamine

and serotonin. These are neurotransmitters, brain molecules that give us a feeling of positivity. In fact, we might be able to say that we actually don't enjoy anything in the world except dopamine and serotonin.

Kim Stachenfeld: 01:32:12 Yeah, I don't have any evidence to the contrary. I guess epinephrin, adrenaline seems to have its perks, but maybe just through dopamine, who knows? Anyway, I'll answer your question more specifically.

Jon Krohn: 01:32:23 But yeah, so there's this reward-prediction error hypothesis, RPEH, which is a leading theory in neuroscience related to dopamine, you know a lot about it. I think it ties into the reinforcement learning stuff that we've been talking about recently. Tell us about it.

Kim Stachenfeld: 01:32:44 Yeah, so the reward-prediction error hypothesis. So I guess first I'll introduce a reward-prediction error. This is a big deal in reinforcement learning. Basically the idea is that as you go through life, you are making expectations about how much reward you're going to get, how rewarding different outcomes will be. And you're tracking this because you want to build a good model of how rewarding different actions are, how rewarding different states can be, so that you can take actions that maximize them. And a lot of these, the algorithms that do reinforcement learning, a key substrate of them is this thing called a reward-prediction error. And this is the difference between how much reward you actually get and how much reward you predicted. And then when you have this surprise signal, this reward-prediction error, you update your expectations. If it was a positive one, you say, "Okay, that actually went a bit better than expected. I'm going to be more optimistic next time." If it was a negative prediction error, you will be more pessimistic and anticipate less reward in the future.

01:33:44 So this is simple learning rule, just make expectations, see how they go, update accordingly. And it ends up being a pretty powerful thing that you can really scale up to some large scale machine learning systems. As it relates to dopamine, there's a hypothesis that the activity of dopamine neurons is encoding a reward-prediction error. And there's some experiments, I think that the original experiment on this was Schultz, Diane and Montague, I want to say. And they recorded dopamine neurons in a task where an animal was getting reward, sometimes predicted by a cue, sometimes not. And what they found was that the firing of these neurons corresponded really well to what you would expect if they were encoding a prediction error. I could go into more.

Jon Krohn: 01:34:39 Go for it, yeah.

Kim Stachenfeld: 01:34:39 Yeah, okay. So basically the setup was the animal is sitting there and then sometimes it gets a little bit of juice, and juice is very rewarding. So you'd see a little firing of dopamine neurons whenever the animal got the juice. Then they set up something called a Pavlovian conditioning setting. And what this basically means is the animal sitting there, it hears a tone, and then a few seconds later it gets a bit of juice. So after a little while it picks up on this pattern, and whenever it hears the tone, it's like, "Ah, I'm going to get some juice soon. It's going to be great." What they see in that case is that when they hear the tone, they get the dopamine firing. That tone is predictive of reward.

01:35:17 They were just sitting there not knowing what was going to happen and then boom, something that's indicative of future reward happens, lots of dopamine. Later on when the reward actually happened, no dopamine because the dopamine wasn't surprising anymore. So this is consistent with this idea that dopamine is signaling unexpected reward. They also found that if they didn't get

the reward later on, dopamine neurons went far below their baseline rate of firing. So instead of just chirping along at their average rate, they would actually go quiet for a second as if saying like, "Hey, where's my reward? I thought I had predicted some reward happening." So that's foundational reward-prediction error, hypothesis story.

Jon Krohn: 01:35:59 That's so interesting because everything you just described, I can appreciate that subjectively as an experience that I have regularly.

Kim Stachenfeld: 01:36:12 Yeah.

Jon Krohn: 01:36:12 It's like you're constantly on the lookout for something new and surprising. And I guess that's how things like the TikTok algorithm have been so touted, I avoid TikTok. I have a TikTok account that I don't think it's really taking off we thought it would. We post clips, 30 seconds, 60 second clips from episodes, of concrete bits of conversation from these episodes on TikTok. And sometimes I know people who are like, "Yeah, I just got 2 million views." And I'm like, "I got 20." I don't know, I don't know. For some reason the SuperDataScience isn't taking off in the TikTok algorithm, but maybe it will. Maybe this is the conversation that will.

Kim Stachenfeld: 01:36:55 Yeah, I'm terrified of TikTok and I think because I'm also terrified of it just hacking my rewards [inaudible 01:37:01].

Jon Krohn: 01:37:01 Yeah, exactly. So it seems to do a really good job of giving you new surprising things that you're actually... So it's a new level of product manager warfare against human minds and hacking our brains and being able to continually surprise us and delight us with unexpected new things, because that is what we're looking for. So yeah, so tying back to this, I have, and I'm sure all of our

listeners have had this experience of something that you know should enjoy, but because it happened exactly as you expected, there was no surprise there. It's just like baseline, no extra dopamine, no extra good feeling. But if that thing that you expected and then routinely happens is taken away from you, then you experience sadness.

Kim Stachenfeld: 01:37:56 Yeah.

Jon Krohn: 01:37:57 Like [inaudible 01:37:57], jonesing for some dopamine.

Kim Stachenfeld: 01:37:58 Yeah. Yeah, I think it's easy to anthropomorphize specifically... Which maybe is fine. I mean, it's in the brain, but it's easy to really empathize with dopamine neurons.

Jon Krohn: 01:38:11 The other really important thing that your animals drinking juice story reminded me of, this might be a uniquely Canadian thing, but we had this song, this camp song, (singing).

Kim Stachenfeld: 01:38:20 We didn't have that in New Jersey, that's cute though.

Jon Krohn: 01:38:30 Probably inspired by that reason. Nice, yeah. So we've covered most of the topics generally that we wanted to cover even if we didn't get to dig into all of the wonderful questions that we prepared. But I think we've had a great conversation in around these topics. Anyway, the last big technical topic area to go over is this idea of compression.

Kim Stachenfeld: 01:38:58 Yeah.

Jon Krohn: 01:38:59 So in papers you've discussed how the brain compresses representations for future planning and previous trajectories, so-called Hippocampal Replay. And a recent paper called Language Modeling Is Compression by DeepMind and Meta explores the relationship between prediction capabilities and compression. So your work

emphasizes predictions and compression as key elements in representation learning. So I feel like maybe we haven't even talked about... Maybe we need to define representation learning, which we haven't probably done specifically. But then yeah, I'm sure you have lots of interesting things to say related to this compression and how it relates to both the brain and artificial systems.

Kim Stachenfeld: 01:39:45 Yeah, so representation learning generally is this problem of learning how to represent your observations. You have all of these things that you experience. What is the format that this experience should take? If I think about this current setting I'm in right now, I'm sitting at this table, I'm mic'd in, I'm in the middle of speaking, these are all aspects of my experience that I want to represent in some way, that have some relevance to what I'm about to do. And I don't want to get distracted by irrelevant things. If I'm in the middle of a sentence, I really want to remember the beginning of that sentence, but maybe I don't need to be aware of the cool strap on that guitar with its bird pattern. That's neat and I could access that knowledge in different situation, but being able to focus on specifically relevant information, not getting distracted by other things is really important.

01:40:39 The flip side of this is if I wait for experience to justify every bit of information before I learn anything, I just haven't constructed enough of a map. I haven't really been using my knowledge or my experience in organizing it in a way that can support any information. So representation, learning is like what sort of objective should you use to learn a representation? What information is relevant? What is irrelevant? What should I keep around? How do I represent it in a way that is both expressive enough and compact enough to support intelligent but also efficient behavior? So prediction and compression are two objectives really that come up in this. Prediction when cast as an objective is saying, "I

want to be as good as possible at predicting what's going to happen next. I want to be able to make predictions about what's going to happen, because whatever is relevant to predicting things in the future, that's probably something that's not just random noise. It's something that I'll at least persist in time for a little bit. Maybe that's an outcome I want to maximize at some point." Compression is more about how do I represent things succinctly? How do I try to have short descriptions of what's going to happen? It's a way to summarize events rather than expressing them in their full detail. This can be useful for efficiency. You just have fewer things that you have to learn about if you have a more compressed representation. It also can be really useful for identifying abstractions. One of the things that emerges when you apply compression, when you try to make your descriptions be more succinct, is that you get more overlap between related ideas. And this is really useful as a way to start getting some elements of abstraction out. So for instance, if I have lots of pictures of elephants that I've seen, I can represent them more compactly if I just have the concept of an elephant. And here are lots of specific instances of it, but I have this more general concept that unifies them all with a single description.

01:42:48 I might also have a compression or a compact representation of all of the things I've ever seen before that are pink. And there's maybe different kinds of instances, they manifest in different ways, but there's one property in common which is pink. And if I try to have a compressed representation, that's a concept that might emerge. And as it relates to abstraction and compositionally, once you have these things pull out and part of a compact summary, you can start reasoning about more interactions between them. I can now have a pink elephant, because pink is something that popped out and elephant is something that popped out, just different compression. So it's useful just for efficiency, but

it's also useful at a deep level for trying to extract patterns and rules in the world.

Jon Krohn: 01:43:28 So this all makes perfect sense. So representation learning is just this general idea that we are probably familiar with anecdotally, subjectively as individuals, where it's just the representation of information. And as we do that, we want to compress that information as succinctly as possible probably, in as many circumstances as possible so that reduces cognitive load. So we want these representations to be efficient, which allows us to have more thoughts and that allows us to hopefully be able to make predictions better, get more dopamine reward per unit of mental effort in the planning that we're doing. And then so you saying that, it seems almost trivially obvious that language modeling is compression, because it just seems extremely obvious that saying the word elephant and that being able to represent a big gray object with four legs and a nose and big ears, it's like you compress all of these ideas about an object into a few syllables, into one word elephant. And we make up words for theoretical concepts like corporation and love. And so you are able to compress huge complex things into a single word, and then that makes it easier to play around with these and have them bump into each other and think about a loving company or something. Which if you don't have these words, it could be difficult to think about these quite disparate concepts together.

Kim Stachenfeld: 01:45:21 Yeah, I think absolutely. I think one thing that was a cool aha moment for me was realizing that... I used to think of compression maybe as just something you had to do because you had limited resources. The brain is not infinite. We have working memory constraints. We have only so many neurons and only so much food to power them, so you have to compress stuff. But I think this idea that it actually has enormous computational benefits too,

that it's a deep part of the reasoning process where you identify commonalities between different situations can be expressed as compression was cool. That it's not just a compromise or something that you have to do in order to make it work, but it actually has real advantages.

Jon Krohn: 01:46:08 Nice. And so this ties back neatly to again, we got to talk about LLMs. Every episode has to be about LLMs. And so this interestingly ties into this idea of you describing tokens as being so widely useful and in large language models we're using tokens to represent tokens of language, sub words per token. And so this seems to naturally tie into a lot of stuff we were talking about a while ago in the episode where we were saying things like spatial representations could be stored in language tokens. You could describe a scene, this often happens in novels where it's describing a visual scene and you don't need pictures in the book to be able to imagine what that might be like. So yeah, so that all makes a lot of sense to me. Really quickly, I think we've covered probably enough technical content for this episode. This has been a long one. Thank you for being so generous with your time. But on top of that, you also just in terms of general interest stuff, you hold black belts in both a type of karate that I'm probably just going to butcher, Isshin-Ryū.

Kim Stachenfeld: 01:47:16 Isshin-Ryū.

Jon Krohn: 01:47:19 Isshin-Ryū karate and TaeKwonDo. So you have black belts in both this kind of karate and Taekwondo, so that must've required a lot of discipline and dedication, a lot of type two reward happening there. Do you think that doing that kind of stuff has been helpful in you being a rigorous academic and achieving the success you've had working at the greatest AI lab in the world, that kind of stuff?

Kim Stachenfeld: 01:47:49 I could wax lyrical on how much martial arts has been-

Jon Krohn: 01:47:54 You could wax on wax off [inaudible 01:47:57].

Kim Stachenfeld: 01:47:57 Yeah, I should have seen that one coming. Yeah, I mean honestly, it's just been foundational to almost every aspect of my adult personality. I've been doing martial arts for a really long time. I started when I was six years old at karate. My parents actually met at a karate school, so it was a thing they had been doing for a long time too. Yeah, I mean I think the main thing that I think about the most often in terms of how karate has informed my scientific life, is that my favorite activity in karate was always sparring. I just could not get enough sparring. Whenever it was my birthday and I got to pick the... We were always sparring and I think-

Jon Krohn: 01:48:42 "Do you want an ice cream cake?" "No, I want to punch you in the face."

Kim Stachenfeld: 01:48:50 Yeah, I mean more or less. And I think a thing that's really different about sparring or fighting, if you're part of a karate school or a Taekwondo club, in TV or in kung fu movies, in the Karate Kid, whenever there's fighting, there's often real animosity between the people fighting. It's a big tournament and the other guy is bad and beating him is good and the fighting corresponds to an actual animosity. But that's so not how it is. If you're actually in a karate school. You are sparring with your best friends and people you really like and people you don't actually want to see hurt or injured. And you're sparring just because it's fun, it's improvisational, you're building skills, you're helping them build skills. It's a much more constructive activity, but it's still really combative.

01:49:34 I think that's just a really excellent thing to get used to for scientific discourse. That you often have disagreements, you have different ideas, you're trying to sort them through, but you're in a community and you're trying to

figure out what's right. You're trying to be skeptical and sometimes adversarial, but in a way that's ultimately really constructive and pro-social. I think knowing the difference between fighting or combat or adversarial situations that are more constructive and ones that are not constructive is really useful. And being able to be adversarial in a way that is constructive is a thing that takes a little bit of practice. And I think that karate really helped me develop more of an intuition for that.

Jon Krohn: 01:50:19 That was such a beautiful way of describing it.

Kim Stachenfeld: 01:50:23 Thanks.

Jon Krohn: 01:50:23 Thank you so much, Kim. And yeah, one last thing here is for folks who want to be a computational neuroscientist or an AI researcher like you, what should they do? What skills should they hone other than karate?

Kim Stachenfeld: 01:50:41 Yeah, so okay, what skills for neuroscience or machine learning? Neuroscience is super interdisciplinary. I think one thing I was struck by when I first started in graduate school is one, there isn't really a standard core curriculum. Every single department has their own intro class and you learn different stuff somewhat just based on the whims of the professors in the department. And also everybody coming into my class had a pretty different experience from undergrad. Some people knew more about psychology, more about math, more about engineering, more about the really low level neuroscience or biology, there was just a massive spread. And I think basically neuroscience is just in this fairly early stage for a science where the fundamentals are still in development. Really the philosophy of the field is in flux.

01:51:30 It makes it a really, really exciting time to be in the field. There's a lot of turnover of different modeling ideas, constant introduction of new methods. I personally had

experience in math and engineering. The skills I found most useful are programming and really just being able to code stuff up as a way to sketch out or check intuitions about a mathematical model. And math is really useful just as a more persistent base of scaffolding ideas, that basically there's these ideas that aren't going to be changing with fads every second. And they help you think about concepts like representation and how to do different logical operations and execute computations in a general way. So that's been really useful to me. The philosophy also honestly, has been really useful just for thinking.

Jon Krohn: 01:52:25 Well, that was unexpected.

Kim Stachenfeld: 01:52:28 I know. I took some philosophy of mind courses in college and I still think about those ideals a lot. I think especially once you, they're not going to necessarily be the skills that help you get a job, but they will help you do the job properly and think about the implications. I think in general as neuroscience affects people pretty directly, people are the havers of brains. And AI is really out in the world doing stuff and questions of ethics and humanities and how these are actually going to affect the world and society. In a lot of ways, people have thought about them for a while and in a lot of ways people feel like they're laying down the tracks as the train is going. So having knowledge in those areas maybe aren't the most sought after for getting a job, but they will help you think about the implications of what you're doing a bit more thoughtfully.

Jon Krohn: 01:53:23 Really cool answer. Did not see the philosophy of mind part coming, but it makes so much sense and yeah, definitely something I would love to dig into more. It sounds like if I can twist Neil's arm to be on the show, then we'll have a theory of mind AI episode sometime in

the near future to check out. Before I let you go, book recommendation. Do you have anything for us?

Kim Stachenfeld: 01:53:46 Yeah, so a book I thought about a lot in really thinking about some of the answers to the simulation questions, what is the role of simulation? What kinds of physical processes are hard to model? Is this book, it's a physics book by Carlo Rovelli. He's a really cool guy. He's a physicist and he just writes these wonderful popular science books about physics. And this book, the Order of Time, it's about entropy, it's about complexity. It is just a really cool book. And he talks a lot about fluid dynamics and why it's such a hard to model system.

01:54:21 There's just a bunch of fun, mind-blowing facts in there. I think one of the things he says about fluid dynamics that I found particularly evocative and I think illustrates why simulation is so important in this domain, is that most of the problems that are the frontier of physics are things that are very far away from our experience. There are things that are microscopically small, massively cosmically huge, so hot that we would just explode if we touch them. There are things that are hard to look at. And fluid dynamics is not that way, but it's still really hard to capture because it's such a complex and sensitive phenomenon.

Jon Krohn: 01:55:00 So annoyingly fluid, those fluids.

Kim Stachenfeld: 01:55:01 It's very elusive.

Jon Krohn: 01:55:01 Just hold still.

Kim Stachenfeld: 01:55:02 Yeah, yeah, exactly. You can pour milk into your coffee cup and watch it swirl around, and that's the frontier of physical knowledge on that subject. It's both proximal but very mysterious and elusive. So he has some really poetic,

lovely examples and ways of seeing the world that I just think is fantastic.

Jon Krohn: 01:55:18 Nice. That sounds really cool. Great recommendation. And then how can people follow you after this episode? Clearly you're brilliant, I hope it won't be too long before we can get you on the show again, because we asked five or 10% of the questions that we had prepared, because there were just so many interesting things that I immediately thought of that I wanted to ask you about and naturally flowed with the conversation. So yeah, I really do hope to have you on again soon. But in the meantime, how can people follow you to get your thoughts?

Kim Stachenfeld: 01:55:46 So I'm on Twitter, my handle is @neuro_kim, and I also have a website, which is neurokim.com. This should be easy enough to remember, first syllable of my name and the word neuro.

Jon Krohn: 01:55:59 Perfect. Love it. We'll be sure to include those in the show notes. Kim, thank you so much for making the trip to record here in New York with me. And I had so much fun, I'm sure a lot of our listeners did as well.

Kim Stachenfeld: 01:56:13 Yeah, thank you so much for having me. This has been a genuine delight.

Jon Krohn: 01:56:15 It feels like we only just scratched the surface of Kim's tremendous knowledge and crystal clear analogies today. Hopefully we can get her back on the show soon to continue the conversation. In today's episode, Kim filled us in on how we can much more efficiently make predictions about the physical world using machine learning models, including learned simulations of mental simulations relative to classical simulations that try to capture all of the underlying physics. She also talked about how our brain's hippocampus is key for memory

formation and is a cognitive map of physical space. She filled us in on the best things for our cognitive abilities, including exercise, sleep, exploring new environments, learning new skills, crosswords in particular, and doing things you enjoy. She also talked about how the dopamine reward-prediction error hypothesis leads us to seek to have our expectations always exceeded. And simply having our expectations met can lead to a feeling of slight disappointment. And she talked about how sparring with her friends in martial arts cultivated her capacity for constructive scientific discourse.

01:57:23 As always, you can get all the show notes, including the transcript for this episode, the video recording, any materials mentioned on the show, as well as the URLs for Kim's social media profiles and my own at superdatascience.com/725. Beyond social media, we could also meet in person this Friday, October 27th at the Scale Up AI conference at which I'll be interviewing GitHub COO, Kyle Daigle live on stage. You can check it out in person in New York, or you can stream it online anywhere in the world. The conference is put on by Insight Partners, one of the world's largest hedge funds and is targeted at folks who are ready to scale AI businesses or scale up their business with AI. You can use my code JKAI35, that's in all caps, JKAI35 to get 35% off on your registration.

01:58:14 All right, thanks to my colleagues at Nebula for supporting me while I create content like the SuperDataScience episode for you. And thanks of course to Ivana, Mario, Natalie, Serg, Sylvia, Zara, and Kirill on the SuperDataScience Team for producing another extraordinary episode for us today. You can support this show by checking out our sponsor's links, by sharing, by reviewing, by subscribing, but most of all, just by continuing to tune 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. Until next time, keep on rocking it out there and I'm looking forward to enjoying another round of the Super Data Science Podcast with you very soon.