THE TRANSMITTER | BRAIN INSPIRED

Kim Stachenfeld on the dance between neuroscience and artificial intelligence

As a researcher at both Google DeepMind and Columbia University, Stachenfeld offers cross-disciplinary insight into how to understand the brain.

11 September 2024 | by PAUL MIDDLEBROOKS

This transcript has been lightly edited for clarity; it may contain errors due to the transcription process.

Kim Stachenfeld
Not only is neuroscience-inspired AI not really super what's going on, science-inspired AI is just still happening, and in lots of different areas, but it's—what neural networks do is they have several stages of processing, and at each one, it's a re-representation of their input. This question of what information they represent at each layer, the neural network gets to figure that out on its own. It's not like you fix it the way we did, but the question of what it does choose to represent has a big effect on what it can do downstream. Neuroscience just has this tremendous variety and diversity and eccentricity and stuff. I just love that.

[music]

Paul Middlebrooks
This is "Brain Inspired," powered by *The Transmitter*. Hey, everyone, it's Paul. You may have just caught that "Brain Inspired" is now powered by *The Transmitter*. That's right. I'm excited to announce a major milestone here. "Brain Inspired" is now a proud partner of *The Transmitter*. For those of you who haven't heard of it, *The Transmitter* is an online publication that provides information, insights, and tools to help neuroscientists at all career stages stay current and build connections. It's funded by the Simons Foundation, but it is editorially independent. I'm delighted to join their team.

What does this mean? "Brain Inspired" will stay the same, but I'll be contributing to and collaborating on various new projects in line with *The Transmitter*'s mission to spread the word about neuroscience. In fact, you can find this and future episodes on their website, thetransmitter.org, where you can also easily sign up for email alerts every time a new "Brain Inspired" episode is released. Actually, if you visit their newsletter page, which I'll link to in the show notes, you can customize what kinds of neuroscience news and topics and columns that you'll receive in your inbox. Trust me, there's a wide variety of options there. Moving forward, I'll also point to stories that grab my attention.

For example, recently I read a summary of a compelling study from Mehrdad Jazayeri's lab about how monkeys' brains build mental cognitive maps and use those maps to imagine things that they've never seen. I'll link to that in the show notes as well. This is really an exciting new partnership between "Brain Inspired" and *The Transmitter*, and I'm grateful for their support.

All right, now, today's episode. Kim Stachenfeld embodies the original core focus of this podcast, the exploration of the intersection between neuroscience and AI, now commonly known as neuro-AI. That is because Kim walks both lines. She's a senior research scientist at Google DeepMind, the AI company that sprang from neuroscience principles, and she also does research at the Center for Theoretical Neuroscience at Columbia University. She has been using her expertise in modeling and reinforcement learning and cognitive maps, for example, to help understand brains and to help improve AI. I've been wanting to have her on for a long time to get her broad perspective on AI and neuroscience.

One of the things that we discuss is the relative roles of industry and academia in pursuing various objectives related to understanding and building cognitive entities. She has studied the hippocampus in her research on reinforcement learning and cognitive maps. We discuss what the heck the hippocampus does, since it seems to be implicated in so many functions, and how she thinks of reinforcement learning these days. Most recently, Kim at DeepMind has focused on more practical engineering questions, using deep-learning models to predict things like chaotic turbulent flows, and even to help design things like bridges and airplanes.

We don't get into the specifics of that work, so I will link to it in the show notes. Given that I just spoke with Damian Kelty-Stephen, who thinks of brains partially as turbulent cascades, Kim and I discuss how her work on modeling turbulence has shaped her thoughts about brains. OK, so that was a lot, I know, but you can find all the details in the show notes at braininspired.co/podcast/193. Again, thank you to *The Transmitter*. This is really exciting for me. Thank you to my Patreon supporters for your continued support.

OK, here's Kim. I moderated a panel that you were on at COSYNE a few, what was it, a month ago, two months ago? I can't keep track of time. It struck me, you even mentioned at one point that you were saying something in service, not just to be defensive, I think was the quote. It was about DeepMind and how-- So the original mission of DeepMind was to use what we know about the brain to make better AI. That has sort of gone off the board because these days it's just, scaling up is what AI is all about these days. What I wanted to ask you is, is it fair to say that DeepMind failed?

Kim Stachenfeld
[laughs] I guess.

Paul Middlebrooks
OK, all right.

[laughter]

Kim Stachenfeld
I don't think so. I'm not sure. OK, yes, I guess there's a couple of things. One, the original characterization of DeepMind's mission. It was something a bit more circumspect than that. It was like, solve intelligence. Here are lots of different directions that we think might be viable. Neuroscience is one direction that DeepMind had that maybe it was unique and that other groups didn't. I think one, the neuroscience research projects were always pretty careful to pick projects that were in a particular zone of impact so that it could be useful to neuroscience and say something useful about neuroscience and had the possibility to say something useful or relevant to machine learning.

The kinds of questions that we focused on were things that we thought had some application to a particular open problem in machine learning. Continual learning, robotics, learning rules and optimization in general, structured credit assignment. We picked things that we wanted to learn something about the brain and were also open problems in machine learning. The other thing that the neuroscientists did at DeepMind--

Paul Middlebrooks
This is all past tense, I'm noticing. Go ahead.

Kim Stachenfeld
A lot has changed in the last year of machine learning. The neuroscience team has also shifted their focus quite a bit from the neuro lab. The other thing is that neuroscientists at DeepMind were, as Matt Botvinick, who headed the neuro team for a long time, put it, were like bilingual. They would work on machine-learning problems, publish machine-learning papers, also work on neuroscience problems, and part of the benefit of this was thought to be a more intellectual or abstract exchange. If you're trained in neuroscience, you just have different ways of thinking about problems.

A lot of cool stuff came out of the neuroscience team that had really very little to do with neuroscience. A lot of the stuff with graph neural networks and using them for simulation, that was a project I worked on for a while. Pretty tenuous connection to anything to do with neuroscience. It was a pure machine-learning project. There was a lot of stuff with concept learning, too, that was a big focus of the neuroscience team, not directly paired with a neuroscience investigation project, but just something that neuroscientists tend to think about.

As you mentioned, this is pretty past tense. I think the neuroscientists at DeepMind, we've largely been pivoting, for a couple of reasons. One is that we're in this, as I said in that panel, the whole field of machine learning is very much in a scale-up mode right now. It's like build it bigger, build it better. Just add more data, add more TPUs, and try to generate it for longer. This is very much the mode. This is for some good reasons. There's two papers that I think really encapsulate this nicely.

One was a paper from OpenAI. Now the guys who wrote it are at Anthropic, but it was on scaling laws and transformers. They basically showed you make the model bigger, you add more data, you add more compute, the model will get better predictably. If you are looking for a place to turn money into good machine-learning models, that's a really reliable knob to turn. That's the signal you want. The other paper was a paper from DeepMind showing that emergent properties start becoming apparent as you scale up a model. This is almost like the opposite, rather than just as you add more compute, the model will get better, this will be like, as you add more compute, the model doesn't get better until suddenly it does.

Paul Middlebrooks
It's different.

Kim Stachenfeld
Yes. It starts doing different things. Some abilities that it used to not have now emerge. Basically, in both ways that are predictable and not predictable, as you add more compute, more scale, the model seemed to get better. I'm a researcher. I'm not an engineer, so not my strong suit.

Paul Middlebrooks
You majored in chemical engineering, right?

Kim Stachenfeld
That's true, but that doesn't actually come up that much in my day-to-day. I guess this is just to say I can see why we're in a period that is more about more-- that we're engineering and increasing scale or having a moment rather than, let's investigate new methods. Let's try to search broadly and find inspiration from different areas of science and make interdisciplinary connections. Not only is neuroscience-inspired AI not really super what's going on, science-inspired AI is just still happening and in lots of different areas. There's a lot to be gained from scale right now. There's a lot to be gained just from making things better. It's just not really like the pendulum has swung a bit.

Paul Middlebrooks
Oh no, the pendulum again. This came up in the panel as well. I made one of these comments, like I stopped everything. I was like, the pendulum--

Kim Stachenfeld
It was an attractive metaphor because it makes it feel like it's coming back. The good news for neuroscience though, is there's a lot to be done there. I think that's the other thing is, there's just a lot of good opportunities for applying data. This is a moment for data-driven methods in doing stuff.

Paul Middlebrooks
But they're tools, right? They're not necessarily models of the system. People like Jim DiCarlo and David Sussillo, there's a lot of people using them as models, but what I'm hearing you say is that it's all tools.

Kim Stachenfeld
I think they're both. I think they're really both. You can interpolate in between. One aspect is to just use them for tools. Say, like, here's a method for getting patterns out of data that's complicated and hard to understand. The brain and behavior is filled with complicated data that's hard to understand. If we train models, maybe we can predict it well. Maybe we can summarize it well. Maybe we can decompose it into some form that is a little bit more-- From which humans can grok patterns. That's one opportunity, is just try to summarize. That's having a moment. There's a lot of methods that people have been working with.

Mackenzie Mathis has awesome work just summarizing behavior, **Bence Ölveczky**. Summarizing behavior, that's on the tool side and it really expands what you can do. I think on the models of the brain side, it's also pretty useful because if you train a model to do something, there's many ways you can structure it. They trade off interpretability and messiness in a nice way. For instance, if you train the same big, messy black-box model with different optimization algorithms, you get different representational properties that emerge. If you have different objectives on a deep neural network, you can say something systematic about what's happening inside of the representations.

This forces us to think a little bit more about like what features are happening, what kinds of cognitive operations are happening implicitly in a big, messy, deep neural network. Rather than structuring something is like, this is the module that does search. This is the module that does reinforcement learning. If you're like, we trained a big model to do all sorts of stuff and we're still looking for those cognitive functions, but we're trying to pull them out of implicit operations rather than structuring them deliberately.

I think it's probably a useful perspective and still counts as model-driven research. It's just that the specific variable that we're modulating is related through a messier system. This has been something that we've been thinking about a lot, is if we're thinking about how to use neural networks to comprise models of the brain, there are certain problems that only arise when you have a neural network system, or there's certain ways that you can combine structure and expressiveness and get in this sweet spot where you are getting something out of the neural network, but also not completely foregoing any ability to say something structured.

Paul Middlebrooks
One of the things, and I'm not going to just focus on the panel that I moderated, but I was revisiting it. One of the things I asked was whether any of the panelists thought that AI has actually hurt neuroscience in any way, and then no one answered. Then you were kind enough to-- You couched it in a different answer that someone else asked a question and you were kind enough to at least address the question that I asked. You mentioned something about biological plausibility or non-plausibility biologically.

I think that's interesting, and I want you to expand on that a little bit because I thought one of the lessons here, and this goes to what you were just talking about, was that biological plausibility doesn't matter. It was odd to hear you say that non-biological non-plausibility, non-biological plausibility, whatever it was, that was in some way how AI has maybe dampened neuroscience or hurt it. Am I making sense?

Kim Stachenfeld
Yes. It's funny because I was like, my research is not like a paragon of biological plausibility.

Paul Middlebrooks
I know that. Yes. We'll talk, because you know, turbulence.

Kim Stachenfeld
I think that there is a level of abstraction that models bring. I say I think, but I guess this is a common framing principle when thinking about model-driven neuroscience is all models are wrong. Some models are useful. Exactly what you're trying to say about the system, and whether or not your claims are justified, depends on the level of abstraction you're using.

For instance, if I say this effect appears to be driven by the statistics of data and I think that it is indifferent to which particular learning algorithm you used, then I can say, "Well, we used a biologically implausible one because we wanted to get the system actually working at the level where I can reason about these kinds of statistics. We couldn't do that with a biologically plausible one, so we did it that way." Somebody could say like, "Well, I don't agree with your claim that this is irrelevant to the biological implementation details. That could make a really huge difference." Then I would have said something really misleading.

I think the reason that occurred to me in that panel is because that's just what I worry about a lot, is I focus on a particular level of abstraction and try to say things that are justified at that level about how the statistics are having an effect, what kinds of computations are supported by different kinds of representations in a specific implementation agnostic way, but those details could really trickle up and matter. Then I would be saying things that are wrong, which of course nobody wants to do.

I don't know, for instance, that we have been critically led astray by that. That wasn't necessarily my instinct, but it would be, that's the risk is that you work at this level, and you were led astray because it didn't obey some constraints, which turned out to really make a big difference. I will say though that there's two levels of biological constraints at some level, the implementation details, but then there's the behavioral level. One thing is the models that relax the behavioral constraints, the advantage of them is that they have an easier time getting this behavioral level, or at least that's ideally the justifying case to use them.

It's not always totally fair to say they're biologically unrealistic because they are pinning to one aspect that you know the biological system is doing, which is doing a good job at some naturalistic behavior. I'm not the first to make that point, but it's almost like they're pinned to different levels of biological plausibility and no model gets both of them, rather than saying, we don't care about the biology at all.

Paul Middlebrooks
I don't know, are you half Columbia, half DeepMind? What is the correct--

Kim Stachenfeld
Basically, I'm at Columbia one day a week officially. I sometimes go up on Fridays too, because Google is hybrid. I'm usually there Mondays. My official appointment is adjunct associate professor. It has a lot of asterisks next to them. [laughter] I'm up there some fraction of the time, and then I'm at DeepMind the rest of the time. I do a little academic stuff, a little industry stuff.

Paul Middlebrooks
The reason why I asked that is because I'm curious, A, what's more fun, and B, how much time do you spend these days thinking about neuroscience versus applications like modeling turbulence and design?

Kim Stachenfeld
Yes. They're both tremendously fun. That's a boring answer. They're both really fun. [laughs] They're fun in different ways. The Columbia stuff, at least the way I have it structured because I'm part-time, is more advising on projects. That's fun because there can be tremendous variety. Because you can work on more stuff when it's students or postdocs who are driving the projects. Those projects also tend to mostly be more neuroscience-focused or theory-focused.

Some of the projects are more, how does the brain do this? Or here's a neural network. It's doing something like the brain. Does that actually compare? Some are more actually analyzing data that's come out of other labs, and one in particular, it's a machine-learning project, but it's a more theoretical and conceptual one. In general, the kinds of stuff that is maybe a bit easier to do in academia, not just possible to do in either place but actually easier in academia, is things that are with smaller models and that are maybe a little bit more conceptual, because Google just has--

Paul Middlebrooks
Like toy problems?

Kim Stachenfeld
Yes. Kind of toy. Yes, exactly. Things that are a bit more toy or at least don't require you to train a gigantic model. You can do stuff with gigantic models that somebody else trained, but you don't want to necessarily be in the business of training a huge model. That's not easier to do in academia at least. The kinds of projects that are fun to do at Google are largely things that benefit more from scale that have some-- like you can use a really big model. You can train a big model to do a different thing. You can really experiment with some of the more conceptual stuff at large scales.

This is coming up, in the context of-- I have been working recently on some projects with memory, with retrieval augmented generation. There are some neuroscience versions of this. We think about hippocampal contributions to learning all the time is like, what is memory adding to a process, and maybe making hypotheses about when you should see different kinds of hippocampal activity, different levels of activity, different styles of activity. That's a neuroscience question. You can often ask that with somewhat toy models. You can get abstract versions of that problem that you can look at in similar systems.

Whereas at DeepMind, you can play around with a large language model retrieving from Wikipedia. That's not impossible to look at in academia, but it's harder. There's infrastructure, and it's nice to be able to play around with the same concepts but at a scale where it's naturalistic and complex.

Paul Middlebrooks
I remember I was riding in the car with my father. I think I was in graduate school, actually, for neuroscience. He asked me or suggested that perhaps industry is better suited to, he didn't use the term solve intelligence, but "to understand brains and

stuff." What is your perspective on this? I think he made a very good point. Of course, it hurt a little bit because I was in graduate school, but he was an IBM guy. I think he made a pretty good point. I'm curious where you fall on that. Should industry just solve neuro?

Kim Stachenfeld
That's funny. I would be curious to ask you more about what your dad's like pros and cons were for academia and industry.

Paul Middlebrooks
OK. I can tell you real quick, basically just that we academics are super slow. There isn't a bottom line to get done because it's an endless search. In industry, there's a thing that you are setting out to accomplish, and I think that's part of it, and academics are just slow.

Kim Stachenfeld
I guess researchers in general are pretty motivated. I think it's more an industry because it's not like people are pulling longer hours in industry because they're getting paid more.

Paul Middlebrooks
No. They're happier. They go home earlier, and they're happier, it seems.

Kim Stachenfeld
Yes. I think the difference is because of this goal collapse in a way, that there's a financial bottom line or that there's an incentive that's sort of aligning people. You get more people working on the same thing and you have things that are broken into projects. I'm not an economist, so whatever, take this with some grain of salt, but my sense of it is that if industry is well poised to work on a problem, if a problem does align with the goals of a company, it's great to have that happen. There is an alignment of people who are getting funding, and they're going to work on the problem, and they'll break it down into parts that will be followed up on tightly with different kinds of progress management criteria.

I think a lot of those things work really well. The question is, of course, just like, is there an alignment of objectives? I think basically when there is, then it's great to have the problem pursued in industry. If there isn't, like some problems are really-- In some sense, there's not an obvious financial motive to understand how the brain works, except insomuch as you could design pharmaceuticals to help it or base technology on it or something. It's these very sideways scant views of it. I think the benefit of academia is you have more smaller groups of independent researchers all trying their own thing. It's not like a bunch of people are going to lock in and run in the same direction with quite as much ease.

The goal that people have is just the actual-- can be the same goal as understanding the brain. It doesn't have to align in these sideways ways. I think that's sort of the pros and cons, like when industry does it, awesome. We should be delighted. I don't worry like, "Oh, no, if industry does it, how will academia compete?" I'm like, "Oh, there's still plenty of things to do that aren't necessarily on the path to making better technology or better pharmaceuticals." We don't have, I think, an alternative to academic research to moving in those directions.

Paul Middlebrooks
I wish my dad was around to see the course that DeepMind has taken, just to circle it back, because incentives change when companies get bought and pivot is a euphemism that we can use. Then all of a sudden you're not trying to solve intelligence. You're scaling up, and I don't know how DeepMind works, but I'm curious what he would actually think about the trajectory of DeepMind.

Kim Stachenfeld
Yes, it's interesting. This, I guess, was what happened with Bell Labs, is they did basic research for a while, and then there was some reorganization that pivoted. DeepMind still does a ton of basic research, but the nature of it, and they do a ton of things that are just like, the goal of them is to contribute a high-impact scientific result.

Paul Middlebrooks
Always has been.

Kim Stachenfeld
Yes. I think that hasn't stopped being true or stopped being a big part of the way that they brand themselves and we brand ourselves. I think that is something that makes it a little bit-- that adds a little bit more complexity to the story about what can industry do? What can academia do than just like profit motive alignment? It's not profit irrelevant. Doing things that are positive contributions to the world are good for a company to do. The company does benefit from that. It's not just like, if you can make a product out of it, then it's beneficial to your company and otherwise it isn't.

It makes it a little bit more complicated. I think DeepMind still does do a lot of these things. You do have to appreciate that a company has different goals. Grant funding can be fickle in its ways too. It's not like anyone is totally immune to this. Everyone warns about industry research, like as things about the company change, the research will change too. That's true of grants and that's true as well, but the time scale is potentially shorter, and universities have been around for a long time.

Paul Middlebrooks
Would you rather have a beer with an academic or an industry person?

Kim Stachenfeld
Oh, it depends. There's a huge variability. [laughs]

Paul Middlebrooks
Come on. You can't wiggle out of that one. Of course.

Kim Stachenfeld
I've had excellent beers with industry and academic people. I don't know. One thing I really like about neuroscience departments is that there is a lot of variety and discipline. When I was in graduate school, my roommate, Diana Liao, she studied how marmosets talk to each other. That was really cool. None of my colleagues at DeepMind study how marmosets talk to each other. Neuroscience just has this tremendous variety and diversity and eccentricity and stuff. I just love that. That is a point in favor of having a beer with at least neuroscientists, [laughs] but it gets away from the more controversial question. I think you wanted me to answer which is objectively more interesting.

Paul Middlebrooks
Yes. Let's be objective about it. I'm curious. I don't know how your interests have changed, but I know that your projects have changed over time. I alluded to some of the work that you've done modeling simulations for turbulence, for design. Are those things that you pick? Are they mandated? You're still doing lots of interesting cognitive map reinforcement learning, neuroscience stuff, but how would you characterize your own shifting interests?

Kim Stachenfeld
What I started off working on in graduate school was computational models of hippocampal contributions to learning. Basically--

Paul Middlebrooks
You were right into hippocampus immediately. That's what you--

Kim Stachenfeld
Pretty much. I think that was actually built on a rotation project I did with Matt Botvinick and Sam Gershman. Really right off the bat. The kinds of computational models I was working on then that we were working on for that project, were things that more had to do with tabular reinforcement learning and linear algebra, which is to say they were models where we could set up the math analytically and compute exactly what we were doing. We weren't training neural networks to do stuff. The motivation was neural networky. It was that we want to understand how a certain representation in the brain supports the downstream computation that's happening.

There's different choices you have for how you're going to represent your experience, and the way you represent it will support different kinds of computations down the line. For instance, a simple example would be if you want to tell the difference-- If you have a downstream task that requires you to tell the difference between different colors, you need to have a representation that is not grayscale, that has color in it. Our work with the hippocampus was about how representations in hippocampus seem to account for what's going to happen next. That they form representations such that different states that are going to the same place end up with similar representations.

If you group things by what outcome they predict, that makes it easier to do certain computations down the line, computations that implicitly need to know something about what's going to happen next. It was a neural networky motivation because what neural networks do is they have several stages of processing, and at each one, it's a re-representation of their input. This question of what information they represented each layer, the neural network gets to figure that out on its own. It's not like you fix it the way we did, but the question of what it does choose to represent has a big effect on what it can do downstream.

The motivation was this representation learning question, but the method wasn't. After I graduated, I was at DeepMind. I wanted to do something with neural networks. I wanted to learn what those were all about. I don't know. I had these kinds of concepts that I feel like I was working on in certain ways that were tractable and let you really chew them up and understand every aspect of them. I wanted to see what would happen if we stuck neural networks on them and made them work that way. At that point, I--

Paul Middlebrooks
Wait. I'm sorry to interrupt you, but what percentage of neuroscientists do you think had that same thought? We have this thing, now I just want to stick a neural network on it and see what happens.

Kim Stachenfeld
I bet a lot. I'm not sure. It's definitely an aesthetic preference. I think some trial and error is required to see if you like it or not. I definitely have talked to some researchers who say sometimes they do some work with neural networks, some work without and say they just prefer working with them because they feel like it's working at scale. Some say they like working without it because they feel like they don't know what's going on, and they want to cut into science to think about stuff. Empirically observing is often what you're-- There's exceptions to this, but a lot of neural network research becomes a little bit empirical and observational.

Yes, I think a lot. Also, I think it's a very good skill to learn. I think a lot of people just want to play with a neural network because it makes them feel safer in their long-term prospects, which is very logical. I certainly felt that way at DeepMind. I was like, a lot more people doing neural networks here than linear algebra.

Paul Middlebrooks
Oh, interesting.

Kim Stachenfeld
Yes.

Paul Middlebrooks
Sorry, I interrupted your train of thought--

Kim Stachenfeld
No, not at all. It's a great question. It was the closest path to looking at the same kinds of principles we were thinking about, like how the brain represents relations between things, how the brain makes predictions about what's coming next, was to work on Pete Battaglia's group where they were using graph neural networks, which are a neural network architecture that reasons about relations between things and using them to make predictions about how a physical system will unfold.

Basically, if you have a physical system with a bunch of interacting entities, like things bumping into each other or fluid particles bouncing into each other, that's a relational system. Interactions between those particles determine what's going to happen. It's also a predictive problem. You're trying to see what's going to happen next. That's how I got there. It was sort of the nearest-neighbor research manifestation of these relational and predictive ideas but in a machine-learning form and a machine-learning application.

Paul Middlebrooks
Do you view the brain as a machine-learning entity?

Kim Stachenfeld
Broadly speaking. At some literal level, I guess I view the brain as a learning entity and not a machine, I guess by definition. Machine learning, I think is the extent to which the brain is a thing that learns stuff. Machines that learn stuff are a good batch of machine-related analogies for that, if that makes sense.

Paul Middlebrooks
Yes.

Kim Stachenfeld
Sure, I'll accept that. I think machine learning is a really great batch of tools for thinking about how learning works. Yes, I think there's just so many concepts in common between understanding how machine learning works, understanding how the brain works, building better machine-learning models, understanding why different brains are the way they are.

Paul Middlebrooks
Do you see the brain as a learning entity?

Kim Stachenfeld
Yes.

Paul Middlebrooks
Primarily.

Kim Stachenfeld
Yes, that's at least the aspect that I study and find most interesting. I think the whole nature versus nurture, I don't have a super creative opinion on that, but I guess the extent to which the brain is a learning thing is probably the extent to which machine learning is a good batch of models for it.

Paul Middlebrooks
OK. You have worked a lot on the hippocampus, and there was a guest speaker that came, I'm at Carnegie Mellon University, and there's a guest speaker that I had lunch with, and he does work with hippocampus stuff. I realized, maybe I hadn't thought of it before, but the hippocampus has been sort of the darling of neuroscience now, since place cells, probably. Since that became popular, would you say that's right?

Kim Stachenfeld
Yes.

Paul Middlebrooks
You're biased.

Kim Stachenfeld
I was around back then, but I definitely got that sense. I was warned when I started hippocampus stuff, people were like, "It's a crowded field. Good luck in there."

Paul Middlebrooks
Oh, really?

Kim Stachenfeld
Yes.

Paul Middlebrooks
I thought visual neuroscience was a crowded field, but I guess hippocampus took it since.

Kim Stachenfeld
Yes, hippocampus is crowded; vision is crowded; RL is crowded. I don't know, a lot of the good stuff.

Paul Middlebrooks
Oh, yes, these days, RL. Do you feel that it's crowded? Because I'm tired of seeing algorithms, new algorithms. I'm like, "Oh, I got to go learn another one."

Kim Stachenfeld
I work on hippocampal contributions to RL, so I guess there's some part of me that just thinks objectively RL and hippocampus are the most interesting and could never be too crowded, but they're definitely real popular.

Paul Middlebrooks
What does hippocampus do?

Kim Stachenfeld
What does hippocampus do? Seems to help with memory, maybe some structure learning.

Paul Middlebrooks
You're good at wiggling out of questions like this, but so there's the learning aspect.

Kim Stachenfeld
I'm not trying to. I'm just also trying not to say things that are wrong.

Paul Middlebrooks
Yes, when you're someone like me, you get used to that real quickly, and I'm fine being wrong, but it's memory, it's learning, it's spatial cognition, it's a cognitive map. Is it all of these things? Do we need to say hippocampus does function X?

Kim Stachenfeld
Yes. It seems like it is implicated in a large number of things. People make fun of hippocampus researchers and say that they think that the cortex is just to keep the hippocampus warm, which--

Paul Middlebrooks
I haven't heard that one. That's pretty good.

Kim Stachenfeld
Yes, it's obviously not doing everything in the world. Hippocampus is unique in certain ways that do justify its supposed ubiquity. It gets projections from all over the brain. It seems like it's processing lots of different kinds of information. It's not going to be specialized on a particular sensory modality. It seems like there are some things about it that are different from other areas. In particular, it seems like it's capable of really rapid learning. I think the complementary learning systems idea makes a specific but non-specific prediction about hippocampus.

Paul Middlebrooks
Let me just say what complementary learning systems is real quick. The idea is that you learn quickly and rapidly in the hippocampus, and then it transfers over time. What is the word I'm looking for? Consolidates, right? Is that right?

Kim Stachenfeld
Consolidate.

Paul Middlebrooks
Yes, in the cortex over time. Hippocampus keeps sending this learned information to cortex, and cortex over time consolidates the information. There are two kinds of learning systems. Did I say that right?

Kim Stachenfeld
At least, yes, that's my take. That decomposition of roles leaves a lot of stuff for hippocampus to do. It's saying like what it's specialized in is rapid acquisition and memory for specifics. That should relate to all these other things, like rapidly learning about spatial environments and where things are stored in them, rapidly acquiring new episodic memories, remembering, preserving specific aspects of memories from which you don't want to generalize. The classic example is the difference

between where you usually park your car would be more a cortex job and where you parked your car today would be a more hippocampus job.

It's something specific. You might change this over time, but you don't want to forget that information. It's useful to preserve. Then also the ability to form new memories should interact with the statistical memories you've formed, the structural memories you've formed in a pretty deep way, because a new thing you want to learn is maybe a reconfiguration of old statistics. If I today put a coffee cup on a bench, I have a concept of bench. I have a concept of coffee cup. Those are probably statistically learned properties, but the specific conjunction of them is something that's new.

That hippocampus's rapid learning should interface with slower learning isn't necessarily a contradiction, but it does make it easier to just have this explosion of roles where hippocampus is doing all of these things when actually we should really be thinking about it more in terms of how hippocampus is interfacing with other areas that are all partially doing these things.

Paul Middlebrooks
You mentioned reinforcement learning, and you've done a lot of work in reinforcement learning. I mentioned the idea of cognitive maps a moment ago. I mentioned how hippocampus has been the darling, and all of these things are wrapped together. Since then, there's just been an explosion of algorithms, reinforcement learning. You've worked with successor representation, etc., model-based, model-free, what is your perspective? It's out of control, isn't it?

Kim Stachenfeld
I think it's nice. I think that it'd be good to get everyone on the same page, I guess.

Paul Middlebrooks
What does that mean?

Kim Stachenfeld
You want model comparison. I guess this is something that you want to see what one model does that another model doesn't do necessarily. Maybe a risk of like proliferation of different models is if you aren't comparing the models on the same data or if you aren't comparing models at all. You're just saying, "Look, this model captures this picture that was in a paper. This other model captures this other picture that was in a paper," and you don't necessarily compare them all together. People are doing this with hippocampus a little bit. There's some toolboxes that people have been developing.

I know RatInABox is one of them by Tom George at UCL. Clementine Domine has one that I'm blanking on the name of, but that's also at UCL. Basically, these things that are trying to take lots of models and try to make the same predictions of them. Brain score is a nice example of this in the visual world. There's issues with making things too score-based and too benchmark-based, but it has benefits that all people are talking about the same data and not just saying this model captures this one aspect. This other model captures this other aspect. Assuming that the brain could just put both together and have it work just as well or not necessarily arbitrating when the models are mutually exclusive.

That's just an important thing to have. On the other hand, I think good-- I almost see that massive proliferation as a success of the RL account rather than a warning that if you make a model that's somewhat right, that's doing a simplistic RL thing, then a bunch of other models will come that do more nuanced versions of that or more particular versions of that. A multiplexed RL, distributional RL, action prediction error, regularized RL, all these kinds of things that add nuance to the picture. If the original RL thing was totally wrong, this would be a really bad use of effort.

If the original RL thing is a bit right, but not capturing everything, that worked pretty well. It's like you have a sort of coarse, simpler picture and that got you into the vicinity where people can bubble around making different models and seeing which aspects you can capture and then try to consolidate them.

Paul Middlebrooks
Yes, I think I remember you saying in one of your talks, maybe, I don't remember exactly the way you phrased it, but you used to think that reinforcement learning, or at least model-free reinforcement learning, was like, "Hey, you did something good." Now you think it's like, "Oh, this is the worst way to--" Well, how did you phrase it? I don't remember.

Kim Stachenfeld
Yes, I know what you're talking about. I use an example when I'm giving RL lectures on what makes reinforcement learning different from other types of learning. It's commonly--

Paul Middlebrooks
It's cruel. I think cruel is the word that you use.

Kim Stachenfeld
I used the word cruel. I think reinforcement learning is a little bit cruel. I think of it as almost very passive-aggressive. The example I have of this is if you were trying to learn biology, one way you might try to learn biology is read a lot of textbooks and try to find patterns and link things across different bits of the text and group all of the things that have to do with the mitochondria and recognize like a, I don't know, causal structures from mitochondria producing energy to enzymes that use that energy to build stuff or whatever. You try to identify structure. That's what unsupervised learning is all about, is it's a large and amorphous field. It's all about trying to learn patterns and structure.

Supervised learning is much more limited in its scope. That is the setting where there exists a correct answer. There's a right answer, and you are trying to find that answer. You can set it up as classification problems or regression problems. You have a picture of something, a human labeled it a picture of a dog, and you are trying to train a network to say, this is a dog. If it guessed that it was a cat, the answer it gets is, no, this was a dog. It doesn't just get an answer 10 points or 6 points or minus 11 points without any context for how many points are possible or whether there was more points available with some other answer.

That latter one is what reinforcement learning does, is trying to learn biology by taking tests and then just getting a score of 34 and not knowing if it was out of 34 or 8 million or exactly which answers you got wrong and which answers you got right. It's a very like restrained signal. It's nice for setting up the problem of autonomous learning. If you can learn from signals like that, then you are pretty good at figuring stuff out on your own. You don't need the handholding of supervised learning. It, I think, also highlights the challenges of it. The credit assignment problem is really hard. You have to figure out which things contributed to your points and which things didn't.

The exploration problem is really hard. You have to figure out how many points were available to you were you to have taken different options. Yes, I think of reinforcement learning as somewhat cruel, rather than this, like, "Oh, you did a good job. You get a treat," which I think I used to conceptualize it as a very gentle, warm way to nurture an agent.

Paul Middlebrooks
Why do our brains employ such cruel passive-aggressive algorithms?

Kim Stachenfeld
It's very flexible. You can learn lots of different types of things, and it's much more autonomous. You don't need information to be labeled for you. If you don't have labeled information, you don't really have a choice. The other thing is there's ways to make reinforcement learning easier. A lot of the research that I've worked on has been about how to do that, how hippocampus might specifically help by doing that, by combining unsupervised learning with reinforcement learning.

Basically, by learning some structure so that you can strain the reinforcement learning problem to a narrower set of possibilities or give it some hints about what kinds of things might be driving the number of points. Basically, read a textbook and then take a test and see how many points you get, rather than just taking the test and hoping that a billion tests will eventually teach you biology.

Paul Middlebrooks
Where is model-based reinforcement learning these days? My sense is that, so I travel in particular bubbles, like we all do. My sense is that the tide turned to everything is model-based reinforcement learning at the pendulum, let's say. That pendulum has swung back and said, well, maybe very little is-- You actually don't need model-based to do a lot of these things. Maybe we're sort of overstepping our bounds in thinking of everything as model-based. Is that accurate at all?

Kim Stachenfeld
It's interesting. The huge tide in machine learning recently has been the rise of self-supervised learning to the extent that it's

sidelined reinforcement learning to a large degree. I think there used to be a lot more appetite for let's try to learn as much as we can with reinforcement learning. We'll rely on self-supervised learning when we need a bit of a crutch. Now self-supervised or unsupervised learning, they mean really similar things. Basically, it turns out that gets you very, very far.

If you just train a model to do next-step prediction on words, that's how language models are trained, you're very close to being able to not just predict the next word, but make the next word be something that you actually want it to be. It's a little bit different from some ways of conceptualizing model-based learning, but it's very similar to others. It's basically like you're using a predictive model to start off the process. Then you coax it to do exactly what you want, which is very much the, I think, key concept of model-based learning. Most model-based reinforcement learning, or at least most that I was familiar with, or the cartoon I had in my mind of model-based reinforcement learning, is tree search. You have a model that can simulate different things. You simulate different outcomes, and you see what happens, and then you pick the action that leads to the best possible things. Model-based reinforcement learning can be structured in other ways, too.

It could be, for instance, learning concepts about the world, learning some model of these things are all the same and are going to behave the same way, and then when you learn about their value with model-free reinforcement learning, you've done some vaguely model-based organization of them that supports it. That's what the successor representation does it more that way, that you represent things in terms of the predictive structure, and then the model-free reinforcement learning that happens on top of that automatically takes into account some features that a model-based model would have told it about.

There's other ways that people use these self-supervised models in conjunction with reinforcement learning. It's a little bit of a different take, but I think it's still the same concepts shuffled around in a different way.

Paul Middlebrooks
Shifting gears, I was going to ask you-- I was going to segue into cognitive maps. I'm not sure. I'll just shift gears, actually, and ask you, and one of the questions that I sent you was, what do you think neuroscience needs more and less of? Forgive me, someone who works at DeepMind, in industry, I think has a unique perspective, perhaps, on these things. I'm genuinely curious, from your perspective, if you think that you have a perspective that might be unique relative to a normal academic like myself. Can you see neuroscience from the outside? I know you're still inside it also, but you're also outside it. Do you have a take on this?

Kim Stachenfeld
It's funny. I struggled with that question.

Paul Middlebrooks
It's an unfair question.

Kim Stachenfeld
I don't know. I feel like it's a very fair question. At some level, every part of picking a research problem is trying to figure out what the neuroscience world needs more of that you are equipped to do. I guess the kinds of things that I wrote down that I tried to brainstorm about this were--

Paul Middlebrooks
Wow, you put effort into it.

Kim Stachenfeld
I put effort into this, [laughs] which is extra embarrassing because I don't know if I have a great answer.

Paul Middlebrooks
Oh, I'm so thankful.

Kim Stachenfeld
I think that the thing that I thought maybe I might have a bit of some unique insight on or that the DeepMind perspective relates to is the extent to which we can build pretty sophisticated models of really complicated processes. The basic revolution in machine learning recently, or the thing that we have increasingly been showing our ability to do, is that we can learn really complicated functions. With enough data, the sky's the limit to how complicated a thing you can learn. Learning

predictive models is definitely something we can do. I think there's a lot of appetite for this, for building predictive models of neural data, of behavior, building foundation models that integrate lots of different data.

The question that arises from this is, what do we do with that? If we distill a black box system into another black box system, what do we do? Was that useful? What is that useful for? I think even on its own, that is a useful thing to know how much systematicity there might be in the data, how much is predictable, how much a model changes when you add new data, what counts as surprising based on the data you've had at the time. Even these black box predictive models, I think could be really useful.

There's some other ways I think that we can make them more useful if we combine the deep-learning model with a model that has a little bit more structure, a little bit more interpretability. This has been a really big theme for the NeuroLab at DeepMind recently. Folks like Maria Eckstein and Kevin Miller have done some nice work on this recently. I've been thinking a little bit about how to relate the learned physics dynamic stuff to this if we were to apply the same techniques to biological data.

I think, basically, while the revolution in machine learning was very much about building good predictive models, we haven't done quite as much work on building predictive models that are also constrained to be interpretable or constrained to interface with existing models that we've built that we know how to say structured things about. That's not really something that's fundamentally ruled out. I don't think we have to consider these models 100 percent black box.

I think thinking about how to how to combine really complex models with some structure, either by including it as an architectural prior that you know something about or interfacing the model with knowledge or interfacing the large black box model with some prior that takes into account structure that could be hypothesized. I think that that's a really useful direction for neuroscientists to be thinking about, particularly neuroscientists who have an interest in building large-scale models and access to doing that. Trying to come up with something that's in between super complex predictive, we're just going to build it all and pretend that that was enough just to imitate the system.

This more like we only care about exactly what we can build in a very attractive way. If we can interface between those levels, I think that would be really powerful. It would also open up the ability to compare models to data in a much richer way. If we can say, "This is the behavior that should be happening at the level of this video of an animal because we've combined the model with something that actually generates low-level behavior," that would be a much stricter way of comparing to neural data and behavioral data than it would be to just say, "This shows on average an increase," or "We see this gross statistical effect on average, but it's not capturing every specific eccentricity of the behavior."

I think that integration between structure and data is the big thing.

Paul Middlebrooks
Is it fair to give the crude summary that neuroscience needs more interpretability on top of the big models?

Kim Stachenfeld
Yes, broadly speaking, or maybe that the dichotomy between interpretability and data-driven is maybe a little too harsh a dichotomy. We shouldn't be fighting if we want models that are predictive, or if we want models that are interpretable. We all want both. We should just try to see how we are-- think of this more as a Pareto frontier that has a more complex trade-off.

Paul Middlebrooks
What about, did you take any notes on the question what neuroscience needs less of? [laughs] I hope you didn't work too hard on these things, by the way.

Kim Stachenfeld
I just jotted out some notes. I wrote specifically, I don't know if I really have an opinion on that. Honestly, I couldn't think of anything creative here. [laughs] I don't know. I think one thing I noticed in my own research is that when I am targeting a journal paper versus a NeurIPS paper, I have a different-- it is easier to do something a little bit bite-sized or more incremental or, I don't know.

Paul Middlebrooks
With NeurIPS?

Kim Stachenfeld
NeurIPS, yes, that it's easier to be like, we capture this one thing, trust me, it's right, whereas with when I'm thinking about writing journal articles, it's like we really need to convince people broadly in a really different way. The goal of a journal paper, I think, is to--

Paul Middlebrooks
Tell a story.

Kim Stachenfeld
Yes, tell a story and talk to-- I feel it's much more about talking to neuroscientists and saying, "I think this is right; do you believe me?" Or "I think this has something to offer. Can I convince you that that's true?" I wouldn't say we need less NeurIPS papers, but I think if we just started only doing NeurIPS papers and not writing in that more journalistic style, it might be a bit of a loss. On the other hand, the journal review process is way overbearing and unnecessary, and I don't like for-profit journals, and everything about open review is, I have a lot of respect for. There's pros and cons to both models.

I think that specific aspect of writing with an idea of impacting the field, rather than writing with the idea of scoring a win and getting a NeurIPS paper in is probably more scientifically minded.

Paul Middlebrooks
There's something nice about the honesty of, let's say the NeurIPS, we'll just call it the NeurIPS method, because you don't have to change the world. You can just do something that's incremental. You don't have to claim that you're solving something that no one is, there's just a lot of overclaiming in the storytelling in journal papers. When I think about writing a journal paper, I think, "All right, how can I trick them into thinking that my stuff is important?"

Kim Stachenfeld
That's funny. I might just have such a different model of journal of yours versus-- because I think that NeurIPS papers over-- I think of them as the worst for the over [crosstalk]

Paul Middlebrooks
Oh really?

Kim Stachenfeld
You know what, maybe it's just safer to say that they over-claim in different ways. I think the thing I find frustrating about-- and I think that's actually probably pretty accurate. There's an incentive to over-claim about how broadly your model captures results in a neuroscience paper. There's an attempt to over-claim about how novel your method is with a machine-learning paper. I think the most frustrating thing about machine-learning papers is that there's just this incentive towards what I call a categorical geometry.

You want to say, "Here, we made a model. It is completely different from all of the other models. It is itself just not-- there's no frame of reference in other models. It's a unique point on the landscape. It's new, and we made it. It's a new method, and it's innovative, and it's novel. I promise you the method is new. We compare it to other models, which themselves are their own little points on the landscape and our model does better than them." You're not really looking does your model have the same thing as this other model at some level and a different thing at this other level?

How is this similar to other stuff? In other fields, like, I assume math, but I'm not a mathematician, and neuroscience, I think you get more points for being integrative. You get more impact if you can say how things relate to other things, rather than just saying, "This is maximally unrelated to other things." I think that makes it really extra work to understand principles about what's working and what isn't working in machine learning.

Paul Middlebrooks
You would rather have a beer with a neuroscientist than a machine-learning industry researcher?

Kim Stachenfeld
You didn't ask that. You asked academic or industry.

Paul Middlebrooks
I know, but I'm changing the question. It sounds like--

Kim Stachenfeld
I'm not into that one either. [laughs] The thing is, at some level, I'd rather have a beer with a machine-learning person because the only way to really figure out how a model is related to another model is to have a beer with them, because it's not going to be in the paper. [laughs]

Paul Middlebrooks
Oh, wow. A little backhanded compliment, perhaps.

Kim Stachenfeld
Yes, I'm not sure. No, I think that they're both super interesting fields. If I could have picked a single one, I would have done it. I think it is true that the behind-the-scenes talking does fill in a lot with machine-learning papers because you learn how they thought of the idea in the first place, what it relates to. This seems true of neuroscience at some level though. You talk to people and they're like, "Yes, this data we feel really sure about. This data was technically significant, but we feel a little shakier about it." There's always a lot to be learned from actually talking to people.

Paul Middlebrooks
Neuroscience feels like it's in a-- in one respect, it's in an awesome place because we have more compute, more data, more models, more everything. It feels like we don't quite yet know what to do with all of the stuff. There's a lot of exploratory work. What I'm doing right now, it's like, "We're going to throw all the tools at it that are around because we don't really know what to do with the data." Does that ring true to you?

Kim Stachenfeld
Yes, that does ring true to me.

Paul Middlebrooks
What does that mean? How do we get beyond a weird spot? Just a bunch of brilliant people that make the right advances?

Kim Stachenfeld
At the very least, I think that's a feature of neuroscience that's useful for recruitment because it makes it an exciting time to be in neuroscience. You're not just answering specific questions; you're figuring out what are the right questions to ask and what's the right way to describe them.

Paul Middlebrooks
That's the hard stuff and the fun stuff.

Kim Stachenfeld
A lot of people don't like that. They're like, "I feel like I spent five years doing a Ph.D., and I have no idea if even the question I answered will be relevant in 20 years, let alone the specific answer to it." I think that's a personality difference, I guess, about risk profiles and what kinds of thoughts you want to have. I think the fact that it's philosophical, at some level, we're not really sure how to even be saying things is really interesting. The aspect of the black box machine-learning predictive models versus more structured models is a really nice example of that.

Is that a philosophical choice about which direction to go? Are there ways to combine the best of both? What would that look like? Maybe there's more structured ways to do it. Maybe there is a right, yes or no answer. At some level, it feels like it's a bit of an empirical question, try different stuff and see what works. What works is itself a really hard thing to evaluate. What are you going for? If you're just going for prediction, then of course the models that just do prediction are going to win. If you're going for insight, that is pretty hard to optimize for. It's really hard to take gradients through it.

I think there's a lot of, exactly how to operationalize what it means to understand something and make progress on a question, have a model that makes progress is really tricky. The RL models you mentioned of, we have this RL model and then there was an explosion of things, you could argue that was a win for the model and that it opened up a lot more research. On a longer time scale, I guess we'll know if that research was useful or not, but I'm not sure how we'll actually be able to use that to restructure the models we're asking.

Paul Middlebrooks
You just made me realize that it happens too frequently on my commute to work. On the train, I'll just be sitting there and then I'll think, "Oh God, what is my question?" That's the important thing. It's to have the good questions. I was like, "Oh, do I have a question? Oh no."

Kim Stachenfeld
Yes, no, I don't know. I feel that way. I feel that way a lot. I put off for a really long time writing my bio for my Columbia page. In fact, I think I still have not written my bio for my Columbia page because I was like, "Oh my God, a one-summary sentence of my research? That sounds like a job for tomorrow, Kim." [laughs] That is really hard.

Paul Middlebrooks
This is again, a very difficult question, I'm sure, but can you articulate how your questions have changed, the nature of your questions, if not the content of the questions, over the course of your career?

Kim Stachenfeld
Yes, I think one big one. I think I originally was motivated, one of the things that shaped some aspect of my research trajectory was in college, I took this philosophy course with Dan Dennett that was on AI, it was called "Language and Mind." One of the thought experiments he raised was this idea of a robot firefighter and what it would take to design a robot firefighter. He just jotted down all of these competing things that the robot would have to figure out how to reason about at the same time.

It would have to figure out how to search a building, to make these complicated decisions about whether it should be looking for people or if it should be taking the people it found out of the building, these high-level goal questions, while at the same time also putting one robot foot in front of the other and actually moving in a direction. We end up parsing this or triaging this hierarchy of decisions, and that allows us to function in the world. There's some organization to behavior, and finding it is an interesting problem, is the way to study the brain.

A lot of the projects I thought about and things I worked on relied on these explicit hierarchies, like what should the high level and the low level be? I think one thing I have--

Paul Middlebrooks
Wait, so what you're saying is you came into it because of that, or you were influenced by that into thinking in terms of hierarchies? Is that what you're--?

Kim Stachenfeld
Yes, that's a pithier way of putting it. I was very thinking explicitly about hierarchies. I think one thing I moved on was exactly when and where to think about hierarchies being useful or necessary that a lot of times if you train a model that is a big model that is trained to do something next-step prediction, it can do some of those things implicitly. You don't necessarily need to hard code those things or the places you thought you would need to hard code them, you don't necessarily. It turned out like, I felt like I had all these projects where I was like, "This obviously will need hierarchy."

Then it was like, "Actually just a neural network doing simple stuff or trained in a simple way is a pretty tough baseline to beat. It's not necessarily that hierarchy isn't useful. I think it just often is emergent or implicit in the system. I think a lot of the ways I shifted to thinking about hierarchy were instead of just thinking about how to put it into a system, which I still do, and I think is still useful in certain settings for getting better generalization stuff. It's not like that's not useful, but thinking about when it's useful, I think I took for granted that it was obviously useful a lot of the time.

Now I think of it as something that you need to think a lot more carefully about when the data tells you enough to generalize, when you actually need hierarchy to reason more efficiently or to make a broader inference. The other thing is I shifted to thinking, "If what appears to be a hierarchical ability is emergent, how can we understand how that is unfolding in a neural network?" Really shifting to it as an evaluation rather than a method. If the behavior is happening, can we understand it? I think thinking a little bit more implicitly about how structure and generalization emerges was maybe a little bit of a soft trend in my research.

Paul Middlebrooks
Has modern AI shifted the way that you've thought about that as well? I know that you're impressed with large language models, foundation models, I've heard you say that, but of course, everybody is. I think I remember you saying that that has shifted the way that you think about intelligence or brains, perhaps.

Kim Stachenfeld
I think large language models are good. That's not the world's hottest take. The fact that they can get as far as they can with next-step prediction is pretty fantastic to me. Maybe it should have been more predictable because it reminds me of this thing that was observed a really long time ago with H.M. that supposedly if you just had a conversation with him, he didn't seem that-- oh, sorry, I should say, yes, for people--

Paul Middlebrooks
I was going to do it, that's OK. Do you want me to do it? H.M. was the most famous patient in history in the neuroscience world because he-- I don't know how he lost his hippocampus. Was it a stroke?

Kim Stachenfeld
No, it was surgery. They took it out.

Paul Middlebrooks
Oh, that's right. Epilepsy. They removed it. They removed a large swath of his hippocampus and some surrounding tissue as well. He had retrograde amnesia. Is that what it's called?

Kim Stachenfeld
Yes.

Paul Middlebrooks
He couldn't remember anything moving forward. Anterograde? I don't remember. He had amnesia.

Kim Stachenfeld
I should definitely know this. It was the one where he could remember stuff before the surgery, up to a few days before the surgery, but he couldn't form new memories, broadly speaking.

Paul Middlebrooks
Neither of us know what it's called. That's embarrassing, isn't it?

Kim Stachenfeld
[laughs] Yes, we have both of our hippocampi, as far as I know.

Paul Middlebrooks
I know. We also have computers in front of us too.

Kim Stachenfeld
I know. I'm worried your listeners will hear me typing if I look it up.

Paul Middlebrooks
Let's just fix this. Let's fix this immediately here. Surgically, anterograde.

Kim Stachenfeld
Cool. I'm glad we checked.

Paul Middlebrooks
All fixed.

Kim Stachenfeld
Anterograde, I totally suspected. [laughter] He had both of his hippocampi surgically removed and a little bit of OFC too, I think, and lost the ability to form new episodic memories. Basically, he lost the ability to form most memories besides slowly acquired complex motor control tasks. Supposedly, if you had a conversation with him, he didn't seem that different. You would think that it would be the ability to form new memories would manifest pretty immediately in a conversation. Instead, it would be if you were just having a conversation with him about the weather, what was going on, or something that didn't require accessing old memories, he would sound pretty normal.

It was really specifically tasks that require things we think of as hippocampal functions, where the deficits would show up.

Paul Middlebrooks
Why is that surprising to you, though?

Kim Stachenfeld
The reason it is, maybe it shouldn't be surprising. I guess I just think that memory informs so much of our ability to have conversations. That when you're talking about new things, it feels like you're searching your memory for the thing to say next. If you say something that reminds me of something that I heard, and then we talk about that for a while. It feels like a process where you keep diving and scooping up memories and sticking them into the conversation.

Maybe his conversations were a little bit boring or something and didn't do that, or maybe, I don't know. [laughs] That's, I guess, harder to write a paper about. I think the reason I found it surprising is just because it feels like memory informs what you're going to say next in a very recognizable way. Maybe that's just the wrong instinct, but it seems like, actually, you can continue in a conversation saying pretty complex things without obviously seeming that different unless you're asked pretty specific questions.

That's, I think, how I now interpret retroactively this surprise about predictive models, that they're going to do a pretty good job predicting what's going to happen next in a sentence. Only on really specific questions about factuality or things that specifically go against the statistics of your experience, specific reasoning questions, will something that's obviously wrong? You can get pretty far on just next-step prediction. That, I think, is maybe a neuroscience-- a thing that should have made something that was surprising to be not surprising by historical context.

Paul Middlebrooks
You think it's all about prediction?

Kim Stachenfeld
I think prediction can get you really far. It's hard to rule out other things. There's things that are not just prediction, I think, especially in hippocampus, we had a model of how hippocampus represents predictive structure. There's a lot of retrospective structure that it represents too. If you came from different places but are going the same place next, hippocampus will be different based on where you came from. It represents information about your past as well. Keeping information around about the past and the future seems like it's important, but I think prediction can get you really far.

Paul Middlebrooks
Do you think you have a better understanding of now, as opposed to, I don't know, whatever, five, six years ago, of what intelligence is? I think I have a worse understanding than I did.

Kim Stachenfeld
Me too. [laughs]

Paul Middlebrooks
Really?

Kim Stachenfeld
I guess I thought of it as something maybe a little bit more structured. Now I'm maybe less sure. I guess maybe a large enough neural network trained on complex enough data, the fact that it could eventually imitate something that looks basically like language comprehension, maybe shouldn't have been surprising because we just know that neural networks are universal

function approximators. They can get any function, even a really complicated one. I think I probably had different intuitions based on just nothing in particular, but my own intuitions about what would be required to make that work efficiently and what we would have capacity for.

I think I maybe felt like more structure would be required to make as much progress as has been made.

Paul Middlebrooks
What do you mean structure? I'm trying to--

Kim Stachenfeld
I mean explicit decompositions, like explicitly saying, "This is a category; this is another category; their operation is constrained in some explicit way." I think that I might've felt like-- I think I had an instinct for us needing some aspect of that. I think one thing that's really interesting about transformers as the model that is currently reigning supreme is they're very beautiful architecture. They have this deep symmetry built into them that processes sequences in a really different way from how people previously thought sequences should be processed efficiently, which is to say they just look at everything in their recent context.

Instead of saying things that happened a long time ago are going to get processed differently from the model, they have been processed more times because you have to keep applying a recurrent neural network that keeps updating itself based on new information, you just say like, "We're just going to have all of the information we have in recent history, and we're going to label it by its position in the sequence. Every point in that sequence, every entity in that sequence is going to go through exactly the same processing pipeline. You're going to have the same weights that look at every token in that sequence.

Only the label on that token that tells you where in the sequence it is tells you information about the sequential structure. You could basically scramble up the sequence, and as long as you kept the labels the same, the model would do the exact same thing. That I think is really cool. It's a really fundamentally relational structure. At a very low level, it's saying, "We're going to process all relationships between entities in our sequence with the same general-purpose relational operation. You can tell us if you are nearby in a sequence or far away in a sequence; you can tell us what features you have."

As layers of processing happen, you can update these tokens with whatever the model has processed about them, but it's still doing the same relational mechanism. It has a deep connection to these models, graph nets that we work with a lot for physics that fundamentally represent predictions about a system in terms of relations between its entities. I think there's been a lot of work on relational reasoning as a powerful mechanism for computation. I hypothesize, although I'm not entirely sure how I would test this, that operation is a powerful one that is partly giving the model some computational object that it can implicitly arrange relational operations into.

Paul Middlebrooks
Does that tell us anything about brains? Sorry, this is a naive question, but we don't get labeled sequence information through our senses.

Kim Stachenfeld
The brain might go out of its way to label them. That's, I think, one hypothesis that's related to cognitive maps, which you mentioned. I think the hypothesis about place cells and time cells in hippocampus is that they're learning to label incoming information with a spatiotemporal tag, which then could be a useful attribute to have on some information. If you want to index it by space and time, that could be really useful.

I think one thing that's useful about transformers is they structure sequence prediction; they structure memory as a problem of attending over your past. I think that's a cool way to think about how you might use memories. Rather than just recalling the thing that is the most similar to your current situation, loading that into memory, and then reasoning accordingly, it's a much more flexible and directed process. It's saying, "You can choose which things in your past you're going to attend to and use to inform your current decision."

It's a more flexible take on something like temporal context model, where you are in a sequence informs how you label and retrieve entities that you remember. Instead of having just an exponentially decaying factor, you have a weight on it that you can learn, something that you can more flexibly modulate to relate your past to your present. I think this model class is useful for expressing questions in that form. How could we attend to our memories in a way that we now know how to train, but is a little bit different and richer than what previous models are doing while still relating to those models?

There's a lot of questions about biological plausibility, but that's true for all the neural networks that we currently reason about.

Paul Middlebrooks
Who cares? Who cares about biological plausibility though? Really, if it's working, it's working, right?

Kim Stachenfeld
Yes; it depends if it's working to explain how the brain works or if it's working to be good at machine learning.

Paul Middlebrooks
That's a different question.

Paul Middlebrooks
I don't know. [laughs] If it's explaining stuff you couldn't explain otherwise, I think that counts as working. I guess the key biological implausibility of transformers is like, how would you keep all those tokens around.

Paul Middlebrooks
I don't. I know that and I don't keep. I'm close to H.M. in that respect.

Kim Stachenfeld
I'm not exceptional for my memory either.

Paul Middlebrooks
How long are transformers going to be the thing?

Kim Stachenfeld
I don't know. State space models are getting some attention. I think they have a property in common with transformers, which is that you can train them really efficiently by parallelizing over hardware. I think the key innovation with transformers was, yes, in some sense, it's less efficient because you're just representing all of your memory rather than re-representing it in the way an LSTM does with a shared batch of parameters. However, you can parallelize the operations that you're training a matrix on.

Even though you have way more parameters, you can efficiently train them on way more data. It's actually nice. State space models too, I think they're a rearrangement of RNNs that allow parallelization to happen a lot more easily. That's essentially my understanding. I think the architectures that lend themselves to parallelization, that's a really big advantage if you've got lots of data and lots of hardware. I don't know how long there'll be the chosen model. Training a big model is really expensive. Once you train a big model, there's a real switch cost.

Paul Middlebrooks
Oh, yes. That's not stopping anyone because I guess there's a lot of money in it.

Kim Stachenfeld
Yes, I think that's the thing is there's a lot of money in getting a good one, but it's really expensive to train a big one. I think the burden of proof to say you should switch your model or you should train a second model just to see if it's as good is a little bit high, but definitely there's a lot of upside. If you're a big tech company and you've got money to burn on really good AI, there's probably incentives to do it.

Paul Middlebrooks
All right, Kim, I won't keep you too much longer. I have basically two more questions to ask you. One, is there something that you are currently struggling with just beyond your reach? Is there something that is gnawing at you?

Kim Stachenfeld
Definitely, the relationship between predictive models and structured models; that's gnawing at me.

Paul Middlebrooks
That's the thing.

Kim Stachenfeld
I feel like there's opportunities there, so I'm like-- I guess something that really annoys you and something that's good research motivation has a lot of overlap. It feels like there should be something there. I find that really exciting. I think that that's probably been the major thing that is gnawing at me at the interface of neuroscience and AI.

Paul Middlebrooks
I mentioned your work on turbulence a few times, and it struck me actually. I just had someone on who thinks of the brain as a system of cascade turbulence. He uses multi-fractality and principles from turbulence to think of our cognition. I thought, "Well, here's a stupid question I could ask Kim." Did your work on the learn simulators for turbulence, did that shape at all how you think about brain processing?

Kim Stachenfeld
I'm so glad you asked. That also is-

Paul Middlebrooks
Really?

Kim Stachenfeld
-a thing that is gnawing at me. One is this idea that we should be able to use-- the same models that we train to capture physics systems should also be able to capture biological systems too in neural data and where that's a good model and where that can show us stuff is also just something that's been gnawing at me. It's the same question. It's like, if we build a predictive model with this kind of structure built into it, when is that structure the right fit for different kinds of problems in neuroscience and also in biology?

I think one structure that's built into the graph nets and the convolution nets that we used for fluid modeling is that they have this repetition, just the way convolutional neural networks, what they do is they learn about a little patch of an image and then they repeat that pattern, that filter for extracting features everywhere in the image. If, for instance, you learn something that's an eyeball detector, then it'll go through your image and wherever there's something that looks eye-like, it'll pass that on to the next layer of processing.

Graph neural networks do the same thing, but instead of looking at patches of an image, they look at patches of a network, like a little neighborhood, and then extract rules about maybe if there's a bunch of particles that are crashing into each other in that window, they learn that there's going to be a collision. Kinds of feature extractions like that. This seems like a really cool model for fitting data in the brain. If you want to try to fit a model of how the brain's connective structure is giving rise to different computations or how the structure of different brain areas relates to different dynamics, this seems like it could be a useful model to apply.

The main risk of it is that the way that they share and repeat their pattern, understanding when that isn't or is a good fit for neural data is complicated because the brain isn't actually constrained to literally repeat itself at every point. However, a model that does repeat itself can be a useful way for saying, "We're going to take the same model and apply it to different motifs across the brain and try to capture within the same model different circuits that happen to be arising in different ways."

Basically say, "We're going to take a model that reasons about local graph structures in general and then says which one applies here or can reason about what would happen if you set up the graph structure in a different way because it's learning general principles about the local structure rather than letting every part of the brain learn its own eccentric structure. Idea of learning a general model of connectivity and then trying to apply it everywhere and make the data tell you how it needs to differentiate itself in order for the model to make the right predictions seems like it could be really powerful.

I don't know if that made complete sense, but that is something that has been gnawing at me.

Paul Middlebrooks
I was thinking cortical columns fit that.

Kim Stachenfeld
Yes, exactly. Cortical columns are-- I think one of the things these models do really nicely is you can train them on a small system and then generalize to a large system. You can train the models on a small window of fluid interactions and then generalize it to a much larger one because you're just learning these little local operations and you can arrange them into different global patterns without breaking anything, without taking the model out of distribution.

Paul Middlebrooks
It's scale-free in that respect.

Kim Stachenfeld
Yes, exactly. This ability to train on something simple, but still work on something complicated or scale to something complicated is the hypothesized property of these cortical columns. That you find a structure that a little of them is good and more of them is better. It's really the scaling property that people hope transformers have or shown that transformers have. I think that would be a really cool thing to capture if any of your listeners want to talk about that.

Paul Middlebrooks
Yes, that was inspiring. That's a great place to end it. Kim, thank you so much for taking time out of your busy schedule. It's great to hear that you're having fun doing what you're doing. I wish you continued fun, and keep producing great work. Thank you.

Kim Stachenfeld
Thank you. This was really fun.

[music]

Paul Middlebrooks
"Brain Inspired" is powered by *The Transmitter*, an online publication that aims to deliver useful information, insights and tools to build bridges across neuroscience and advance research. Visit thetransmitter.org to explore the latest neuroscience news and perspectives written by journalists and scientists. If you value "Brain Inspired", support it through Patreon to access full-length episodes, join our Discord community, and even influence who I invite to the podcast. Go to braininspired.co to learn more.

You're hearing music by The New Year. Find them at thenewyear.net. Thank you for your support. See you next time.

[music]

Subscribe to "Brain Inspired" to receive alerts every time a new podcast episode is released.