**SURYA GANGULI:** (VOICEOVER) A lot of people, of course, listening to your podcast and out there in the world have become incredibly successful based on the rise of digital technology. But I think we need to think about digital technology now as a suboptimal legacy technology. To achieve energy efficient artificial intelligence, we've really got to take cues from biology.

**[MUSIC]**

**KEVIN SCOTT:** Hi, everyone. Welcome to *Behind the Tech*. I'm your host, Kevin Scott, Chief Technology Officer for Microsoft.

In this podcast, we're going to get behind the tech. We'll talk with some of the people who have made our modern tech world possible and understand what motivated them to create what they did. So join me to maybe learn a little bit about the history of computing and get a few behind-the-scenes insights into what's happening today. Stick around.

**[MUSIC]**

**CHRISTINA WARREN:** Hello, and welcome to the show. I'm Christina Warren, Senior Cloud Advocate at Microsoft.

**KEVIN SCOTT:** And I'm Kevin Scott.

**CHRISTINA WARREN:** Today, we're excited to have with us Surya Ganguli.

**KEVIN SCOTT:** Yeah, Surya is a professor at Stanford working at the intersection of a bunch of different, super interesting fields. So, he's a physicist. He's a neuroscientist. He is actually quite familiar with modern psychology, and he's a computer scientist. And so, he's applying all of this interest that he has across all of these different fields, trying to make sure that we're building better and more interesting biologically and psychologically inspired machine learning systems in AI.

**CHRISTINA WARREN:** I love it. It’s so cool. So, I know that you and Surya are going to have like a totally serious, totally informed and thought-provoking conversation about all things AI, but I wonder if maybe, first, we can kind of go into, like, the pop culture aspect of AI, you know, with all the different movies, and comics, and various cult classics.

**KEVIN SCOTT:** That would be great.

**CHRISTINA WARREN:** Right? Because a lot of times, these things really kind of indulge our imagination, but also our fears about AI. Like, you know, I'm a really big fan of the film *Blade Runner*, the Ridley Scott classic, which is based on the Philip K. Dick book. And there are other things, like even *Westworld*, and films like *2001 A Space Odyssey*. A lot of these kind of views of AI aren't necessarily positive, right?

**KEVIN SCOTT:** Yeah. I think it's a really challenging thing because I think when you think about AI in general, like, AI done well disappears into the background, and it is a thing that exists to empower humans to augment us, to support us, to enhance us. And like, it isn't a substitute for humanity.

**CHRISTINA WARREN:** No.

**KEVIN SCOTT:** The thing that I've been really thinking about is that fiction, to a certain extent, has always been a way that we reflect the hopes and anxieties about the impact technology has on our lives. So, it's this like really beautiful way that we express our struggles with sort of this unknown future. And this imagination that we've got right now about AI, like it's sort of really playing out in fiction.

So, I remember when I was growing up, we were still in the halo of the space race, which had created this incredible canvas for all these writers and artists to make all sorts of amazing literature and films about what it would be like for humans to go beyond earth, beyond our terrestrial origins.

And the thing that really gave all of those artists permission to create the stories that they created I think is, you know, sort of the speech that President Kennedy gave in 1961 when he announced the Apollo program.

So, like if you look at the text of that speech, which I have because I've been writing a book, you know, one of the things that he -- you know, that he said in the speech is, you know, “We're going to set sail on this new sea because there is new knowledge to be gained and new rights to be won, and they must be won and used for the progress of all people.”

It's like really funny, like you read the text of this speech and you could just sort of substitute out like all of the things that he said about the anxieties that people were having over, you know, sort of rockets and the space program and all of these technologies, and like replace it with AI and the speech would totally make sense.

And like because we got ourselves unified as a country around this bold goal that we had to, you know, go to the moon and start the space race, like I think it gave this incredible permission to all of these artists and thinkers to like sort of imagine like what this future could be. And I sort of feel like we're missing that a little bit right now with AI. Anyway we could wax poetic about all of this for a really long time, but like I –

**CHRISTINA WARREN:** We could go on and on, but—

**KEVIN SCOTT:** And on and on--

**CHRISTINA WARREN:** But we have -- we should probably meet our guest, right?

**KEVIN SCOTT:** Yeah, let's talk to Surya.

**[MUSIC]**

**KEVIN SCOTT:** I'm very pleased to introduce today's guest, Surya Ganguli. Surya is an Assistant Professor of Applied Physics, and by courtesy of Neurobiology and Electrical Engineering at Stanford University. He's considered by many to be one of the leading experts in the field of artificial intelligence. Welcome, Surya, and thanks for being on the show.

**SURYA GANGULI:** Yeah, thanks for having me.

**KEVIN SCOTT:** So, I got to be familiar with your work through the work that I've started doing with Stanford's Human-Centered AI Institute. And so, we're going to get to all the cool stuff that you're doing right now, but I'm super curious. You’ve got this crazy interesting educational background where you just -- you seem super curious. Where did that come from? How did you start in tech as a kid?

**SURYA GANGULI:** Yeah, that was kind of my misspent, wayward youth. (Laughter.) So, you know, I kind of always wanted to be a scientist. And then, you know, I grew up in Irvine. I read all the books in the public high school about artificial intelligence, and they were all written by old professors at MIT. So, I kind of wanted to work in AI even in high school.

So, I went to MIT. I took my first AI course. This was, you know, end of the '90s, or so. It was all old-school expert systems, logic-based systems, all that.

**KEVIN SCOTT:** And what year was this?

**SURYA GANGULI:** This was around -- in the middle of the '90s, essentially.

**KEVIN SCOTT:** Okay, great.

**SURYA GANGULI:** So then, you know, I asked my professor, "Shouldn't we try to reverse engineer the brain?" and I'll never forget his answer. He told me, "No, no, no, Surya. Just ignore the brain. It'll just confuse you. All you've got to do is figure out the software program the brain is running.” And even as a freshman, that didn't feel right to me.

And so, I wasn't sure I wanted to do AI anymore. I stuck with a CS degree, but I had friends taking math and physics courses, and I kind of enjoyed that. So then, I ended up triple majoring through, just serendipitously, in math, physics and computer science.

**KEVIN SCOTT:** How do you serendipitously triple-major?

**SURYA GANGULI:** I just took courses for fun, and then I checked the requirements for the courses in my junior year, and I realized, okay, if I just take the junior physics lab, which is a terrible experimental course -- I mean, it's a good experimental course, but not if you want to be a theorist -- then, I could get a physics degree and the math degree would come for free, based on what I'd done. And so, I just, you know, I sucked it up and took the experimental physics course.

**KEVIN SCOTT:** That's awesome. And were your parents scientists or engineers?

**SURYA GANGULI:** No, my dad was an engineer. Yeah, he's a mechanical engineer, and my mom was actually a philosophy major in undergrad. And then, she became practical and became a certified public accountant for the IRS.

**KEVIN SCOTT:** Oh, interesting. And so, were they encouraging you to pursue a particular path?

**SURYA GANGULI:** Yeah, they were -- they were super encouraging. I think my dad was, like, an amazing mentor. He still has quotes that he would say all the time that stick in my head, like we were watching a video about neurosurgery, right? There was a neurosurgeon who was working his butt off doing this surgery. He was tired and he made it happen. And my dad said something like, "Surya, the power of a concentrated mind is limitless."

**KEVIN SCOTT:** Wow.

**SURYA GANGULI:** And it just, it just stuck. And he was full of these kinds of inspirational things. And I was kind of naturally inclined to study science and math, and he totally, like, encouraged that.

**KEVIN SCOTT:** What was the most interesting course that you took as an undergraduate? So, you were sampling, like, this great breadth of things. What -- what was the most interesting?

**SURYA GANGULI:** You know, I loved quantum mechanics. It was amazing. And then --

**KEVIN SCOTT:** And what about it really interested you?

**SURYA GANGULI:** It was just the, just the power of mathematics to penetrate into the microscopic world in a way that human intuition could not. And then, slowly, you think about it, and think about it, and think you gain intuition. And, and it was just amazing.

And then, you can predict how the, how the microscopic world will evolve, verify those predictions. Just the power of mathematics to penetrate nature I really felt that for the first time as an undergraduate then. Because up until then, everything was studying topics that you could sort of reason about, like electricity, magnetism, waves and so on. But quantum mechanics was different.

**KEVIN SCOTT:** Yeah. It's still something that I have a hard time getting my head fully wrapped around. I mean, at least for me, you sort of nailed it in the -- your normal intuitions as a human being with a normal set of human experiences are all wrong for, like, trying to understand the quantum world.

**SURYA GANGULI:** Yeah, and you construct stories in your mind. But a famous physicist once said, "If you think you understand quantum mechanics, then you don't." (Laughter.) So, it's this kind of Catch-22 nature to it.

**KEVIN SCOTT:** Yeah, I've got the Feynman lectures sitting, sitting by my favorite reading chair at home, and when I’m feeling especially energetic, I will grab one of them and, like, try to understand. And I think, you know, that was one of the geniuses of Feynman is, like, he was so good at relating these, like, very complicated, in many cases, completely non-intuitive concepts in a way that you could understand.

**SURYA GANGULI:** Absolutely. I love those lectures. I lived off of those lectures when I was an undergrad. It was super fun.

**KEVIN SCOTT:** That's awesome. And so, like, this spark that ignited for you around quantum mechanics and physics sort of led you to pursue a graduate degree in physics, right?

**SURYA GANGULI:** Yeah, on a whim I decided to not go to graduate school in computer science. I decided to go to graduate school in physics because I felt like it'd be fun. And on a whim, I decided to do my PhD in string theory because, at the time, it felt like the most fundamental subject in physics.

**KEVIN SCOTT:** (Laughter.) So, so explain to our audience what string theory is. Explain to me what string theory is. (Laughter.)

**SURYA GANGULI:** Yeah, so -- so basically, you know, we have these twin theories, right? General relatively, which governs the curvature of space, of space and time on cosmic scales. And we have quantum mechanics, which governs the temporal evolution of the NanoWorld, right, on very microscopic scales. And there's no one theory that can really unify those together where you can get both gravity and quantum mechanics together.

String theory is one where each particle becomes a little string, and different modes of vibration of the string become different types of physics. Like, one mode of vibration becomes the graviton. Another mode of vibration becomes photons, and so on. And so, in a quantum mechanical way, so you can really unify these two things. And it's a mathematically self-consistent theory. It's absolutely beautiful and very difficult to connect to experiments.

**KEVIN SCOTT:** Yeah. One of my colleagues at Google was, had gotten his degree in string theory at Stanford.

**SURYA GANGULI:** Oh, was this Yamaton Zummer?

**KEVIN SCOTT:** Yeah. (Laughter.)

**SURYA GANGULI:** Yeah, I was in classes with him. Yeah, a super fun guy.

**KEVIN SCOTT:** (Laughter.) And, like, he did absolutely nothing with string theory at Google. But he was, like one of the more interesting and brilliant people I've had the --

**SURYA GANGULI:** It's great training. And to be honest with you, if I had to do everything over again, I would still do a PhD in string theory as well.

**KEVIN SCOTT:** That's interesting. And string theory's like a little bit out of, out of vogue right now, or is that a controversial thing to assert?

**SURYA GANGULI**: It's evolving, actually. I mean, I still talk to the Stanford string theorists, to some extent, and there's very interesting ideas about holography and tensor mathematics. And now, applying string theory quantum condensed matter systems in ways that, like you can take table top quantum condensed matter physics experiments and describe them using a dual viewpoint using general relativity. And that, that idea came out of string theory.

**KEVIN SCOTT:** That's super interesting.

**SURYA GANGULI:** Yeah.

**KEVIN SCOTT:** So, so PhD in string theory. Then what?

**SURYA GANGULI:** Yeah, so you know, I kind of had a battle in my soul between do I want to become a pure mathematician or a scientist, right? If I stayed in string theory, I worked on the more mathematical topics in string theory.

But part of me always wanted to connect to nature. Everything was driven by understanding, you know, nature, right? And I didn't feel like I could really connect to nature in string theory.

So, I took my first course in computational neuroscience at the end of my PhD. It was a fascinating course taught by great professors at Berkeley, Yang Dan and Frederick Tunison, Bruno Olshausen, who discovered sparse coding, which is, you know, foundational in machine learning now.

And it was an amazing course. It was all about reverse engineering the brain, trying to understand it, and so on. And I just completely fell in love. And that was what my freshman kind of soul was yearning for. And I only discovered it at the end of the PhD and then, I couldn't even look back. That's what I did all my post-docs in.

**KEVIN SCOTT:** It’s sort of interesting. I mean, I was watching the talk that you gave at the launch of the Human-Centered AI Institute at Stanford. And you said this thing that I think is really one of the impediments to, like, getting people to more fully embrace the connections between neuroscience and machine learning, which is, like, all of us, based on our background, sort of reduce the complexity of problems in ways that are sort of convenient to our training.

And so, like, yeah, you were sort of saying that, you know, computer scientists, like when we think about artificial neural networks, we have this very reductionist way of looking at you know, how to model a synapse. And it's like this single scaler. It's like this weight. You know, when we talk about billion parameter models, you know, it's sort of like the very loose moral equivalent of like a billion synapse system. But like, we model it with a single scaler. And if you look at the neurobiology of what a synapse is, it's like this incredibly complicated system.

And so, I'm just sort of interested. Like, you go from this discipline, string theory, where like a part of what you're trying to do is, like, develop this beautiful, elegant model of the universe, and like, you jump into this neuroscience world where everything is just complicated and messy. So, like, how -- how did you go from one to the other? Like, they just seem to be like opposing ideas to me.

**SURYA GANGULI:** Yeah, it's easy. Polia had a famous statement. Whenever you're attacking a new problem, you both know too much and too little -- too much of the wrong thing, and too little of the right thing. So, I actually decided to temporarily forget my training in physics. And I went straight to USCF, a medical school, to do my post-doc, where I was surrounded by experimental neuroscientists.

And I really spent a lot of time trying to understand how experimentalists think. What are the questions that they're interested in? What are they asking? What's important to them? Because your success as a theorist in any field will partially be determined by your ability to change what the experimentalists do, what the practitioners or engineers do, and so on.

**KEVIN SCOTT:** That -- I mean just from a career perspective. Like, let's forget about all the complicated science stuff. Like, that was sort of a brave thing to do, right? Like, I mean, for like listeners who haven't been in academia, like academia is a very, in many ways, a rigid system. Like, you go get your degree. You get your PhD. Like, hopefully you can jump into tenure track position. You know, you may use a post-doc as a thing that gets you to -- and, like, switching disciplines, that is a horribly risky thing to do. Like, how -- like, where did you get the courage to do that?

**SURYA GANGULI:** It's either courage or idiocy. You know, I never worried too much about the future, the long-term future. Honestly, what I was thinking then was I'm pretty sure I don't want to do string theory. I'm super excited about this neuroscience thing. Somebody had offered me a fellowship to just learn about it for several years.

And I was like, let's do it, and if it doesn't work out, Wall Street is always waiting. (Laughter.) But I was not that excited about that at all. So, I, I just kind of jumped in. Honestly, I thought this is kind of ridiculous. It's probably never going to work out. But at least, for the next couple of years, I can have a lot of fun.

**KEVIN SCOTT:** Yeah, there's like the career risk. There's also, you know, you are sort of making yourself vulnerable, in a way, right, because you spend all of this time accumulating expertise, and with this PhD in physics, and now you're jumping into this brand new domain where you have to go bootstrap yourself again. Like, that's also a thing that people sometimes have a hard time doing.

**SURYA GANGULI:** Yeah. I mean it does sound like a lot, but actually, there is a well-worn pathway from physics to many other fields, especially neuroscience. If you look at a lot of the top theoretical neuroscientists out there in the world, a lot of them are actually trained in physics, to begin with. And increasingly, they're being trained in computer science.

And so, what was really nice is that the neuroscientists were super welcoming. They needed the help of quantitative people. So, there's lot of opportunities for people trained in the quantitative sciences -- computer science, physics, mathematics -- to really make an impact in neuroscience. And you can hit the ground running pretty quickly, compared to what it takes to do research in string theory.

**KEVIN SCOTT:** And so, as you entered this brand-new field as a post-doc, what were some of the interesting connections that you were able to make that were only possible because you had this you know, sort of unique point of view and background?

**SURYA GANGULI:** You know, what I always kind of thought about slightly differently from some of my colleagues at the time was, you know, thinking about a high dimensional dynamical systems view of the brain. So, my first project in string theory that led to a publication in *Neuron* was how to monkeys pay attention, right?

We get distracted by bottom up distractors. We also have top down attention. Monkeys have both of these things. They can both focus on a particular location of space and get distracted. And there was some strange neural dynamics that was occurring in the monkeys' brain, the part of the brain that allocates attention. And nobody could understand it.

When I attacked it, I thought about it from a higher-dimensional perspective, and that cracked open why the brain was operating that way. So, to make a long story short.

**KEVIN SCOTT:** Interesting. Yeah, well actually, let's go into the long story. So, like, how do you approach a problem like that? So, even though a monkey brain is not quite like a homo sapiens' brain, but it's still a very complicated mechanism. How do you like even get the data that you need, or like to build a better understanding or, like, a quantitative model of what's going on?

**SURYA GANGULI:** Yeah, so I had fantastic collaborators. Mickey Goldberg and Michael Shadlen. These are experimentalists who can record many, many neurons from the brain. In this particular experiment, they recorded from the parietal cortex of the monkey, which is sort of -- it has a map of visual space. And there's patterns of activity in one-to-one correspondence of locations of visual space. And wherever the pattern happens to reside is where the monkey is allocating attention

And so, you can make this bump move around -- this bump of activity -- move around in the brain by flashing distractors. You can make it move around in a top down fashion by having the monkey allocate its attention by doing a task at a certain point in visual space.

And so, they did both of these manipulations while they recorded from the brain. They had lots of neurons. They didn't have a simply theory for why the dynamics and the neurons (crosstalk/inaudible).

**KEVIN SCOTT:** And the recording is something like --

**SURYA GANGULI:** Electrophysiology recording. So, they stick electrodes into the brain, and they measure -- they eavesdrop on the electrical signals that neurons emit when they fire.

**KEVIN SCOTT:** Gotcha. So it's not directly observing the firing, it's sort of observing some sort of secondary affect, like a bunch of things firing?

**SURYA GANGULI:** Yeah, it's pretty close. So you have electrodes. They eavesdrop on a small number of neurons near it. And you can de-mix that signal because each neuron firing has a different shape.

**KEVIN SCOTT:** Interesting. It's super fascinating. I've always wondered, you know, one of the things that has driven a ton of progress in machine learning over the past 15 years is that a lot of our systems are benefitting from things that are growing exponentially fast. So, like, data for training, the compute that you're using to run the training. And you're able to do large scale experiments with, like, very quick turnaround.

And so, like, you sort of take all of those together and you can turn the crank on an experimental cycle really quickly, and you know, just sort of drive to larger and larger scale in your models. But like, when you're doing these biological experiments, you're sort of missing some of these things.

**SURYA GANGULI:** Definitely. Yeah, that's why there's still a lot of room in deep learning and machine learning for the small data problem where you have small amounts of data that's very expensive to collect. How do you detect patterns in high dimensional datasets where you don't have that many data points?

**KEVIN SCOTT:** Yeah. The first time that you and I chatted, you gave me a recommendation for a book to read whose title I'm totally spacing on now, but like is sitting in the front seat of my car, about sort of the design of biological neuro systems.

**SURYA GANGULI:** Oh yeah, *The* *Principles of Neural Design*.

**KEVIN SCOTT:** Yeah, and it's this like fantastic book. I read, you know, the first 50 pages or so of it, and the thing that's I think really fascinating is some of the big models that we're training right now that are the things that are sort of sensational in the world of machine learning take an unbelievable amount of power to train.

And so, we just finished training a model the other day that was sort of a three petaflop day run (laughter). You know, and so, like you're doing this run on hundreds and hundreds and hundreds of 300 watt chips and you know, it's like rows and rows of servers and data centers connected by miles of cabling.

The power envelope in these things are like a cluster. These machines might be sort of a megawatt of power consumption when they're at full utilization. And like, the human brain is what?

**SURYA GANGULI:** Twenty watts.

**KEVIN SCOTT:** It's 20 watts in the steady state. So, it's just unbelievable to me, like, what this machine that sits inside of your head is able to do relative to the things that we're doing right now that are, like you know, just the vanguard in machine learning.

**SURYA GANGULI:** Right. Yeah, we've actually been thinking about that. We've actually been looking into a way to get inspiring directions from researchers is to think about order of magnitude discrepancies between what the brain does and what machine learning systems do. You hit it upon the head that the energy -- or the power dissipation is in order of multiple orders of magnitude discrepancy.

Part of that has to do with biological systems operate very differently from our computers, right? In, in digital computation, every single bit has to flip with very, very low probabilities of error and very, very fast. So the laws of thermodynamics extract a high energetic cost for every fast and reliable bit flip.

But biological systems operate very differently. You look at them, and they look like they're noisy, chaotic, out of control. But what they've done is they've made every intermediate step of the computation just good enough for the final answer to be just accurate enough, thereby not spending excess power at intermediate substance computation.

A lot of people, of course, listening to your podcast and out there in the world have become incredibly successful based on the rise of digital technology. But I think we need to think about digital technology now as a suboptimal legacy technology. To achieve energy efficient artificial intelligence, we've really got to take cues from biology.

The other aspect of what you just asked, the data hungry-ness of current AI algorithms, I suspect that's because the existing framework of training bigger models on bigger datasets might be a little bit like climbing a tree to get to the moon if the moon is considered the goal of like a general intelligence, right?

I've looked into the numbers on this. If you look at the data requirements of AI systems compared to humans, you know, like AlphaGo Zero. It practiced about 33 million games, right? So if a human were to play 33 million games, it would have to play -- I did the calculation recently -- around I think it was 300 games a day every day for 30 years, right? And our top Go Masters do it in much less, right?

And then, these systems probably won't be able to generalize. If you change one rule, a human would do, still, very well. Or if you change the size of a board, the human would do well, but these games wouldn't. So, I think humans have a very different way of leaning that allows them to be much more data efficient.

**KEVIN SCOTT:** So, this is awesome. What are some of the inspirations, like specific things that you have examined that are, like, sort of in this gap between sort of neurological and artificial neural networks?

Some of them are like purely theoretical things, like this sort of saddle point work that you've done, which is a way to try to accelerate the convergence of the actual numerical optimization systems that sort of sit at the, you know, core of modern machine learning training algorithms. And then, some of the things that you're doing are, you know, like much more closely associated with the biological systems.

**SURYA GANGULI:** Yeah, I can give you kind of two examples. We draw inspiration both form the physical world and the biological world, when we work on AI problems. You also directly attack neuroscience problems and physics problems, as well.

But one example from the neuroscience world is really taking the complexity of synapses seriously. Synapses are incredibly complex signal processes devices and in artificial neural networks, we just treat them as a scalar value.

If you take into account the dynamical complexity of synapses, you can create different types of artificial neural networks where the synapses can retain a memory trace of all the changes that they've undergone while solving a task.

**KEVIN SCOTT:** So, talk about that a little bit more. Like I understand a little bit about the neurobiology of, of synapses, but like, I don't think I fully understand this whole notion of being able to have this memory trace.

**SURYA GANGULI:** Yeah, I'll give you an example of that. So you know, one way that we used a potential memory trace in the synapse is to try to attack the problem of catastrophic forgetting, right? So, the catastrophic forgetting problem in artificial neural networks is that you train on task A. You learn all your synaptic weights. Then, you train on task B and you re-learn the weights, but they erase whatever information was in the weights about tasking for you.

**KEVIN SCOTT:** Yeah, and so like a good example of this is, like the early reinforcement learning systems that people were building to play videogames. Like you could play it, you could train the system that could get really good at playing Pac Man, for instance. But if you then tried to get that same neural network to play another video game, like Q\*bert, it's completely clueless.

**SURYA GANGULI:** It's completely clueless, yeah. So --

**KEVIN SCOTT:** And like, there are even some really interesting things. You know, someone was showing me a demo the other day where, like, some of the training that we can do for game playing is extremely, extremely brittle where you changed the luminosity of the screen, or like you move like an element in the game, like one little bit, and it just really hasn't learned anything general at all about the structure of the game.

**SURYA GANGULI:** Yeah, that's a lack of robustness, which is another key issue with these current ML systems.

**KEVIN SCOTT:** So anyway, I interrupted you.

**SURYA GANGULI:** Oh no, no, yeah. So going back to the catastrophic forgetting problem, right, if you could have each synapse retain a memory trace of how important it was for solving a particular problem, and we developed an online learning algorithm within the synapse that kept track of its own importance at almost no additional computation cost compared to just gradient (descent?) training.

Then, as you learn subsequent problems, you could slow down the learning rates of the important synapses and speed up the learning rates of the unimportant synapses. And that way, you could learn the second task without forgetting the first task. And we demonstrated that this actually worked. So, this was kind of inspired by taking synapses more seriously, taking the potential power of them more seriously.

**KEVIN SCOTT:** Interesting. And so, this is a little bit related to this whole idea of transfer learning, right?

**SURYA GANGULI:** Yeah, exactly. Yeah. We, we also tried to use it for transfer learning. So, transfer learning is -- you know, humans are really good at learning one task, say, like, ping pong. And then being really good at another racket sport, compared to how they would have been if they had no experience with a racket, right? So, they can transfer stuff from task A to task B.

A major impediment to transfer learning is can we come up with a mathematical formulation of a pairs of tasks such that we can predict when structure in one task will transfer to structure in another task through a neural network.

And so, we were able to develop a recent theory of that that'll be presented at iClear. And what we found is, we found a mathematical formula for pairs of tasks that took into account how common the features were that are important for solving the two tasks. It's kind of intuitive, in hindsight. If there's enough common features, then transfer learning will be successful.

What's interesting is it doesn't matter what you do with the features, right? As long as there's some particular function of the inputs that's important for solving both task A and task B, but task A says you do one thing with that function, and task B says you do another, that's not a problem. So, you just need a notion of commonality in the input feature space.

**KEVIN SCOTT:** Interesting.

**SURYA GANGULI:** And we were able to formulize this all mathematically, which is super fine.

**KEVIN SCOTT:** Interesting. Yeah, this is just fascinating stuff. So, where do you think the next set of breakthroughs are likely to come from? Like, what are you thinking about right now that you're --

**SURYA GANGULI:** Yeah, we're thinking about kind of all of it simultaneously. We're really interested in sort of unsupervised learning -- you know, what can be done in that direction. We're thinking about more mathematical theories. We're trying to understand -- develop algorithms for interpreting neural networks, especially neural networks that were trained to mimic the brain, because what's happening in neuroscience is we're increasingly starting to model more complex neural circuits under more complex tasks. And our models themselves are complicated deep neural networks.

So then, you know, we're placing something that we don't understand, i.e. the brain, with an artificial network that we don't understand. And so, we're developing algorithms to do that, and we're actually using ideas from physics involving course grading and things like that where let's say you have a very, very complicated model. Can you extract from it a simpler model where the individual connections in the model, and the neurons in the simpler model are in one-to-one correspondence with what we think is already there in the brain?

And we've recently actually, there's a test case we've been working on the retina, which is a deep neural network of its own accord. And people have been showing, like, for 40 years, different artificial stimuli to the retina, and they come up with these ad hoc models for each of these artificial stimuli. Yet, nobody has come up with a really good model of the retinal response to natural scenes, the very scenes that sculpted the evolution of the retina.

We recently came up with a state of the art model of the retina that involved a deep neural network. It's a complicated model, and now we're looking inside it to see how it responds to all of these artificial stimuli.

**KEVIN SCOTT:** That’s really fascinating. I mean, one of the things that I've thought for a long while is that, even though, in many, many ways, the connection between artificial neural networks and biological neural networks is tenuous, that done right, you could do things exactly like what you're describing, and sort of use the DNNs as like a way to get insight into how the the biological systems work versus, you know, like, what we typically try to do, which is, like, derive this inspiration from the biological for the artificial.

**SURYA GANGULI:** Yeah, exactly. My colleague, Daniel Yamins at Stanford, has done some great work on that where he's come up with models of the ventral visual stream all the way from retina to V.1, to V.2, to V.4 to IT, which has these object detection cells. And he trained deep neural networks to do object classification, nothing to do with neuroscience.

And then, he looked for patterns of activity in different layers of the deep network that would optimally match the patterns of activity in different layers of, say, a monkey's brain when neural activity patterns were measured in different layers in response to the same set of objects. And he found a great match there.

**KEVIN SCOTT:** Interesting. I mean, I want to go back to this -- you mentioned unsupervised learning just a minute ago, and it is one of the things that is really very exciting right now. So, there's been a bunch of breakthroughs just over the past nine months or so with applying the techniques of unsupervised learning to natural language processing tasks. So you know, document summarization, question answering, translation, like a ton of things.

And so, you know, for everybody, unsupervised learning is this notion that you can train a set of machine learning systems. Like, they typically involve deep neural networks. But you don't have human beings in the loop providing, corrections to training, or like labels for particular things. it just sort of like learns like a general conceptual model of some domain. And then, you try to figure out how to apply it to, like, specific problems.

In some cases, like, that application is, like you -- you do transfer learn from, like, the unsupervised general model to some small supervised model that sort of specializes the, the unsupervised model to a task. But like, in some cases, and this is the work that Open AI did with their GPT-2 model, like they were doing one-shot learning. So, like, just with no supervision whatsoever, like, getting this thing to do useful things.

And so, the reason that that's super interesting, as you well know, is that one of the things that constrains how fast you can sort of scale up the ambition of classical machine learning systems is, like, this data labeling, data engineering task is very, very difficult. So how do you think about this, because like, in some ways, human beings can be very good at unsupervised learning.

Like, a toddler, just by absorbing the universe around it, can learn things way more advanced than what a machine learning system right now is able to figure out.

**SURYA GANGULI:** Yeah, exactly. Yeah. I think you hit the nails on the head with everything you said. Going back to the first example you gave, natural language processing, unsupervised learning has been incredibly useful in natural language processing because we have a simple principle for doing unsupervised -- solving unsupervised learning problem, which is predicting the next word, or maybe the next character in sequential text, right? So if you can predict, then you can understand.

And what's been really amazing is the internal representations of these neural networks used to solve this prediction are actually very useful for subsequent training unsupervised tasks. And actually, I'm working with a computational linguist, Chris Manning, actually, an NLP person.

**KEVIN SCOTT:** Yup, who's written my favorite NLP textbook of all time.

**SURYA GANGULI:** Exactly, yeah, yeah. I've been studying, actually, his textbook recently. We're jointly advising a student who's analyzing how these unsupervised trained networks work, John Hewitt. He actually came up with a really cool result that showed that if you look inside these neural networks, they implicitly build up syntactic trees associated with sentences, right?

**KEVIN SCOTT:** Interesting.

**SURYA GANGULI:** And so, what he did was he -- there's something called the dependence parse tree, where you can take a sentence and come up with a dependency parse, and that gives you a distance between all pairs of words in the sentence. He showed that he can learn a simple quadratic form going from the internal representations of these networks, Burt and Elmo, to a scalar which predicts the dependency parsed at this many words in the sentence.

And so, now what we're doing right now is we're analyzing the dynamics of this model, trying to figure out how, on an online, word-by-word fashion, it builds up this parse tree, which is an interesting computation because conventional computer science algorithms to build up dependency parse trees need a stack to do it. So, we suspect that a stack machinery in hiding in the dynamics of this network. So, playing around with that.

**KEVIN SCOTT:** That also -- yeah, that's super fascinating. I mean, like, it's interesting on a bunch of different dimensions. Like, the thing -- as you were, as you were sort of saying all of that, that I was thinking about is, like I was a compiler in programming language person when I was a younger computer scientist. And I remember -- I remember long, long, long ago reading Chomsky's work. So, like, Noam Chomsky is like a very, very --

**SURYA GANGULI:** Another MIT person. (Laughter.)

**KEVIN SCOTT:** Another MIT person, one of the most famous sort of linguists and social philosophers in the world.

**SURYA GANGULI:** That I didn't agree with, even when I was a freshman. (Laughter.)

**KEVIN SCOTT:** Yeah. And like, look, I also didn't agree with him because, like, one of the things that he asserted a long, long time ago is that human beings had some built in notion of grammar that was, like, you know, sort of in their brains. And what you just said is sort of striking, that like, even when you're evolving an artificial system, that like maybe some fundamental notion of grammar, like, manifests itself.

**SURYA GANGULI:** Yeah, exactly.

**KEVIN SCOTT:** Like, that's super interesting.

**SURYA GANGULI:** It's -- it's an emerging property of -- and we've, we've actually worked out mathematical solutions to the dynamics of learning in deep neural networks that show how hierarchical concepts can emerge naturally in a deep neural network.

So, for example, babies, when they learn concepts, they learn course grade distinctions first, like animals versus plants, even if you control for perceptual dissimilarity. And then, as they get older, they learn finer distinctions, different types of animals, different types of plants.

We were able to prove mathematically that deep neural networks have to do this when they're exposed to it. And so, a paper that's going to be appear in *PNS* soon compared our mathematical theory of deep learning and semantic cognition to many, many experiments on babies and semantic cognition. And we achieve a match.

**KEVIN SCOTT:** Interesting. Yeah, so --

**SURYA GANGULI:** Actually, I have one more thing I wanted to say about unsupervised learning. Is prediction good enough, okay? That's the driving principle -- one of the driving principles in unsupervised learning today. I don't think it's good enough. If you go back to what babies do, there's a famous experiment, right?

If you give a baby two magical objects, like object A where, if you drop it, it doesn't fall, right? Say, through a video or something. And then, object B, it seems to go through walls. You give these two objects to a six-month baby sitting on a highchair. What will it do? The object that didn't fall, it'll throw it off the highchair to check if it falls. The object that seemed to go through the walls, it bangs it on the highchair -- on the table to see if it'll go through the table.

So, this is incredible, right? Babies, even at six months, have an implicit model for the physical evolution of the world. They, they pay attention to violations of the world model, and they actively choose -- experiments to gather specialized training data to test those violations further, right?

So, this business of building world models, using those world models to imagine the future and make decisions, looking at violations to modify the world model, actively doing experiments -- that's the next frontier in machine learning, not just passive training data.

**KEVIN SCOTT:** Yeah. Totally, totally agree. I mean, like, we've gotten really, really good at prediction and classification, and we're not really great yet at, higher order things, like you know, deduction reasoning.

**SURYA GANGULI:** Yeah. And reinforcement learning is starting to get there in terms of at least formulating the problem because you have a sequential decision making problem where you have to exploit, and explore, and all that. I think the methods for exploration are not that efficient.

**KEVIN SCOTT:** Yes. Yes.

**SURYA GANGULI:** So, model based reinforcement learning is kind of the direction that people are trying to go where you use world models to, again, imagine a plan and, and learn.

**KEVIN SCOTT:** Now, so this has been a super fascinating conversation, but one of the things that I want to pick your brain on is -- everything that we've talked about so far is, like, incredibly technical, which I love. Like, you know, I could spend hours, you know, just trying to learn from you and talking about some of the super technical stuff.

But, like, one of the things -- and this sort of gets back to your involvement and my involvement in the Human-Centered AI Institute at Stanford -- that we are sort of grappling with right now is, like all of this AI stuff is increasingly having an impact on everybody's day-to-day life.

And so, you know, for the majority of folks, you know, we probably lost them like half an hour ago in this, this conversation that we're having. So how do you think about our role as scientists and engineers, and technologists in helping the public better understand this big bag of complicated stuff so that they can make good decisions about it?

**SURYA GANGULI:** Yeah, I think it's incredibly important. And more than that, even, it's, it's coming to terms with ourselves as a society, and how we can optimize the development of AI so the outcomes are good for society, as opposed to bad. And so, there's incredibly thorny issues involved in that, that touch on labor economics, and political science, and regulation, on ethics, and so on, and so forth.

So, it's an incredibly complicated set of problems, so we need to bring together, you know, in a real way, scientists from many different disciplines, like the ones I just mentioned.

And that's what HAI is actually trying to do, just to talk a little bit about the structure, we kind of have three focus areas. One is building next generation AI inspired by the power, versatility and robustness of human intelligence. And so, we bring in ideas from neuroscience and psychology, and machine learning to work together to get to that new technology

The second is building AI systems to augment the capabilities of humans. Think like building intelligent hospitals, working in domains where companies might fear to tread, like development in Africa, or things like that.

And then, the third branch, which is very important which speaks to the question that you just asked is can we guide and design the impact of AI in society. So, we're bringing in economists, social scientists, historians. Historians, for example, can study biases that exist in the training data that we're using to teach AI systems; social scientists who can study the impact that AI could have in different sectors of our society; economists to deal with the displacement of jobs that I think is proximal issue that we're going to have to deal with very soon.

So, I've been having a lot of fun in this initiative actually meeting, economists and lawyers as well. Like, how do we regulate AI systems? What are the ethics involved? If a self-driving car hits somebody, who's liable, right? These are incredible issues that we don’t have all the answers to. We need to build an institute to bring people together to convene the stakeholders to figure out, you know, what we should do. And that's what we're trying to get at, at Stanford HAI.

**KEVIN SCOTT:** I'm involved, so like obviously, I think it's a worthy undertaking. I think there's this really urgent need for storytellers to basically bridge this gap between an incredibly complicated, technical world where it's like, very, very easy (and I've allowed myself to do this on multiple occasions)--like, the thing that you, that you really, really strive for sometimes when you're a scientist or engineer is to, like, just get into this flow state where you're completely immersed in a particular problem. And like, psychologically, when you get into, into flow state, like everything else around you just sort of disappears.

**SURYA GANGULI:** Yeah, yeah. I love the flow state. (Laughter.)

**KEVIN SCOTT:** Yeah, no, we all love the flow state, and it's, like, where you do your best work. But you know, we also are like working in this discipline where, you know, it is equally important to, like, pull yourself up and connect with the greater context around you, and to make sure that you're giving enough time and energy to, like, help bring along everybody around you.

And, to also understand what all of the sort of complicated set of social concerns are with all of this stuff, and to make sure that we're sort of pointing all of this intellectual energy that we're focusing into these very interesting problems right now in ways that sort of have net human benefit to everyone.

**SURYA GANGULI:** Absolutely. So, we as professor are doing that through HAI. We're talking to media. We're talking to people who have decision making power in industry and governments. I've actually been -- you know, in the past, I've tutored CEOs of portfolio companies for a VC firm on AI, which is actually super fun -- incredibly smart people who may not have all the technical background, but they're very curious.

The other thing that we're trying to do is really nip this in the bud and train the next generation of leaders to take ethics and social science, and all these other issues into account at the very beginning, say freshman CS courses, and things like that.

We need to bring in the societal implications of AI directly -- you know, bring that into the, the consciousness of students who are going to be the next generation of AI practitioners. And that's part -- we have a very seriously educational goal there at HAI, as well.

**KEVIN SCOTT:** Indeed. So, let's change directions completely. So, you have so many interests across such a breadth of different things. So, do you have any interesting hobbies outside of your professional life that you get obsessed about?

**SURYA GANGULI:** Yeah, I used to have a lot more. I have a three-year-old kid at home (laughter), so that takes a lot of time.

**KEVIN SCOTT:** Yeah, that's the new hobby. (Laughter.)

**SURYA GANGULI:** I love, I love playing tennis. I was actually on a varsity tennis team, so I used to be a college jock, albeit it at MIT. (Laughter.) So, MIT was very proud of its football team, for example, because it made it into *Sports Illustrated* for fun stats, which is the highest ratio of IQ to weight, body weight (laughter), which means they weren't that great on the field.

**KEVIN SCOTT:** And isn't there a professional football player right now who's getting his PhD in Applied Math at MIT?

**SURYA GANGULI:** Yeah, yeah. You're right. I'm blanking on his name, but yeah, that’s absolutely true.

**KEVIN SCOTT:** Yeah, he's like super amazing. Like, I'm ashamed that I can't remember his name. So anyway, like you were a tennis player.

**SURYA GANGULI:** So yeah. I love playing tennis. I love swimming. We love hiking and so on. Yeah, just, I like getting out into nature and just, just exercising the body as oppose to the mind.

**KEVIN SCOTT:** And I'm sort of curious. I know my hobbies help me to get my brain reset. So, like, yesterday, for instance, I'm working on a book right now, and I had to get myself psyched up to finish writing the last chapter of this book, which I spent, 12 hours yesterday, like putting the finishing touches on.

**SURYA GANGULI:** Wow, you found 12 straight hours to work on it? That's awesome.

**KEVIN SCOTT:** By hook or crook. It, it involved staying up until very late last night. (Laughter.) But before I could even get my head clear enough to, like, sit down and write, I had to go do something with my hands. And like, that's sort of the way that I -- it's almost like, you know, some people meditate. I go into my shop, and I make something.

And it doesn't need to be a complicated thing. It just needs to require 100 percent of my attention. Like, if I'm not focusing on this, I'm going to cut my fingers off, or you know, something.

**SURYA GANGULI:** Yeah, yeah. You know, for me, that's swimming. Like, I, I leave the terrestrial surface. I go under the water. I swim for half an hour, 45 minutes, and I'm a completely new person. It's, it's really weird.

**KEVIN SCOTT:** That's awesome.

**SURYA GANGULI:** So when I was in grad school when I had the freedom, my schedule was roll in at 10am, right, socialize with my grad students friends, do something -- then, go swimming at 6pm. Like, I used to swim a mile a day and then show back up in the lab at like 8pm. You know, a bunch of grad students are still there. Then, work till 2 or 3am. That was great. And then, roll in at 10am.

I can't do that now. Now, I wake up at 4 or 5am before my kid wakes up, and I can barely make it to the pool because I have to be home by 6:30pm. (Laughter.) So, you know, it's tough.

**KEVIN SCOTT:** Yeah, it's really interesting. I, I used to think when I was a graduate student, that I was very busy.

**SURYA GANGULI:** Oh, yeah.

**KEVIN SCOTT:** And, and if I had known what my life was going to be like at 47, like I would have --

**SURYA GANGULI:** I tell my grad students in post-docs that the best time in the academic chain is grad school or post-doc. Actually, post-doc, I think is the best. They don't believe me. (Laughter.)

**KEVIN SCOTT:** Yeah. No, and like, I wouldn’t have believed you either because I thought my life was crap. (Laughter.)

**SURYA GANGULI:** Yeah. You value different things at different stages. Yeah.

**KEVIN SCOTT:** All right. Well, thank you so much for being with us today. This was an awesome conversation.

**SURYA GANGULI:** Thanks. It's my pleasure. Thanks for having me.

**[MUSIC]**

**CHRISTINA WARREN:** Well, thanks for joining us for Behind the Tech. That was Kevin Scott speaking with Surya Ganguli.

**KEVIN SCOTT:** Surya is this -- yeah, as everyone just heard, like this really, really brilliant polymath.

**CHRISTINA WARREN:** Yeah. Well, one of the things, you know, that you two were talking about towards the end of your conversation that I found really interesting is, you know, we're getting all these really intelligent models. We're starting to train these things in really interesting ways. We're able to do a lot of really great things. But with that comes obvious ethical questions.

**KEVIN SCOTT:** Yes.

**CHRISTINA WARREN:** So, how do we ensure that what we're building is going to be used in a way that doesn't turn into that dystopic, you know, vision of, you know, the sci-fi novels that have defined, you know, our entertainment or even worse, had like a negative impact, you know, on our -- on our lives?

**KEVIN SCOTT:** Yeah, I think one of the most important things that we can do is to try to make some of the complexity of the field more accessible. So, it can't be the case that a small number of people -- I won't even call them elite, because like in a way like being clever and, you know, sort of having the willingness to sort of commit yourself over a long period of time to accumulate a bunch of expertise is like admirable, but like it's also like not this unattainable sort of thing.

And so, but like what we have to do a much better job of than we've been doing throughout the history of AI as a defined discipline, is making sure that we are able to convey some of the complexity of the field to other folks, so like we have a big swath of people participating in the conversation about AI in a rational, reasonable way.

You can't just expect like a bunch of scientists and engineers to universally be able to make a set of good decisions on behalf of the rest of society. Like everybody needs to have a voice in this thing.

**CHRISTINA WARREN:** Yeah, you're definitely right, we need to have more voices. And we need to be more transparent with how these things are being built, too, right, because, I mean, I think that that's one way that you do both A) improve the models that are being built, but B) and correct me if I'm wrong here, but it seems like maybe it would make people more comfortable if they had better insight into what actually is happening.

**KEVIN SCOTT:** Yeah, I totally agree. And like the irony of this field is like in a way it is a very modern discipline. So, in some senses it has far more transparency than some of the scientific or engineering disciplines that have preceded it, because you have open source software where people are able to take the code that they're writing to create some of these models and share them with the rest of the world, like you have transparency there, because so much of this data that people are using to train models exists on the open Internet, you have transparency there.

And, you know, because like publications in this area are so much freer now than they've ever been before at any point in human history, like you've got a real transparency about the ideas.

You know, we have some work to do there, like, you know, there's sort of a reproducibility crisis that you've got happening across all of science right now where the experiments and -- and results that people are publishing are becoming increasingly difficult for other people to replicate, and I think that's certainly a problem here in AI.

But like I do really actually believe that, you know, that a big part of the transparency here is about like awareness, because like I can make arbitrarily complicated things super transparent, like I can send my mom the proceedings of the -- you know, the big deep neural network conference, NeurIPS, the Neural Information Processing Symposium, and like she's not -- that's not going to help her be involved in the conversation about AI, like even though it is like the very foundation of, you know, transparency about what's happening in the field right now.

**CHRISTINA WARREN:** Okay, so what we need is like a Netflix show or an HBO show that shows the reality of the situation, right?

**KEVIN SCOTT:** Totally. Like we have got to figure out like how to get people more excited about connecting with this stuff, and we've got to get my posse of folks, like the scientists and engineers, like more excited about boiling these things down to their like clear understandable essence.

**CHRISTINA WARREN:** All right, so we're going to write a screenplay, but until then -- that's what we're going to do, Kevin, but until then, I think that is all the time that we have for today.

**KEVIN SCOTT:** Yeah, indeed.

**KEVIN SCOTT:** And be sure to join us next time on Behind the Tech when we speak with Fei-Fei Li. Fei Fei is the Co-Director of the ***Stanford Human-Centered AI Institute*.** I hope you'll join us.

**CHRISTINA WARREN:** Yes. And please help spread the word about this podcast. See you next time!

**[MUSIC]**