Hi, I'm Craig Smith, and this is Eye on AI. Terry Sejnowski, an AI pioneer, chairman of the NeurIPS Foundation, and co-creator of Boltzmann Machines - whose sleep-wake cycle has been repurposed in Geoff Hinton's new Forward-Forward algorithm, talks in this episode about how advances in deep learning, particularly in the scaling of large transformer models, may help us understand our own brains.

Terry is a wonderful communicator and makes these complicated concepts accessible to everyone. I hope you find the conversation as interesting as I did.

**CRAIG:** Normally I have the guest introduce themselves. You're certainly somebody that people know. But nonetheless, for uninitiated listeners give a little bit of your background and then we'll talk about how AI has developed, and what is happening now in deep learning and where you see it going. So, Terry, introduce yourself.

**TERRY:** Well, hello, my name is Terry Sejnowski and I'm currently a faculty member at UC, San Diego, where I'm the director of the Institute for Neural Computation. I also direct the computational neurobiology laboratory at the Salk Institute.

**TERRY:** And my primary field now is computational neuroscience. That is to say of trying to focus the tools that we've developed here in the NeurIPS community to try to help us understand how brains work, which is a very major scientific problem. Of course, there are engineering spinoffs that, come out of that as we learn more about the brain that's going to help us create better network models.

**CRAIG:** And that link to neuroscience between deep learning and neuroscience. Can you talk about how the NeurIPS conference developed, and why it developed.

**TERRY:** The conference grew out of a movement that began in the 1980s, which was actually instigated by physicists and people in computer vision from many different areas of science and engineering, which was a dissatisfaction with the current tools that were available at the time for handling really difficult problems.

**TERRY:** large scale problems, things like computer vision, speech recognition, language translation. And one of the things they had in common was that they were very, very high dimensional in terms of all of the variables in the features and the traditional tools were based on trying to develop relatively small models.

**TERRY:** In statistics, you have a small model with few parameters, and you try to find the optimal set of parameters. We were going in the opposite direction. We wanted to come up with large scale models that had many parameters and instead of trying to find the optimal set of parameters, let's find a set of parameters, which didn't have to be unique, but nonetheless gave good performance.

**TERRY:** So that was our initial impetus and it kind of snowballed in the sense that our first meeting that was open to the public, we had a bunch of earlier meetings, small groups that got together informally, but it was always self-generated, and we focused on learning.

**TERRY:** Very early it became clear that how do you find the parameters? You've got to find the parameters somehow. So, by having lots of data and by using the data to gradually improve the performance through optimization, was the approach that we took. So, the first meeting that we had was under the aegis of the Neural Information Processing Systems Foundation that was founded by Ed Posner at Caltech.

**TERRY:** And at the first meeting in it was 1987, had a total of 90 submissions and there were a few hundred people who attended it. and it was really exciting.

**TERRY:** Why? Because we hadn't been together that many of us hadn't been together at the same time. And it was unbelievably diverse in terms of where people were coming from. They're coming from math theory, people from statistics, people from cognitive science, from computer science, from neuroscience. It was like the ultimate interdisciplinary group.

**TERRY:** And that's very rare as you know, because we all have these silos that we live in and each one of the disciplines has its own specialized meetings. But we realized that we were tackling some difficult problems that were common to many, many fields. And so, we could help each other by coming together and sharing some of our algorithms, insights, data sets and so forth.

**TERRY:** So that was how things got started.

**TERRY:** Like I said, was started by Ed Posner, but he tragically got hit by a truck when he was bicycling and so I was called and asked if I wanted to take over as the president of the foundation, and I agreed and little did I know where it would go,

**TERRY:** But it's been exciting to see how this meeting and the workshops especially have grown enormously over the last 10, 20 years. And it's also interesting how it has morphed. So it started out as a focus on models that resembled networks in the brain, Hopfield networks, for example, which were relatively by today, standards, small networks, but they exhibited some of the non-linear dynamics that we see in a recurrent network that you need in order to recover complete memories where you're given a little bit of information you want to complete that, get of the entire output out. But it very quickly grew in the sense that many other algorithms were discovered.

**TERRY:** And as they were added year after year, graphical models, support vector machines, Bayesian networks, they quickly became clear that we had here a toolkit. And that became machine learning as the kind of the central focus although again, these are tools that a lot of people were using and adding to. And then about 10 years ago, a NeurIPS meeting that was held at Lake Tahoe, Geoff Hinton had a paper on which he applied convolution networks to ImageNet, which at the time was the biggest and best image collection, for categorizing tens of millions of images into 10 of thousands of categories. And the improvement there was so dramatic that it got picked up by the New York Times and a lot of other outlets. And within a few years, computer vision was transformed because of the fact that the performance was so much better than the traditional approach, which was by hand crafting features where performance improved by half a percent per year. And it was like overnight, 20% improvement. The next year, 10% reduction. And now, it's, taken for granted. But back then it was really a wakeup call.

**CRAIG:** That AlexNet and the ImageNet competition was kind of the kickoff for deep learning, and the conference has really zeroed in on deep learning. Right?

**TERRY:** I disagree with that.

**CRAIG:** Okay.

**TERRY:** And I'll tell you why. Because if you just go through the posters, you'll discover that there's still a great diversity of algorithms and approaches that are still being explored.

**TERRY:** And if you really look at the way that the meeting has evolved, almost four decades now. There's always been, at any given time, the tool, the algorithm du jour, which everybody got excited about, and a lot of papers came out. I remember support vector machines took over NeurIPS for about five years.

**TERRY:** And it takes its course. That is to say it's explored and the limits are found. A lot of applications are solved at various levels, and then something else pops up that was being, incubated at NeurIPS in workshops or something. And then that takes off. And then, you know, every five years you get a new breakthrough in another area.

**TERRY:** And each time that happens, it's a bigger breakthrough in the sense that it covers more applications and has more flexibility. I see deep learning as just another step along that direction. I think we have a long way to go. A much, much longer way to go. The good example, okay, is, transformers, right?

**TERRY:** So, transformers are another form of a neural network, but it's gone from feed forward to recurrent to these transformers with self-attention and every time you improve the architecture or enhance it, enlarge it, new capabilities appear, right? And so, I expect that to continue.

**TERRY:** And I think that the space of algorithms out there is infinite. We're just scratching the surface.

**CRAIG:** Yeah. That space of algorithms out there being infinite is something I've often wondered how you think about it or people in the field think about it, particularly with regards to neuroscience and this impulse to build a model of the brain or figure out how the brain works or build a, an artificial analog to the brain. Do you think these algorithms, as in mathematics, they exist in nature and we are discovering them, or do you think that we're creating them?

**TERRY:** A little of both. Like the Boltzmann Machine was a version of the Hopfield net. We basically heated it up so that it was fluctuating. And from that we got a learning algorithm with Geoff Hinton, and we had wake and sleep phases during which you gave it data that you were trying to train on. Then he let it free run and he subtracted the two.

**TERRY:** And Geoff Hinton is coming back to that. So, it's still, that concept is still permeating. And by the way, you know, humans also have wake and sleep cycles too, right? But what we're looking for are principles that we can extract from nature. And one of the principles that is probably the most important principle is this principle of scaling.

**TERRY:** So, all algorithms, as you increase the size of the problem, they scale with some exponent, for example, if you're doing some kind of a search and you have to compare all pairs that goes as N squared with N objects that you're searching, trying to find the optimal one. But the problem is that, as N gets bigger and bigger, a million, a billion, which is where we are today, then it just runs out of steam.

**TERRY:** You run out of memory it's not practical. And most of the algorithms in traditional AI have that problem. And when we started in the eighties, we had tiny little networks. I worked on something called Net Talk, which was text to speech. And by today's standards it had a couple hundred units in it, maybe few tens of thousands of parameters, weights.

**TERRY:** But it did something which was really quite remarkable. It took something that is as irregular as English pronunciation, and it learned the regularities and the exceptions. And there are many exceptions, as you know. And so, it is an architecture that absorbed all of the complexity of phonology, it's called in linguistics.

**TERRY:** And that was a shock because, linguist at the time were using rules to explain all the regularities. So, they had books with thousands and thousands of rules, and with each rule there was lists of hundreds and thousands of exceptions. The whole classes of exceptions, like French derived words, but then within the exceptions there were rules.

**TERRY:** But then within the rules, within rules, there were exceptions. You know it's rules all the way down and it's, that's a very awkward representation. First of all, it takes a lot of effort because you have to figure out what the rule is and then you have to find how to deal with it. So here we go. And now with learning, it seems like we can do both.

**TERRY:** We don't have to separate it and it doesn't take labor. That was the big, ultimately savings is that Instead of, human manual labor figuring out the rules and working out the details of the features, you could automate that and that could be applied to any problem. It's not that each domain has a separate set of rules.

**TERRY:** It's like, we could learn them, and now here's what we didn't know back then. The eighties we had these small networks, and we had the learning algorithms, the same ones we have today by the way.

**TERRY:** But we did know how well it scaled. What would happen if you scaled it up by a factor of a thousand or a million? Well, now we know. It turns out they scaled beautifully. And this is a principle that we actually had observed in nature. If you look at the size of the cortex of different species, what you discover is that, especially in primates, more and more the volume of the brain is devoted to the cortex, cerebral cortex that is on the outside and in humans, it's so abundant that it gets convoluted, it gets all these folds and so it looks like a walnut from the outside. Right? It's a very, very dense because you need more surface area.

**TERRY:** You got to get the same surface area and the same volume. And so how do you do that? And it's very, very rare. Even in biology, because most of the other parts of the brain don't scale that way. So, there's something special about Cortex. More is better. And that's what we've discovered about some of these neural network models, that it scales linearly.

**TERRY:** That is to say as you add more units, more parameters, just scales with that number. It's not N squared, it's not N cubed, it's N.

**TERRY:** Biology has this additional advantage that you can do it all in parallel. In other words, you have N neurons. That's a big number. It's a hundred billion, but they're all working together in parallel, which means that it's order one, doesn't matter how big your brain is, it's working in real time, or you know how small they all work in real time because of the fact that they scale.

**TERRY:** It's scalable algorithms. So, these are the principles we're looking for in nature is scalability. We're looking for ways to reduce energy. Nature manages to run our brain on about 20 watts, right? These large-scale computers now that are being used to generate transformers, right? They're megawatts.

**TERRY:** We're talking about a completely different order of magnitude. Six orders of magnitude. Everybody knows that, look, we can't continue to scale up. We're not going to get a power plant, a gigawatt power plant for creating networks. We need to be able to have a technology like the ones that the brain uses for being able to scale within a reasonable energy budget.

**TERRY:** That's where we are. We're extracting principles. And there, many others now that we found similar to what happened, by the way, early in this development of man powered flight. In the early days of AI, everybody laughed at us because, oh, what are you going to learn about artificial intelligence by studying the brain. what are you going to learn about flight by studying birds, right? Ha ha ha. The Wright Brothers spent a lot of time looking at how birds glide because somehow without much power, they're able to go a long distance.

**TERRY:** Right? Well, that's what they were trying to do. And the other thing they looked at the feathers, very light, but very strong, stiff. So, they had their plane not built out of metal, but out of wood spars with canvas in between. Very stiff light weight. So, this is how you extract principles from nature and then incorporate it.

**TERRY:** You have different materials, you have different ways of accomplishing the same thing, but you get inspiration. You have a new way of thinking about the problem, and that's what we're doing with the brain.

**CRAIG:** On the scaling problem. So, there's no end in sight in terms of the improvement as you scale parameters, but there is an end in sight, and I think we're close to reaching it in scaling the power required to run the GPUs.

**TERRY:** So, it's interesting because what's happened is that there's so many applications and there's so many people are using the traditional hardware is von Neumann architecture. trillions of dollars, have been put into developing, fast and cheap computers using the von Neumann architecture.

**TERRY:** But now what's happening is that the hardware manufacturers are beginning to realize, oh, we can build these massively parallel networks better, much more cheaply than simulating them on von Neuman n machine. And not only that, but it's less power so that there's going to be a whole set of generations of hardware that is going to come out.

**TERRY:** And it's already out there. There's one company that is building a wafer scale, Cerebras, and has to be water cooled because it's a lot of cores. But the point though is that It's much more, much more efficient in terms of power usage and for the architectures that we're using right now, and that will continue

**CRAIG:** and more than the hardware to me.

**CRAIG:** I had a conversation yesterday with a young PhD student about data pruning and different strategies for identifying what you can throw out and still train efficiently. And as you prune the data used for training, you use less power. So that's, that sounded promising to me.

**TERRY:** See, what's happened now is that there's like thousands of people out there who are figuring out how to simplify, how to reduce the training time or the training set.

**TERRY:** There's a very powerful method called distillation, which is interesting. So, it turns out you need a huge network and a huge amount of data to train it. But once you've trained it, you can now take the input and the output of that complex network and use it to train a much smaller network. It's called distillation, and you use it to train it not just on the correct output, but on the probability distribution of the output.

**TERRY:** And you can come pretty close to having the same performance for much, much less hardware. What that's telling us something, an important principle, which is that maybe why we have such a huge brain is not performance, it's for learning from this huge amount of data in the world, but once you have distilled it, somehow you can now run it really fast with much less hardware.

**TERRY:** And that opens up new resources in the cortex for learning new things. And it is actually evidence for this, which is really a mystery for a neuroscientist, is called drifting representations. So, it was always thought that, well, if you have a neuron, you're recording from it someplace in the visual system.

**TERRY:** That it has some feature that responds to like a line vertical line that will be there fix forever. It turns out interestingly, that now we can record over long periods of time for weeks, for months, and we discover that, oh, the feature responds to changes. It's no longer vertical. Now it's some other angle or some even more complex, stimulus.

**TERRY:** And that's called drifting representations. And what is that telling us?

**CRAIG:** The feature changes or its location in the brain?

**TERRY:** The same neuron distracts optically over many, many weeks and months. And if you just do the same experiment over and over again, you discover that you get a different answer every time over time.

**TERRY:** And that has shocked us because, it's a very different view because now it means that the cortex is not fixed piece of hardware, like a digital computer, right? Where each transistor has a function and it's going to stay that way. Right? Unless it breaks. No, it's the whole brain is shifting all the time.

**TERRY:** It's malleable, it's adaptable. As the environment changes, it shifts. but it takes time, right? It takes, changes occurring over times that are long enough so that you wouldn't notice it if you're just doing a single experiment or just a few experiments.

**TERRY:** You'd have to monitor it. So that's a big shock. And now we're beginning to think that gee, this is one of the problems we have with these networks is you train them up and you use them, but they don't longer learn. Right? So, you need something called lifelong learning. Well, we, our brains do lifelong learning.

**TERRY:** Maybe this drifting representation is an indication of how we do that. So, this, again, we're extracting a principle from something we've observed in the brain and that could generate the next generation network that based on similar principles.

**CRAIG:** You said that Geoff Hinton is bringing back Boltzmann machines.

**TERRY:** Okay, so first of all, let me tell you the history. The history there was that my background was in physics and neuroscience and his was in psychology and AI.

**TERRY:** So complimentary, and we just hit it off. We were at a small meeting, 1979. It was a workshop at San Diego, and we both had the same beliefs about how nature was the only existence proof you could solve difficult problems like vision and speech in language. So, we should be looking carefully at how nature solved those problems and trying to understand that in a way that I was telling you about the Wright brothers, although we weren't thinking about the Wright brothers back then, but it's clear that this is something that has happened over and over again in history.

**TERRY:** And we came across just like I say, through the Hopfield net, heating it up, and looking for learning algorithms. We came across this principle, which we call wake sleep. In other words, you have two phases. One phase is when you expose the network to the inputs and outputs of the task you're trying to solve, and then you let it go, and then let it free run.

**TERRY:** Now why should the Boltzmann machine learning algorithm need to do that? Here's the reason. The trouble is if you're a neuron somewhere in the middle of the brain or a unit in some hidden layer and you're getting correlations coming in on the inputs, you have no way of separating out whether those correlations are due to the outside world, the task itself, or some other neuron, internal correlation. And so, the danger is that if you run it without any sort of way of canceling the internal correlations, then the brain, the network will basically spend more and more of its resources. The weights are going to be focused on internal correlations, and it'll ignore what's coming in from the outside.

**TERRY:** We call that fully ado. People start self-referencing. The beauty of having the sleep phase is that that will be purely generated by internal correlations, so you subtract that off. Now this is the principle. Like I say, it's a principle that came out of the Boltzmann machine, now Geoff has discovered how to use the same principle for feed forward networks.

**TERRY:** A very elegant, very beautiful, approach, which is, competitive with back prop and also much more biologically plausible. So that I think is going to be a real advance, both, in terms of the way that we think about learning, because the back prop is extremely heavy handed. You have to have information going backwards over connections.

**TERRY:** You have to be able to hold onto the state of the network on the forward pass. So those don't exist in the brain, but it is very efficient. It's very efficient, and it has proven to be scalable. So, it's universally used. However, if nature has found a way to do that without the heavy hand.

**TERRY:** And it does it with wake sleep and we do after all, spend a lot of time sleeping, then that's a big win. And also, it may help us understand how the brain works. So, this is another step along the way toward a better understanding, new versions of old algorithms that we can build on, and again, it's the first step.

**TERRY:** He'll be the first to admit that. He's focused on small problems, relatively small problems in order to be able to make progress. But there'll be many, many ways of building on that.

**CRAIG:** That's fascinating. Could you use this forward, forward network in a model the size of GPT-3 or

**TERRY:** I suspect so. He hasn't demonstrated that yet, but I don't see any problem with applying it to that kind of architecture. In fact, it is a feed forward model for inference. And so, he's showing now you can do the same thing with learning a feed forward learning model too. So, you never know. I mean, when something is new you have to explore it, it could lead you in many different directions. But you know, it smells right from my background and training, I can see how it could be implemented in the brain. It doesn't violate any principles.

**TERRY:** And also, many people have tried come up with versions of back prop in the brain. There are dozens of papers. And none of them as successful as Geoff's version. That's a good sign. It's a good sign because it means that this may be the roadblock, right?

**TERRY:** There was a roadblock, and by the way, the last time this happened was back in 1983 when the perceptron learning algorithm that was introduced 1959 by Frank Rosenblatt and it was, it's a beautiful, elegant, simple learning algorithm for a network that just has one layer of weights between the input and the output.

**TERRY:** And Minsky and Papert studied it mathematically and came to the conclusion that it was very limited in terms of what it could represent, namely linearly separable functions, which is to say ones for which if you have a space, the dimensions of the units, you could find a plane running through it that separates the positive and negative examples, right?

**TERRY:** Those are the weight space. And Minsky and Papert pointed out it's a real strong limitation, right?

**TERRY:** And they thought that there wouldn't be, in their view, any learning algorithm that could generalize this to a multilayer perception with hiddenness. And that's fair enough. I mean, that was their opinion. But others who read the book or didn't read the book got the impression that, oh, it's been proven.

**TERRY:** That this is a dead end. And so, I remember, when we were starting out, we were told continually, don't you know that it's impossible, so why are you even bothering? Right? This is like 20 years where this really elegant beginning was not pursued, and that's where Geoff and I came up with the Boltzmann machine and David Rumelhart and Geoff came up with back prop and many, many others. I mean, it really exploded. That's what was so exciting about the eighties, was that we were making these discoveries and so that's an example of where there's a bottleneck in thinking where, it just stopped a whole generation literally from making any progress because people had believed something that was wrong.

**TERRY:** And now there's this bottleneck having to do with Brain Can't Do Back prop right? So, it's a dead end. and now what Geoff has done is yet again shown that, well, there's a way of getting around that. And just by switching the architecture around a little bit and incorporating some old ideas from the Boltzmann machine, we can make progress.

**TERRY:** And so that's the way history progresses in these paradigm shifts, right? You go from one to another, you don't solve all the problems, but you, you're on a different level. And what you do is incremental improvements.

**TERRY:** You just get better and better and better and better. And of that's what's happened over the last roughly 20 years, 30 years is incremental improvements. More computing power, more data. But those are just incremental, right? We could see how that was going to progress, but when you get with these jumps, the transformer was another jump, then it's a whole new ballgame.

**TERRY:** You have a whole new set of tasks, applications, things that you didn't expect, and I think that that'll continue for a long time. I mean, that's the way it's always all through science.

**CRAIG:** Whether or not it's this new forward, forward algorithm, or it's scaling the transformer. It's mind blowing how much has happened in five short years.

**TERRY:** Or even less, actually the last three.

**TERRY:** Yeah.

**CRAIG:** Yeah. Since GPT-3.

**TERRY:** I've written an arxiv paper on this, and the title is Large Language Models and The Reverse Turing Test.

**TERRY:** So, here's the insight, I was reading an Economist article on large language models, but at the end, the author said, well, I have to give credit to GPT-3 that helped me write this article. And it came along with two interviews, one by Blaise Agüera y Arcas, and the other one by Douglas Hofstadter and they were diametrically opposed. One of them came to conclusion that it has very sophisticated understanding of theory of mind and social interactions and then, Douglas Hofstetter said it was totally clueless.

**TERRY:** It had no idea what's going on. And I'm not going to go into the details except to say that when you get something that is so different from two experts, right, it means there's something interesting here. And so, I started, playing with it myself and I came up with the following conclusion that it's pretty clear that it's been trained on a huge amount of data from many, many different people.

**TERRY:** Things that range from short stories, technical reports, poems, computer programs. I mean, it is just endless, right? It's just huge, huge, huge, huge, from the internet. And as a consequence, it doesn't have a persona. It's not like us. You have a certain personality. It may change a little bit over time, but it's different from mine, right?

**TERRY:** And we all have our own personalities, our own, personas. But what it can do is very interesting. It can adapt any persona, depending on how it's primed. You can ask it to be a persona, you can ask it, or it'll just pick up your persona. So, what was happening with the interviews is very interesting, is that if the person who was talking to us very highly intelligent knew a lot about social interactions and theory of mind.

**TERRY:** And was being given a story about that, it immediately picks up on that and it raises its level persona to match that. But if someone's interviewing it with nonsense things that don't make any sense, well, what's it going to do? You know, you ask a nonsense question, you get a nonsense answer. That summarizes Douglas Hofstadter interview, right?

**TERRY:** So, this is called the Reverse Turing Test. That is to say instead of the human trying to detect whether an AI is human or not. The AI is testing the human for how intelligent the human is.

Uh, excuse me, Terry, just a minute.

 Sorry about that.

**CRAIG:** Where do you think this scaling, whether or not it's with a more efficient algorithm, is likely to go and you reference the Wright brothers and birds a few times, but the Wright brothers didn't build a bird.

**CRAIG:** They built an airplane that did something like a bird. But it's very useful. I mean AI doesn't have to be an artificial brain, but it's something that is extremely useful.

**TERRY:** In fact, that's the point I'm making in my arxiv paper, which is that there's this huge debate that's raging.

**TERRY:** Does GPT-3 understand what it's saying? Is it intelligent? Some people think it's just a parrot, a stochastic parrot, which is very easy to dismiss because there's no way that it's just replicating something that is out there because you can ask it questions that have nothing to do with what's out there and it can generalize.

**TERRY:** That's the essence of a neural network, is it generalizes from a relatively small number of examples to data that it's never seen before, but from the same probability distribution. The networks have to be able to absorb a tremendous amount of data information and they got to extract. These hidden units have to be able to distill what are called latent variables. That is to say things that are like concepts. For example, if it was, being trained on faces, the concept that you have two eyes and a nose, right? Something that is of common to all faces. It has to represent that and the relationship between those features, right?

**TERRY:** And that is a way of compressing all the data and extracting what's regular and statistically significant.

**TERRY:** And so, language has only been around a few hundred thousand years. You know, it's something that evolved and there's not a whole new brain area called language. Right. It's embedded in existing sensory motor structures in the cortex and in the basal ganglia that generates motor sequences.

**TERRY:** So maybe what happened over evolution was that humans, through social interactions and being driven by the value of being able to communicate, became large language models embedded in a primate brain that was around and had all kinds of fancy stuff in it for sensory motor interactions. And here we are.

**TERRY:** Right. So, maybe it's telling us something about ourselves.

**CRAIG:** Yeah. I've had kind of a similar conversation a number of times in the last year. Everything I say to you and everything you say to me is kind of an accrual of everything that you've read and everything that you've experienced and seen, and your brain knits together

**TERRY:** some high-level understanding, which then is used to generate new sentences. Right.

**CRAIG:** And the semantics of the word reasoning aside, maybe we're just language models that are, whether you want to call it mimicking or what that's processing and integrating.

**TERRY:** No, this is not to demean humans, right?

**CRAIG:** Sure.

**TERRY:** These are tiny, by the way, compared to the human cortex.

**TERRY:** So, as I say in my archive article, one thing we know for sure that they're not human, right? Whatever they are, they're something different from human. But, what's really interesting is what I think the impact is going to be on philosophy. It's already clear that philosophers are very, very interested.

**TERRY:** In fact, at the end of my arxiv paper, I present an interview with philosophers and not too surprising, gPT-3 sounds a lot like a philosopher. But if you ask, what's the definition of these words, right? Consciousness. Well, you go to a dictionary, and it has a bunch of other words, and you look those up.

**TERRY:** It's all circular. There is no real definition. And the same is true of intelligence, understanding all those words are ones that we use in a colloquial way, and we kind of understand what you mean by consciousness. Cause we experience something. But if you want to study it scientifically, you're in trouble because it's just hard to figure out some more fundamental level, like in physics, right?

**TERRY:** We have theories that explain the properties of matter. And long before physics or chemistry, we had words for it, right? We had; Phlogiston was a word that was supposed to represent fire. Some substance that comes out of the wood, right, is called Phlogiston. Well, now we understand physics, we understand combustion, we understand the oxygen, so we have a theory.

**TERRY:** So, what I think is going to happen over the next, who knows how long, decades, is there'll be from understanding transformers and these large language models, we may actually come up with a real theory

**CRAIG:** of consciousness

**TERRY:** for all of these words to, extent that you can find them in these networks. Well, one thing for sure, we're going to figure out linguistics, right?

**TERRY:** Because one thing you can't deny is the fact, they're talking to us. not even Gary Marcus denies that. Right?

**CRAIG:** And that would argue then that at some point with enough scale and maybe tweaking algorithms that you actually will be able to build a system or a network of systems that approximates consciousness.

**TERRY:** So, listen very carefully to what, Dave Chalmers told us. What he told us is that it's not a binary decision if you're conscious or not. It's a gradient, it's a spectrum, and there are different kinds. Animals. Dogs have dog consciousness. Whales have whale consciousness, and he talked about fish.

**TERRY:** I'm not sure about fish, but they have their own version of consciousness and they're all different, and they're all adapted to whatever their environment is and their social structures. And so why not these large language models that they have a different kind of consciousness from all the rest of us.

**TERRY:** And all the other species, it's just a different one. But the real difference is that we can go in and examine it. We have access to every single unit, every single weight, every single activity pattern for every input, right? And it's just, it's open. It's a function. It's a mathematical function.

**TERRY:** This is not a mystery here. Right. Mathematicians know about functions. They can help us. They should be able to help us understand what mathematical theorems or approaches, tools are needed to extract from that complicated network that we created, how it works, and that will give us some clues.

**CRAIG:** So, these models will teach us about What consciousness or intelligence is, but they also will attain some form of consciousness or intelligence. Just look at what's happened as you said in the last three years. If we can continue to have that kind of progress,

**TERRY:** You're right.

**TERRY:** I mean, that's the part that is so unpredictable is that if we've made that much progress in three years, where will it head? And I've given you one scenario is that we'll be able to dissect it and understand it mathematically. That's one scenario. Another scenario is instead of having to type in a keyboard, you'll be able to talk to your computer and it will understand you. Even things like automobiles be able to talk to your automobile it's a universal appliance that will be taken for granted a hundred years from now.

**TERRY:** People won't even remember a time when we didn't have them talking to us, it'll just be taken for granted.

That's it for this episode. I want to thank Terry for his time. If you want to read a transcript of this episode, you can find one on our website Eye on AI.

That's eye-on.ai. We love to hear from listeners, so I encourage you to email us your thoughts at craig@eye-on.ai. We have listeners from all over the world, and I'd love to hear from whomever is listening in Nepal, for example. I see several of you show up on the map of listeners.

 In the meantime, remember the singularity may not be near, but AI is about to change your world, so pay attention.