AI and You

Transcript

Guest: Chris Summerfield, part 2

Episode 117

First Aired: Monday, September 12, 2022

Hello, and welcome to episode 117! Today we are going to conclude the interview with Chris Summerfield, Professor of Cognitive Neuroscience at Wadham College, Oxford, and also a staff scientist at Deepmind, the UK firm that put deep learning on the map with achievements like AlphaGo and AlphaFold, and attracted an acquisition from Google. He runs Oxford University's Human Information Processing lab in the Department of Experimental Psychology. Last week we talked about parallels between human cognition and AI, plus how Chris started out in psychology and then integrated AI to become a combined neuroscientists and AI researcher. One of his studies actually studied human sleep and a hypothesis that its purpose is to replay events that have happened during the day and integrate them for future learning, and then do the same in AI. And he found that doing a replay-like activity in a Deep Q Network, which is the model that's most commonly used for playing video games, in this case Atari 2600 games, that the AI did better. Isn't that interesting?

So let's get back into the interview with Chris Summerfield.

We're talking about these large language models here and I'm sure I wasn't the only one that was surprised by what they're able to do in that. Essentially, they're autocomplete algorithms on steroids and I wouldn't have thought from that description that they would be able to do the things that I can see them doing now, where even not just completing things from a prompt of a few sentences like finishing a poem. One of them was used to finish *Kubla Khan*, for instance, did very well, but now there's models that open AI has way you can ask a question, and it will do it I've asked it to compose a limerick and it didn't scan but you know, that's getting picky and or answer this question or give me ten bullet points on this, and it does it. And that's hard for me to put in the box of "completion model." That's where my ignorance of how it works is really showing up. Has any of the work that's been done by those transformers any of the results, illuminated anything in the other direction about neuroscience about just how we might be doing those kinds of things? Maybe we have completion models inside our head that we don't know about?

I mean, the first thing to point out is that kind of the model is just a prediction model. But then kind of in a way, all machine learning tools are prediction models in one way or another, right? So, when you build a supervised model, what it does, if it's classifying objects, for example, doing object labelling, it makes a prediction about the correct label, and it gets some feedback and then it adjusts on that basis. And, that's really, I think kind of the primary basis of how biological systems learn too; biological systems can learn by other sort of non-error corrective learning. So, one example is Hebbian Learning, which is just a sort of learning statistical patterns of association between inputs. But like much of our learning is, is error corrective. So, your question is about whether the specific innovation, which has led to these large language models has reflected back anything on neuroscience? I think the answer is not yet. But I can tell you my wild speculations in this area, and I think it will maybe join together some of the strands of the conversation that we've had so far. So, one of the things that's really remarkable about what both the language and the image synthesis models that have been built by places like Google and OpenAI, as you mentioned, the really remarkable things about them is that they seem to be compositional in nature. So, by compositional, what I mean, is that they have the property which language intrinsically has, which is that they're able to essentially take disparate elements in a sentence and kind of join them together in some new way. So, an example of composition would be, you know about purple things, and you know about pineapples, but maybe you've never thought about a purple pineapple before. Well, now you have because, I took two things, which are completely separate and unrelated, and I put them together, and I made a new thing out of them. So, neuroscientists, for some years now, and cognitive scientists actually going back decades, have argued that this property of being able to think compositionally is really at the core of our intelligence and the reason for that is because composition allows infinite generativity right? So, just like in my purple pineapple example, I could have chosen any kind of adjective and any kind of noun, and I could have generated a new thing and I can also add, right, via recursion, I can add new adjectives and nouns and I can make infinitely complex novel constructions. And so cognitive scientists going back decades have suspected that these properties of composition and recursion are at the heart of our intelligence and just by the by they're also kind of the properties that Chomsky identified as being the sort of unique intellectual innovations which allows human allow humans to use language in the first place. So, the question is, how what is it about the large language models we have today, that allows them to engage in in composition and I think that the answer is as follows. So, the transformer is a little bit different to previous algorithms which have been developed for machine learning applications. So the transformer is one of a class of algorithm, which is capable of mapping a sequence of data onto another sequence of data. So, in the case of a language model, the sequence is like some initial prompt in text. And the output is the continuation of that text, what was said next, essentially. And what most sequence prediction models do is they essentially learn a function, which takes each step in the sequence, and then uses the kind of learn some weights, which will predict over time what the most likely next step in the sequence is going to be. A transformer does do something a bit like that, but it does it rather differently and it does it by, rather than taking each item in turn, and using each item in turn to predict what's going to happen next, it takes a segment of the entire input, and it buffers it. So, what that means is that it now has a representation of a large tranche of the input and it uses a very clever algorithmic approach in which it learns, essentially using this is something which is called self-attention and it's essentially a method which allows the network to make predictions about what should occur conditional on everything else that is in that that segment of information and it does that by explicitly representing both what is in that segment and where it is in that segment. So, it has in my example of like on Tuesday, the postman delivered a *blank*, right, it has a representation of that whole sentence and what it does is it learns, it basically learns functions which encode where each of those words are, and what each of those words is. So, the fact that it has separate representations of *what* and *where,* is something which I think makes that algorithm more like what the human brain does. And that's because as I mentioned, when I was talking about spatial cognition, one really important thing that the brain does, is it explicitly encodes where things are in the world. So, it explicitly encodes where we are, at any one time, we have cells in the brain region called the hippocampus, which encode where we are. But we also encode things like where an item is in a sequence of data, or how many items there are in a sequence of occurrences, and so on and so on. And I think that the idea that you build systems that explicitly encode what and where naturally lend themselves to this property of composition, because if you have item A in position one, and item B in position two, by explicitly encoding A and B and one and two, you can now imagine, item be in position one and Item A in position two and that's not actually possible to do unless you explicitly encode kind of what and where or item and item and position if you like. So, one potential answer to why these models is so successful, is because they have this algorithmic property and you see it really prominently as well, in these texts to image synthesis. Many people may have seen text image synthesis models, like Dally, from open AI is a very famous one and kind of you can ask these models to generate compositional images, like kind of, please draw me a picture of Bart Simpson riding a skateboard down a mountain of marshmallows, right? So, that's a scene which is composed of multiple different elements and maybe as I was making that up, you were sort of simultaneously imagining all the different components and how they fit together. And in order to do that, you needed to know certain things, not just what's in the image, but where. That the skateboard goes on top of the mountain and not underneath it and that the marshmallows are in the mountain and they're not in Bart Simpson's head, right. And so, that property of composition is I think, really something which the transformer naturally lends itself to and just as an aside kind of very often, when we talked earlier about kind of symbolic computation and the fact that some people argue that there are properties, key missing ingredients that AI has not AI researchers have not yet come up with that. I very often refer to composition as an example of something that an AI system isn't going to be able to do very well and I think that kind of may well be the case that by innovating on the algorithmic side, we're able to produce systems which can actually solve problems, which, in ways which look a little bit allow behaviors, which look a little bit like those kind of elusive human like cognitive processes, which kind of people have argued that AI would never be able to produce.

Thank you for bringing up the examples of the image generators, because they, they've astounded me with their compositional powers, in that they will take these elements that they can put into an image that has never been seen before and yet, they will have the correct aspect ratio, relative sizes in three dimensions, occlusion, and so forth, which just astounded me that that could come out of an engine like that, after seeing after being used to seeing the image synthesis things like this a lot of things that Janelle Shane was generating, where you ask a network to generate the idealized, say, image of a barbell, and it would always have an arm attached to it because it never seen one that didn't and you know that it's got no idea of what a barbell is and a lot of these images will be Daliesque nightmares of things and somehow, we leapt from that to you can say "purple pineapple being eaten by a corgi riding a skateboard" and it looks flawless. I mean, a lot of the time it isn't, but the fact that it sometimes can generate flawless ones is astounding, is that illuminate anything for you?

Well, I mean, I think, it is instructive to look at the errors that these models do make very often, when they make errors, their errors, I mean, that there are some details so often not so good with faces, or very detailed, or maybe they are good with faces, but we're better at spotting the errors in faces, people are so attuned to what faces should look like and we notice weirdness right away. But for the most part, the sorts of errors that these models make are what you might describe as misbindings, right? So, kind of, if you ask for a pink triangle and a green star, then you might get a green triangle and a pink star by mistake, right? Incidentally, that's a very common mistake for *people* to make in even in recognition, if they their attention is diverted away from a stimulus, they will sometimes misbind the information in a stimulus. So, misattribute, what with where, right. So, if A goes with one and B goes with two, to mistakenly attribute A to two and B to one, to common error. But yeah, I mean, I think that, kind of the there is also just no overlooking the importance of scale, right? I mean, these models are trained on really hundreds of millions of images and, we don't know what the data set that was used to train DALL-E was, but it's very large. And there is there is no escaping the fact that a machine learning system ultimately is going to be largely bounded, its performance is going to be bounded by the data to which it is exposed, right and this goes back to the point that we made earlier about kind of what is in the data, right, so what do you need in the data for intelligence? and so, earlier, we talked about kind of the idea that, there may be things that are integral to intelligence, which will only be in data, which is generated by humans, right? So, my social or cultural sensitivities, insofar as I have them, may be gleaned from a lifetime of interacting with other people, right? That's my data, right? Can we give that data to an AI system? I think if we could give that data to an AI system, then it probably would show something, ultimately - maybe not today's systems - but it probably would show something that resembles human intelligence. But I think that it won't, because it's very difficult to give that sort of data to an AI system and when you look at these language models, what they're trained on is kind of a huge chunk of the internet? And the discourse that takes place on the internet is, of course is human generated, but it lacks many, many aspects of human-generated discourse which primarily amongst them that most human generated discourse, is personalized in ways which are disclosed not merely by temporal proximity, right. So, there's, when you speak, when two people speak to each other, the way that they speak to each other reflects the situation in which they find themselves, right. So, maybe, I'm a teacher, and I'm, or I'm explaining to a five-year-old, how to do X Y, or Z, right, I'm going to use very different language, to kind of, I don't know, if I'm in the pub, talking about politics, right? And the language that we use is grounded in the situations in which we use it and that information is totally absent from the internet. The only information that the large language models have about situation is temporal proximity; kind of the pub conversation is closer in time to other pub conversation, then to teaching five-year-old conversation. But I think that's probably insufficient, because language is such an instrumental process, we use language to influence others, and to learn from others, right and we use that we use language in an extremely kind of dynamic and responsive way and it's really just not possible to capture the full richness of that even by having a dataset of a trillion tokens from the internet. So, at present a language model, and this is why at present, these language models very often come up with things which are inappropriate, or things which are broadly inane, or things which are kind of really quite blind to the subtleties of the question that you asked, or the context in which you posed it, and I think in order to generate, to build models that can really go the extra mile, and kind of really resemble the way that we interact with each other, you're going to need higher quality data, you're going to need data that resembles more the sorts of interactions that we have with each other.

And I think that illustrates, again, one of the differences between AI and us in that in order for me to hold this conversation, I did not have to read all of Wikipedia first.

You did have to go to school for 20 years to be fair.

That's true. But there's a difference between what I learned there, and if I wrote it all down it wouldn't, take up as much space as Wikipedia. It might if there was some way of representing it completely, that I'm completely unaware of. But it doesn't seem to map into that amount of text. That's all I'm getting at. So, you mentioned earlier about the differences between human intelligence, human general intelligence and what we might want to do with AI, there's no reason that artificial general intelligence should come out the same as a human. Even that might be one of the goals now, because it gives us at least some purpose of comparison, like running the Turing test. So, but we already know what human intelligence is, like, we've got 9 billion examples of it. So, if we created an artificial general intelligence, if you speculate, what sort of qualities it will have that are nonhuman?

Well, I think kind of that's up to us really it depends upon the data that we expose it to, and the objectives that we give it in large part, you can give we haven't really been talking about the details of different sorts of approach within machine learning, but a lot of work today is based around a method known as reinforcement learning. So, reinforcement learning is a machine learning method in which an agent is trained to produce behaviors which will satisfy some sort of value function, which is a designation of some sort of score or value which is given to the agent as a function of the behaviors that it produces. And in reinforcement learning, you the experimenter, the researcher, get to define that reward function, you get to say whether it's good to eat apples and bad to eat lemons or good to eat lemons and bad to eat apples and so you get to define what the agent will learn to do and, I think there isn't really any one answer to this to this question the agents, agents are, they're going to do what they're trained to do and if you train them to do something useful and human, like they'll do something useful and human-like if you can find the right data. If you train them do something else, they'll do something else.

And obviously it was an unfair question, but I can't resist asking those kinds of questions when I've got someone like you here. And incidentally, while I do have you here, I want to run by you some, an idea that I expressed in my book that was born of pure wishful thinking and ignorance, but looking at brain computer interfaces, and recognizing that the ones that we have now that prototype, the prototypes that can give people visual inputs, blind people by stimulating an array of electrodes in their brain so that they end up being able to see this pixelized images that let them actually recognize a letter and realizing that the visual cortex couldn't be mapping our visual system, what we see, to a rectilinear array of neurons in the brain and yet somehow the brain was able to interpret those rectilinear stimulations in the way that we wanted. That to me suggested that that was neuroplasticity, the brain adapting to that to be able to see what we wanted it to see; and I then went on to speculate that maybe AI could meet neuroplasticity halfway to be able to transfer more complicated types of abstraction between brains. Does that sound like it's something that could be in our future? Or am I on the wrong track?

Well, first of all, you're absolutely right, that the success of many BCI applications, both in experiments and in the clinical setting is due to plasticity. So, the brain is remarkably plastic, especially in our early years and it's that plasticity, which allows information from different modalities to come to be processed in parts of the brain that wouldn't normally process it. So, the classic example is that, in Braille readers, for example, you often see that parts of the visual cortex come to code for somatosensation. Because essentially, that the what would be the job that would be performed by those neurons in vision or in reading is now being performed by the fingers that in the haptic sense. In a way, when we do BCI, particularly in the clinical setting, we're already using machine learning methods. So, kind of, you will have seen perhaps experiments in people who fought very unfortunate to have syndromes like ALS, also known as locked-in syndrome in which they're unable to move but largely cognitively intact and, in those settings, the algorithms which map brain activity on to a movement, often movement of a cursor on a screen, for example, perhaps to select a letter so that they can communicate with people outside. That's the algorithms that are used to do that are standard machine learning approaches. So neural networks, often they're slightly simpler than the very large neural networks, which are used, which we've just been discussing, but they're neural networks nonetheless. In theory we don't know the limits of what can be communicated or read out from the brain. But we know already that it's possible to decode abstractions in the form of words or lexical items; we've known that for about 15 years, we can do that from data recorded invasively, patients who are for various medical reasons, researchers are able to record directly from the surface of the scalp. That's something called electrocorticography. It's also possible with fMRI and we know that we can read out those abstractions. So, if we can read out those abstractions, then they can in theory be communicated to another individual. I don't know what the future holds for BCI. I know that, obviously, there are companies like Neuralink, which are very excited about the opportunities that might be afforded by BCI, I don't know what exactly they want to do with it. Maybe they want to make us all telepathic, perhaps, a bit like you had in mind. But we will see whether that's possible or not. But I think the main thing to bear in mind is that the real challenge with BCI is actually a really mundane one, which is that it's actually quite difficult and dangerous to have a chronic invasive implant in your brain and there needs to be an aperture through which signals can get out and that very often that provides a way for bacteria to get in, and bacteria in your brain is a very bad thing and, kind of in experimental research with experimental animals, it's one of the leading causes, one of the leading challenges is actually if you're doing invasive electrophysiology, to ensure that it's done in a way which is safe for the animal. So, I don't think it's something that we're going to be seeing immediately.

Good point. So, don't try this at home, folks. I would love to talk about this for hours and hours and days and days; we would run out of several things in the process. So, you have a book coming out in a few months, at the time we're recording this, and perhaps you can tell us when it will be out and what it's about and how people can find it when it does.

Sure. If you've been interested in the ideas in this podcast, then maybe you would be interested in my book because it touches on exactly the same topic. Book is going to be published in December by Oxford University Press and it's called *Natural General Intelligence*. So, as you probably guessed, the title is a counterpoint to the idea that maybe we might be able to build artificial general intelligence and the book really discusses ideas, themes from neuroscience, which might be relevant towards that endeavor. So, like I said, it'll be out later.

Terrific and how can people follow your work or keep up with you, or monitor a place where they can see when the book comes out?

Sure, so you can follow us on Twitter, so it's @SummerfieldLab on Twitter, or you can look at our webpage on the Oxford University website.

I'll put a reference to your website in the show notes and transcript. Chris Summerfield, it has been fascinating, wonderful discussion. I've enjoyed every moment of this and thank you for coming on the show.

Thank you very much, Peter. It's been a pleasure.

That's the end of the interview. A lot to get excited about, certainly what will be possible with brain-computer interfaces in the future.

In today's news ripped from the headlines about AI, Shenzen has brought in the most progressive of autonomous vehicle regulations in China so far, allowing registered vehicles to operate without a driver in the driving seat across a large part of the city, but there still has to be a driver in the vehicle.

Now, you may have heard me mention AutoX, or read about it in my book, which is a robo taxi service that has been operating in Shenzen for months without the need for any driver, but apparently those have been permitted in a limited basis by some local authorities, but what the whole city of Shenzen has now done is establish a framework for liability in the event of an accident. If the AV has a driver behind the wheel, the driver will be liable in an accident. If the car is completely driverless, the owner of the vehicle will be responsible. If a defect causes an accident, the car owner can seek compensation from the manufacturer. In a story from Reuters, Maxwell Zhou, CEO of robotaxi company DeepRoute, said, "If you want more cars, eventually there will be accidents, so these regulations are very important for mass deployment. This is not true driverless but it's a big milestone."

I am still waiting to see where the edges of the AV deployment will be. I'm convinced that we are many years from the general deployment of level 5 cars that anyone can buy and take anywhere, but it's clear that parts of the AV space are being nibbled away gradually and I'm still not sure what the shape they're going to carve out will be.

Next week's guest is Robbie Stamp, Chief Executive of Bioss International, and friend and business associate of the late Douglas Adams, joining us with perspectives ranging from AI definition and governance to the Hitchhiker's Guide to the Galaxy. That's next week on *AI and You*.

Until then, remember: no matter how much computers learn how to do, it’s how we come together as *humans* that matters.

http://aiandyou.net