AI and You

Transcript

Guest: Chris Summerfield

Episode 116

First Aired: Monday, September 5, 2022

Hello, and welcome to episode 116! Today I am talking with Chris Summerfield, who is 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. His research page says, "My work is concerned with understanding how humans learn and make decisions…. We are interested in how humans acquire new concepts or patterns in data, and how they use this information to make decisions in novel settings. We simulate learning processes using computational models, including deep neural networks, that are tasked with similar challenges."

So there was a *lot* I wanted to talk about with Chris, as you might imagine, and this interview was really about comparing how humans think with how AI thinks and exploring the similarities and differences. Here we go with the interview.

Chris Summerfield, welcome to the show.

Thank you.

So, can we start out by finding out a bit about your story how you got into neuroscience to begin with? And then where did that intersect with artificial intelligence?

Yeah, so, I studied psychology as an undergraduate and whilst I was a psychology undergraduate, I became like many people fascinated by how the brain works and I ended up being a neuroscientist. If you want a story, I could tell you a story, which is that I began as an English literature major and I was very dissatisfied with what I perceived to be kind of the absence of ground truth in the knowledge that I was acquiring. So, I decided to leave that degree, and I wanted to become an anthropologist. So, I went to the anthropology department, and they said, sorry, you can't change course, it's very, very complicated. You know, you have to jump through lots of hoops and so, I came out of that department, and next door was psychology and I thought, I'll try that. So, that's how I became a psychosis.

That interesting. At first blush, it doesn't seem to me like, I would thought that I would have thought that psychology had more ground truth than literature. How did you see it that way?

Well, so psychology purports at least to be an empirical discipline and so you can design experiments, which test predictions and idea being that you can generate a prediction which you can test kind of on some novel held out data, which you haven't you haven't seen before,

I see, there's some objective nature to it, where there's not really objective nature to critiquing Proust.

Well, I think, you know, kind of this is a really fascinating question, right about, you know, kind of what sorts what varieties of knowledge we get from disciplines which are more STEM-like or more humanities like, and actually, you know, it's very interesting that if you look at contemporary AI research, we also have different sorts of models that treat knowledge in different sorts of ways. So, one class of model, which people kind of often think of as a discriminative, or a supervised model, what it does is it does a bit like what we do in the humanities, sorry, in the sciences, right. So, what it does is you learn something from some data and allows you to make a prediction, and then you go away and look at some new data, and you test that prediction. So, that's a little bit like what we do in the scientific method. But we also have models, increasingly popular models, which are often called generative models, and those are models which what they do is they imbibe some large body of knowledge, and then they resynthesize it in new and interesting forms. So, you may have come across recent successes in building models, which generate human like text or human like art. So, those are generative models, they're generating new knowledge, but what they're not doing is they're not, they're not testing some theory, or they're not learning some discriminative function. Really, what they're doing is they're kind of imbibing a body of knowledge and they're using it to generate new and interesting forms, right? So, in the case of language, new sentences. Of course, it's trained in a predictive way, both models are trained in a predictive way. But I think it's sometimes it's useful to think about that distinction between discriminative and generative models as being related to the sorts of knowledge synthesis that we do in the sciences and the humanities.

Wow, all kinds of thoughts and for people listening, have heard of GPT-3, the G is the "generative" part and the rest is "pre-trained transformer "and let's talk about that. But can you tell us how AI entered the picture for you when you were getting into psychology?

Yeah, so I grew up in an era of, in which psychology became really dominated by the new techniques, which were based around functional brain imaging. And so in my PhD, I learned a lot about how to use methods to kind of scan the brain to understand how various processes are localized in the brain, how they localized in time in different neural signals. But when I got my first faculty position in Oxford - and even a little bit before - that I was already becoming increasingly dissatisfied with the idea that the answer to questions about the brain might be, that this bit of the brain does x and this bit of the brain does y, or even that the brain does x before y rather than y before x. And so when I set up my own lab, I undertook to really focus on trying to understand computational principles. So, trying to build mathematical models, which describe the transformations that information goes through, as it is routed from our sensory systems, to our decision and motor systems. And so of course, the fundamental premise of all contemporary machine learning systems is that they're learning a mapping function from inputs to outputs. So, they're doing exactly that. Now, contemporary deep learning systems, which tend to be large with lots of trainable parameters, they're doing that in a much more unconstrained and much less interpretable way than models that we build in psychology and neuroscience. But nevertheless, the principle is exactly the same and so as I was becoming interested in computational

neuroscience and becoming interested in computational neuroscience, and doing more and more of this in my own work, I one day received an email from a friend of mine, whom I had known for many years, because we both worked on memory research in as PhD students. And that friend's name was Demis Hassabis, and he went on to set up DeepMind and he asked me very early in the process if I wanted to be involved and so I was involved, first as a consultant, and then later as a permanent staff member and, that's how my interest has really kind of grown and blossomed over the last 10 years.

And so, let's look at now: you're in these two worlds, you're neuroscience at Oxford and you're at Deep Mind, who says on their About page, "our long-term aim is to solve intelligence developing more general and capable Problem-Solving Systems known as artificial general intelligence, AGI." That's putting a large stake in the ground they are when there's a lot of people saying them think that AGI will ever be solved and there's DeepMind saying that's their mission statement. Do you think that the current approaches in machine learning are on the right track to get there?

Well, I mean, this is the million-dollar question. If I knew the answer to that question, then kind of I don't know. We would the field would be in a very different place, I think if the answer were known to that question. So, at the moment, you can really see, answers to that question are sort of lying in two different camps, I think. So, on the one hand, you have people who say that there's a vital missing ingredient, there's something really significant, which we haven't kind of discovered yet, which is really going to transform our systems from being these kinds of relatively undifferentiated, and, really quite brittle prediction systems, to transform them into something which has much more the versatility that we see in the behavior of biological agents, and particularly humans. So, there's some magic ingredient which is missing. And people have different views about what that magic ingredient is. But very often, people appeal to something which is sometimes called "symbolic computation," or a notion that the brain kind of rather than dealing with continuous variables, deals with discrete quantities, like objects or propositions and that that's fundamentally missing from our current systems. On the other hand, you have the kind of inevitable progress, the march of progress, which is largely attributable to three factors. I think it's attributable to the increase availability of cheap computation. It's attributable to the increased scale of machine learning models, but most importantly, it's also attributable to algorithmic innovation. So, innovation, which is perhaps incrementally, but improving the efficacy of the algorithms that we kind of we join up to build large scale machine learning systems and so a really powerful example of that: you refer to the GPT-3 told everyone what the G means it stands for "generative," but the T stands for "transformer" and that's a computational innovation, which was made just five years ago and it's literally transformed machine learning since then, and allowed much more capable systems to evolve. So, in the other camp, people want to extrapolate from that progress and say, well, if that progress continues, then there's absolutely no reason why eventually, we shouldn't have a system, which is able to do the same things as you and I. Now, what is the answer to that question? I think it's very difficult to know. I'm not sure I know the answer to that question.

I how you've framed that there, calling out the two camps, I think, was perhaps exemplified by the Twitter exchange between Ilya Sutskever and Yann LeCun, when earlier said that, he thought that some machine learning models today were slightly conscious, and Yann said no, not in any way and that we need something more and you just eloquently explained those two camps and I've talked to people in both camps, and correct me if I'm wrong or misleading about any of this, the deep learning that neural networks started in an attempt to reproduce what we thought the foundation model of the human brain was. That's why they're called neural networks and they seem to be doing the same kind of thing in some ways, in that we can identify neurons in the human brain, or the visual system that can select diagonal lines, say, and we can find those in an artificial neural network that's doing image recognition as well. So, far, so good, maybe it's doing the same thing and then there's the fact that we need to train the network on 100,000, or a million images where the human brain needs nothing of the sort and that you can game the system by altering pixels, invisibly, to make it think that a horse is an ostrich and that it's not even recognizing the discrete object that we know to be the horse, but it might be looking at part of the background as well and that's fundamentally at odds with what we think about how humans do this. So, what are your thoughts around this kind of dichotomy, this discrepancy between the way AI, deep learning is solving the same kind of problems that humans do, but doing it apparently radically differently?

Yeah, I mean, I think, one way of thinking about this problem is that it's not kind of clear whether the fundamental determinants of intelligence are primarily located in the data or in the algorithm, right? So, on the one hand, what you might have is a very simple algorithm, but very complex data. And so you can see this in the vision algorithms that you refer to, right, so there's a class of algorithm known as a convolutional neural network, which has been proposed – first of all, it's a very successful tool for doing, for example, object recognition in the context of natural images - but secondly, it's also been proposed and, kind of as widely now, I think, accepted to be a good model of what the ventral stream of the primate that's the part of the posterior brain, which is primarily concerned with vision and with object recognition. And so, if you look at that type of model, the, what you can see is that as you say, as you train these models, with lots of natural scenes, just like we during our infancy, for example, are exposed to lots and lots of natural image data, if you like, through natural experience. If you train these networks, they form representations, which look a bit like those that we infer when we do neural recordings in experimental animals, or in humans using brain imaging methods and what I mean by that is that units within the network tend to respond in ways which reflect objects in the external world like faces or animals or tools or houses or whatever you want in exactly the same way as cells in the ventral stream do. So, very often this is proposed as evidence that kind of the convolutional neural networks are really kind of what they're doing is the same as what we see in biological systems. Something that people often don't focus on, is the fact that even if you don't train the convolutional neural network, you actually still get a really good correspondence between the representations that are formed in the network, and the representations that you see in biology. So, what that means is that the fact that units in the network respond to faces or houses or tools or animals or whatever, is not simply a property of the fact that they're trained to do object

recognition. It's a property of the way in which information in an image is transformed through the really complex, but essentially, random weights of the network. And so one way of thinking about that result is that actually, a lot of the complexity which is necessary for intelligence, like disentangling images into faces and people and so on, a lot of that complexity is actually there in the data itself. And so advocates of the kind of more what's often referred to as the empiricist view, that's the view that we can learn, really, we just need a simple algorithm, and we need to scale it big and that's something that in the end is going to give us general intelligence. Advocates of that view, can point to the fact that there is, in fact an awful lot of information and structure in that data. So, of course, people in the other camp are going to continue to argue that you're going to need specialized algorithms, and in particular, perhaps algorithms that involve, as-yet undiscovered computational motifs, to produce generalized intelligence. But there's a lot of stuff already there in the data and you referred also to quantities of data. And I think, this is something that's really hard to draw comparisons really hard to know; the claim is often made, that humans are really, really data efficient, we just need to, children often just need one or two exposures to learn a new word, for example, whereas, you know, neural networks trained from scratch will need lots of data to learn the same thing. That's absolutely true. But it's also kind of not quite a fair comparison because your brain, even the brain of a relatively young human child has already been exposed to lots of data, including probably hearing the words they're then being taught, maybe they've heard them passively being spoken by adults in conversation and even if not, then they are the product of millions of years of genetic heritage which has shaped the connections in their brain to be kind of exquisitely tuned to learn new information really effectively. And so it's very difficult to sort of draw that comparison and to one thing that I think is a real challenge. It's a real challenge for the field and it's not something that answers the question in one way or another is the fact that it's very hard to put your finger on exactly what is the behavior, which you would need a neural network or an AI system to produce in order to say definitively, "Ah okay, we've solved it, we've got that." We don't actually know, the reason that doesn't exist somewhere a list of like criteria, which you could put a checkbox, a tick mark next to and say, okay, well, if the system can do that, then it's solved, it's all done. It's just that doesn't exist and so that makes it really hard for us to ever know what are the criteria that we would need to satisfy for that to be true?

Do you think we could get those criteria? Would you like your research to construct those criteria?

Well, I think that one of the really deep challenges here is that whilst we have a sort of cod sense of what an AGI would look like. We don't really have any kind of generalized agreement, but that's not a research question, what an AI would look like. It's really kind of much more, it's a sort of general societal question, right, of what is it that we expect an AI system? To be? Or to do, right? And one of the reasons why this is so challenging is that of course, we are not just some sort of raw calculation machine in which you stuff inputs in and you get some smart answer out, right. Whether we construe an answer to be smart or not in the first place, is heavily tinged by contextual variables, social variables, cultural variables and we see this, of course, in the history of intelligence testing, which is something, it's one of the corners of experimental psychology, which is slightly more controversial and difficult to talk about. But it is true that for

hundreds of years, psychologists tried to measure intelligence and what they found was that the measures that they came up with, were always going to be sensitive to social, cultural and demographic factors in ways which clearly didn't say anything about the ability of different groups to solve those types of problems. Well, it didn't say anything about the intelligence of different groups, but what it did say about was the, the applicability of the tests, which the psychologists had come up with, to measure intelligence in a fair, and neutral way. And we have exactly the same problem with AI, which is that we don't really know what that test should consist of, so much of what we think of as intelligence is intrinsically linked to our social behavior, or to our cultural behavior. But an AI system, if we were ever to build a general intelligence system, it wouldn't really belong in our culture, I don't think we would want to think of it as another human or another member of our society with equal status to humans and so, in that sense, we wouldn't really expect it to have the same sort of intelligence measurable in the same sort of way.

It sounds like you're branching out into philosophy there and I'll just put in a plug here for people want to learn more about intelligence tests from the perspective of someone who has to design them; we had an earlier episode talking to Kristof Kovacs, who is the International supervisory psychologist of Mensa and that's his job. If we look at the architecture of computers, the Von Neumann architecture is, you've got a processor, you've got memory, and you've got things that transfer between the two and that's just not anything like the human brain of course, which is these like a sponge of neurons connected. And when we build artificial neural networks, we are simulating that with many layers of abstraction over the Von Neumann architecture and a human brain can famously operate on 20 watts of power, you can run it all day on a burrito and we haven't yet replicated that with 20 megawatt compute of information theory. So, entropy says that the 20 watts must be enough to do what we were doing. So, there's a lot of energy left on the table somewhere. You wrote a paper called neuroscience inspired artificial intelligence that argued that we're overlooking experimental and theoretical neuroscience in the design of artificial intelligence if I've got that right. Could you speak more to where you think we are that you are arguing in that paper that we should redirect our efforts?

That paper is five years old now. So, things have moved on a lot since then. But I think that some of the things - that was a paper that I co-authored with people from DeepMind - and I think some of the things that we said in that paper have stood the test of time. The paper highlighted a number of different areas that might be sort of fertile avenues for investigation in AI research and these areas are largely drawn from an understanding of the different functions from which the human or the mammalian but primarily human brain are made up, which correspond to things like having highly structured memory systems, while having a process of attentional selection whereby not all the information that is available to our senses is kind of given the same priority during information processing, the need to care about spatial cognition, and to think about how we're oriented in in the world, both what we neuroscientists referred to as allocentric, that means kind of where you are in a kind of map-like representation of the environment and egocentric t, that is with respect to your own body. And in particular, another area of focus, which we highlighted in that paper, is the need to learn powerful abstractions or concepts, right. So,

educated people educated humans understand, concepts like *mammal*, or concepts like *infinity*. These are all concepts, like the *economy*. These are things which don't have a single definite form. But they refer to groupings or abstractions over information or objects, which are nevertheless useful for communication and for reasoning. So, those are all areas in which I think there is still scope for progress in AI research. I think that in terms of what I just talked about, so just what I just referred to, so in terms of concepts, actually, it turns out that building large scale models, particularly large scale models that are powered by really great algorithms, like transformers, allows networks to form something that resembles the types of abstractions that we see in humans, at least it allows these networks to use these concepts to express them in natural language, to generate them in images and also in the process of doing so, to form representations inside the network, which pertain to really quite abstract things, almost exactly likie we think that humans are able to do. So, concept learning is really an area where actually, scale alone has got us quite far and so that's expressed, as you mentioned earlier, GPT-3, there are a number of large language models that are available today and these models, when you interact with them, really seem to be able to kind of use quite abstract concepts in a way which is comparable to how they might be used in human conversation. Now, on the other end of the scale, I just highlight something else, which is a memory system. So, human memory, in particular, is really highly structured, right? So, most people will know that we have a short-term memory and a long-term memory. But there's also actually information is stored and used over multiple different timescales in the brain and different regions seem to contribute differentially to those different stores. So we have some memory systems, which are really important for very short timescales, like the sort of timescale that elapses between initiating a movement and executing that movement, for example, with your arms, we have others that are useful over timescales over several seconds, such as the amount of time that it takes to process a sentence. And we have others that operate over really long timescales, like you might want to remember where you went on holiday last year. And thus far, I don't think we've really harnessed the power of really the complexity and structure of those different memory systems, to build AI systems that are able to smoothly integrate information from the distant past, the near past and the immediate past, just like we are able to bring all of those different timescales of information to bear upon our conversation and our actions and our decision making and so on. So I think that remains an area where there's great opportunity for innovation in terms of the architectures and the algorithms in AI research.

Do you think the human brain is organized along those different timescales in order to optimize the use of its wetware? There are people that have phenomenal either facial recognition skills or memory skills, they can tell you what they were having for breakfast, Tuesday, five years ago, and perhaps their timescale differentiation is different from the rest of us. If you had more hardware than would fit inside a skull, but AI is not constrained that way, would you still want to? Would this still be value in differentiating those timescales?

Yes, I think the word and that's because the structure of our memory systems, I think, is dictated largely by how the world is, right? So, this goes back to the point that I was making earlier that in the data that is provided to our senses is already like really highly structured, right? Faces are a thing because there is structure in the way that the world is organized such that when we

perceive a face, we always perceive it to have a particular structure with two eyes, and a nose and a mouth and so on. And the same is also true for the way that information covaries over time and so as you go about your daily activities, the information that you need to pursue your goals will change over different timescales, right? So if you're trying to bake a cake, for example, you will need to perform a sequence of activities, or sub-goals, but within each one of those goals, it may have different parts to it, right. So, kind of, maybe you need to beat some eggs, to add to the cake, but the beating the eggs involves getting them out of the fridge and cracking them on the bowl, and getting a fork and, and so on and so on right. And each of those activities is itself broken down into individual muscle movements and specific parts and so on and so, in order to effectively execute that plan, what the brain does, is it maintains information about different levels of the plan in different stores. So rather than you having to remember exactly the muscle movement that you need, in order to to mix the flour, whilst you're beating the eggs, what it does is it maintains one representation of the plan, which is very slow, which is like all the different parts, the four or five different parts of things that you need to do to make a cake. And then it has another representation, which is kind of within each one of those parts, which is on a much shorter timescale. The same is true, by the way in language, right and it's very interesting that the great breakthrough with GPT-3 was the discovery that with scale, what you can do is, you can essentially, because if your model is really very very big, then what it can do is it can effectively predict information predict what will follow in a body of text over a much broader window, so it can essentially, rather than if I gave you a sentence - previously, I mean, NLP has been around for a long time and previously, our NLP systems were able to do quite clever things like, if I said, on Tuesday, the postman delivered a blank, the system would be able to know that letter was a likely a valid completion to that sentence. But what the advanced with large language models is that, they're now able to do that not just over individual words, or phrases or sentences, but over whole paragraphs and they're able to predict what's a plausible continuation to a body of text many sentences into the future. And they do that by using a very big model. But I think that when you look at the human brain, what it does when dealing with language, it really deals with language has an intrinsic hierarchical structure, just like baking a cake. So, when you sit down to write an article, or whatever, right to write to write a piece of work, then you have an abstract idea and then you have different sections and each of those sections condition contains different ideas and those are composed of sentences, which are composed of words, and you really have that kind of breakdown and the explicit the explicit organization of information hierarchically by time is something which I think is really critical to the way that humans control or humans and other animals, behavior is controlled, and something which is not really properly exploited in our large language models to date, so there's no explicit partitioning of predictions at a more abstract level to predictions that are kind of the level of individual words, but rather, what you're just trying to do is trying to predict individual words really, really far out into the future.

That's the end of the first half of the interview; we split this one over two episodes so we keep the episodes of more of a uniform size and don't overwhelm your bandwidth or your brain.

I always like hearing how people were drawn into the AI realm from another one, because we keep hearing how many different fields people have come from into AI; in Chris' case, psychology. And each

time, they bring along the knowledge of their first field and cross-fertilize the AI field with that one and we see how much AI intersects with so many other areas of human knowledge.

In today's news ripped from the headlines about AI, Google has an called Gato, that has learned how to do over 600 tasks, 604 to be precise. If you listened to the Monty Python episode you heard me lampoon that, but of course it's really impressive. Because it's not 600 models in the same box, it's one model that's been taught to do 600 things, from image captioning to language completion.

Of course, many people are going to hear this and think that anything that can learn 600 tasks can learn a lot more, which is likely true, but then they may think that means artificial general intelligence is here, and it doesn't. Despite the fact that DeepMind refers to it as a general-purpose system. At the granularity, the specificity, of those tasks, humans are performing millions, probably billions of different tasks. There's no way to count. So while they consider it a step on the path to AGI, the remaining distance to cover is enormous.

Next week, we'll conclude the interview with Chris Summerfield, when we'll talk more about large language models, tie those to neuroscience, and talk about the new image generators like DALL-E-2, and then what might be possible with brain computer interfaces. 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