AI 4 Science Discovery Network+

AI4SD Interview with Dr Will McNeill
24/02/2021
Online Interview

Michelle Pauli
Michelle Pauli Ltd

11/08/2022

Humans-of-AI4SD:Interview-28

AI4SD Interview with Dr Will McNeill
Humans-of-AI4SD:Interview-28
11/08/2022
DOI: 10.5258/SOTON/AI3SD0242
Published by University of Southampton

Network: Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery
This Network+ is EPSRC Funded under Grant No: EP/S000356/1

Principal Investigator: *Professor Jeremy Frey*
Co-Investigator: *Professor Mahesan Niranjan*
Network+ Coordinator: *Dr Samantha Kanza*

Contents

1 Interview Details 1

2 Biography 1

3 Interview 2

1 Interview Details

Title	AI4SD Interview with Dr Will McNeill
Interviewer	MP: Michelle Pauli - MichellePauli Ltd
Interviewee	WM: Dr Will McNeill - University of Southampton
Interview Location	Online Interview
Dates	24/02/2021

2 Biography

Figure 1: Dr Will McNeill

Will McNeill: 'There are lots of connections between how we model the mind and how we think of computers'

Dr Will McNeill has been a lecturer in Philosophy at the University of Southampton since 2016 and is part of the Philosophy of Language, Philosophy of Mind and Epistemology Research Group. Prior to this he lectured at Kings College London, the University of York and Cardiff University. His research interests are centered on the epistemology of perception, social cognition and inferential knowledge.

In this Humans of AI4SD interview he discusses how clusters of cells become thinkers, the swamp man thought experiment, the excitement of artificial neural networks, and his excellent advice for early career researchers.

3 Interview

MP: What's been your path to where you are today?

WM: I'm currently a lecturer in philosophy, and I study our knowledge of others' minds, the nature of inference, and the limits of explanation. When I graduated from my Bachelor's degree, I said I was never going to do any philosophy again, but within a few months I realised that I was still thinking about philosophical problems. The one that has always bugged me is how we started off as a cluster of cells and became thinkers. It kept niggling me since I finished my undergraduate degree and, in one way or another, that's what I've been thinking about since I returned to philosophy.

MP: How have you approached such a large question?

WM: We can chart the biological processes which turn one cell into multiple cells, but the question is where the thoughts are there. One of the thought experiments which interested me after my undergraduate was swamp man. You're asked to imagine that a bolt of lightning comes out and forms a complex object that we'll call swamp man. Swamp man is physiologically identical to you, but they have just appeared out of nowhere. The question is: do they have the same thoughts that you have or not? The reason to say "yes" would be because they're physiologically identical: if your thoughts are simply a matter of how your cells are organised, then there can't be any difference between you and the swamp man. But it's not obvious that this would be the case.

Imagine that you make a machine that creates things that look like pound coins and are made of the same materials as pound coins. They would not count as pound coins because what individuates a pound coin is partly the fact that it's made by the Royal Mint, that it has a causal history determined by its environment and context. So it doesn't simply follow from the fact that something is a physiological copy of you that it would necessarily share the same thoughts.

As soon as there's that possibility that merely understanding how all of the cells are put together doesn't give you a full explanation, then philosophy needs to get involved. In these situations, you cannot get all of the answers to the questions you're asking from biology or developmental psychology, for instance.

MP: How did your interests develop after your undergraduate degree?

WM: I went on to take an MPhil, a two-year graduate degree, in London. It was tough, and quite a brutal awakening. There were 10 people on my course, and two failed — it was very scary! During that degree I looked at the philosophy of psychology, things like connectionism, computationalism and how we understand representations. Learning about those areas is probably one of the reasons why I'm interested in AI: there are lots of connections between how we model the mind and how we think of computers. That was the first time I learned about artificial neural networks, which excited me a lot. At that time, most people didn't think artificial neural networks were cool at all, but I did. I feel somewhat vindicated that they're now officially cool!

For my MPhil, I looked at Aristotle, the philosophy of psychology, and the philosophy of the mind. My gut instinct was always that society was somehow constitutive of your thoughts and affects your ability to mean things with your words. There's a great philosopher called Donald

Davidson, who was primarily a philosopher of language, but who also came up with the swamp man, among other thought experiments. My dissertation ended up looking at radical interpretation: how we should understand the possibility of there being thoughts. For example, if you're trying to work out what thoughts an object has, it's really difficult. They might make noises, but you can't be sure what those noises mean. The same noise might even mean different things, like "bank" and "bank".

Davidson's thought was that you basically have to assume that everything I say is true and linked to the world, then you have to work out where the links are. So if I make a certain sound only when it's raining, then you can deduce that I mean that it is raining. But there's all sorts of ways of thinking about the world. By making that sound, I might be thinking "It's raining," or I might be thinking "There's water falling from the sky," or even something like "God is crying again." There's an awful lot of thoughts that could correlate extensionally with there being rain. If there's no behavioural difference that allows you to discern what someone was thinking then there's no way to determine someone else's meaning. Davidson's answer is that you have to look at two people interacting with each other. With one person, there's a single line from their mouth into the world, but with two, you have two lines, and you can triangulate what they're thinking about by looking at where their lines cross, how they interact, and what they are coordinated around. It's a constitutively social view of meaning, because you can only really have meaning when you are interacting with other people or organisms.

During my PhD, I started to look at joint attention, which is essentially a real world version of triangulation. For example, when a child is growing up, they interact with each other, their parents and carers, and they start jointly attending to objects: looking at objects that others are showing them. In that way, they're triangulating with the other person. From there, I became interested in not simply attending to the same object as another person but whether or not you can attend to someone's attention: in other words, if you can attend to their mental state. The standard way to model how we know about each other's minds is very sophisticated. It takes a lot of cognition to work out, to infer, and to get to a stage where you are thinking in terms of mental states. To me, it looked like those must be the wrong models. If joint attention was important here, then it looked like it must be quite easy to find people's mental states in the world. My PhD ended up looking at just that: how we perceive mental states.

MP: How did that research lead you to AI?

WM: It's actually a really quick jump from there to AI. From a philosophical point of view, the question is: how do you get justified beliefs? It's a question about knowledge, an epistemological question. How do we know about each other's mental states? The mainstream view is that we infer them: we get knowledge of an inferential nature like, for example, scientific knowledge. We never see atoms, but we infer their existence from other things that we know. When it comes to mental states, the argument would be that all we see is behaviour, and we infer from that what we know.

But from an epistemological perspective, there are loads of serious problems with that model. It doesn't look like the inferences are any good, and if they're not good, then they cannot be justified. So, however much you believe that there are other minds around, it doesn't look like you can know it. In other words, we don't have an explanation for the sense in which you know there are other minds because the inferences look rubbish.

So here was the puzzle. The influential model of our knowledge of others' minds looks bad, but you can, of course, describe the processes that lead to our beliefs as inferential. You can describe any series of events in the universe as an inference: an inference is just really a move from one state to the next. From a cognitive point of view, you can think of anything that goes on in the mind as being inferential, but what's missing is any explanation of why the results of those inferences are good.

The interesting thing about artificial neural networks is that it looks exactly the same: you can see all of the inferences at work, you know what they are, what their results are, and what their patterns are. But what you can't find is anything that shows you why the results are any good, why they're justified. That's the interesting way in which the machinery of the world and the justification that we get seem to come apart, and we find it in perception in general in humans. But it looks like we've got a cool model for how true that is when we look at artificial neural networks. Think of them as being a model that shows you that the discovery of the mechanism is not always an explanation of the results' epistemic value.

Think, for example, about a belief that there's happiness taking place in the world. The question an epistemologist would ask is why that has a particular property, where that property is being justified so as to count as knowledge. Inferential mechanisms, if you can find them, tell you why the belief exists. What's weird, however, is that they cannot always tell you why that's a good belief. We find that among humans and artificial neural networks alike.

So the patterns we're looking for, mental states, are not related explicably to the information we get through our senses. We're unable to give a good inferential explanation of why our beliefs about mental states are good. You can chart all the mechanisms you like, and there's lots of interesting work on that, but, as far as I can see, the mechanisms don't tell us why we're getting the right beliefs. That's the real – and quite surprising! - issue.

MP: In asking artificial neural networks to be explicable, to be ethical, are we asking too much?

WM: As far as I can see, it's a sign of success that systems are inexplicable. All that means is that they're managing to pick out high-level features of the environment whose relationship to the low-level inputs is itself inexplicable. When you have a system that's doing that reliably, you should have something which is inexplicable. There's actually some lovely data on the inverse correlation between systems' explicability and their reliability: the more reliable they are the less explicable they are.

While the systems themselves aren't transparent, there's still plenty of room for transparency. You can be transparent about how you've trained them, what data you've trained them on, the tests you've subjected them to, how you treat the outputs and what contexts determine their credence. These are all ways in which we can be transparent, but the systems themselves will often prove intrinsically opaque.

MP: Do you think there is enough understanding of these issues?

WM: When I talk to artificial neural network designers, they seem to recognise these issues, and there are lots of sensible suggestions as to how to mitigate them. When I talk to neuroscientists, however, they don't seem to understand it at all, because they're still desperate to find explanations in natural neural networks that we clearly can't find in the artificial models we have. There's still this strange hope in neuroscience that when they find out more about neurons they will get their answers to these questions. They might in some domains, but in others, they definitely won't.

MP: What advice would you give to early career researchers?

WM: The first bit of advice I'd give is always ask a question. As a student, I was very nervous and quiet, and one of the faculty members said, "Look, this is a career. You've got to learn how to do this, and part of being a professional philosopher is that you should be asking a question at the end of every talk you go to." It's an excellent piece of advice. By asking questions, you not only gain confidence and experience, but you get interesting answers. Further, you get noticed a little more and respected by people. It's a very helpful confidence booster, which helps with networking.

That leads me to my second piece of advice which is just network, for goodness sake! In the humanities, we can tend to think that networking is beneath us because we have a "higher calling", which makes us special. It can be a very weird culture, and perhaps it's different in the sciences. Ultimately, though, it's total rubbish. You should meet other people who are interested in things, and talk about the things that interest you. It's easier than you think, and intrinsically good.