Tapescript

Can machines ever be like us? **(Interview with Michael Wooldridge and Nicola Davis)**

(I=Interviewer; MW=Michael Wooldridge; ND=Nicola Davis)

MW: The way that Chat GPT and all these, what are called large language models are trained is just by feeding them, giving their neural networks enormous amounts of text. And that text is obtained typically from the internet, I mean, you start by downloading the whole of the world wide web to obtain that text. And you spend months with very expensive AI supercomputers training the neural network with that text so that it can produce plausible and realistic text. If we really succeeded in kind of the Hollywood version of AI having a truly general-purpose AI, then it would be able to, for example, tidy my house and cook me a meal, it would be able to drive me to the pub and back safely, it would be able to do anything that a human being could do. We are, I think, still a long way from that. And people imagine that because we've seen a lot of progress in tools like Chat GPT, that robotic AI must be close behind. And actually, the reality is, it isn't, it's a long, long, long way behind. But nevertheless, I think we're now looking at some version of a competent general-purpose AI system within a reasonably short space of time.

I: Nicola, it's so interesting that it's the robotics that's lagging behind, while the AI is speeding ahead. And it really is becoming very convincing, you know, people are forming relationships, as they see them, with some of these tools. And even a Google engineer last year hit the news when he said that he thought Google's AI system had become sentient, which is quite a claim. I mean, this must be something that Michael has thought about a lot.

ND: Absolutely. And I think it's a really interesting area of research. I mean, not just in terms of technology, but also a philosophy: what does it mean to be sentient? What are the criteria that we need to meet? And I think there's, there's a difference between intelligence and consciousness. Knowing lots of things, or

seeming to know lots of things is sort of different from having the ability to, you know, shape your environment, or be original, all that sort of thing. And I think also, there's this sort of question I talked to Michael, about, to what extent does it involve interacting with your environment or having feedback from your environment. So that's something we really delved into.

MW: The problem of consciousness remains one of the great scientific mysteries, and we don't understand really at all, how human consciousness really works. There is, I think, some agreement that the concept of experience is a really important part. So, I've got a cup of coffee in front of me. And I experienced the aroma of that. And I can describe to you what the taste of my coffee is like. But even though we might use the same words to describe the coffee, we might describe it as bitter and whatever, I can't be sure that you are experiencing coffee in the same way that I am, because our experiences are inherently private. And the only way that we can relate our experiences is through communication, because we don't really have any evidence that you're experiencing anything, but I'll give you the benefit of the doubt. So, these systems, these large language models, Chat GPT and co., have never experienced anything. So, they will have read 1000s upon 1000s of descriptions of drinking coffee, and the taste of coffee and different brands of coffee, but they've never experienced coffee. And they've never experienced anything at all. That's fundamentally not how the technology works. All they've done is ingested some text. So, for those reasons, just to start with those reasons, there are other reasons. I don't think that we can view these things as, as being conscious. They're very plausible. They can describe experiences to us, but they never actually experienced anything. But there's another reason I would argue that these tools are inherently not conscious. You have a conversation with Chat GPT, and then you go on holiday for two weeks, and leave the conversation in mid flow. It's not wondering where you are, it's not worried about you. It's not thinking where's Michael gone wise? Questions. It's not thinking anything at all. It is a computer program, which is just paused in the middle of a loop. It's just literally not doing anything at all. And when you come back and you re-join the

conversation, it's not aware of the passage of time. And for all those reasons, I think that we couldn't credibly think of them as, as being conscious.