Two Conversations on Understanding, Sentience and AI

Towards the end of last year, a friend posted a video on our Facebook group comprising a dialogue between a human and an Artificial Intelligence (AI), in which the AI claimed to 'be conscious'. My first reaction was 'how can anyone be taken in by this?', and certainly my friend wasn't. But then I realised that if you regard mind and consciousness as synonymous, which is the position taken by the majority of people, it would be hard to demonstrate that a future AI could never potentially 'be conscious'. It is clear that eventually, if not already, AIs will be able to pass the Turing test. In other words, if you are chatting online, you would be unable to tell whether you are chatting with a human or an AI.

On the other hand, there is a deep intuitive feeling in most of us that a computer could never experience consciousness. So perhaps by looking more deeply into the amazing capabilities of the best of the current AI systems which communicate in an impressively human-like way, we could arrive at a clearer understanding of the difference between mind and consciousness, thereby dissolving one of the most subtle barriers to the recognition of our true nature?

So having experimented with ChatGPT, a very much better AI than the one in the aforementioned video – and probably the best of the current ones – I took advantage of a podcast with the cognitive scientist, Professor Donald Hoffman, to spend some time exploring this with him:

> *JB: I want to hear more about your views on artificial intelligence. … Obviously, there are ways in which AI computationally can outperform human minds, but are there any ways in which human minds will always be able to outperform AI?*
>
> DH: Well, that's a great question. I think it's also related to the question of 'are the AIs really conscious or not?'. For me, those are two separate questions. On the outperforming humans, I agree that what they're doing right now is astounding, and we haven't seen anything yet. They will be having access to terabytes after terabytes of data that we couldn't ever digest, and computational power that will eventually dwarf even the 86 billion neurons in our brains and so forth. So at some point, just like we have cars that can go faster than we could ever run, and we have just learned to accept that, we will have machines that will beat us at every intellectual endeavour. And so I have little doubt that there will be no intellectual endeavour in which we will be able to compete. In short order. I still remember, maybe 15 years ago, people thinking that the game of Go was going to be forever beyond the AI. But they trounce us hands down. So all the naysayers have been shown to be wrong repeatedly. So in terms of intelligence of that kind, we'll be outwitted. And most of us are already outwitted. There are very few holdouts in any small areas right now.
>
> But the other issue, though, about consciousness is a bit more subtle. Many of my friends and colleagues who are studying AIs from a physicalist framework, I would say 90 plus percent of them, assume that consciousness is just a product of complex physical systems and their interactions, for example, nervous systems. And so they would argue that an AI with the right kind of complexity will also generate consciousness. So not just intelligence, intelligent behaviour and intelligent speech and so forth, but real-life consciousness. Because you could imagine, for example, a dumb program that plays tic-tac-toe and beats me, but you might not say, well, just because it can beat Hoffman at tic-tac-toe doesn't mean that it's conscious. It could be more intelligent than Hoffman, but it's not conscious. But they want to make the stronger claim that it's not only, in practical terms, more intelligent than us, but in reality, it's conscious.

And there I think, it's a different story. There I would disagree with them. I think that physical systems like neurons and brains and chips and circuits and software don't even exist when they're not perceived. Those entities are simply icons in our headset, and they're not independent realities. … The reality is far more complicated, but all we've got is the gloss. And so we have to study the 86 billion neurons and all their trillions of synapses and so forth, not because that's the final reality, but because it's the best pointer we have to a much more complicated and deeper reality that projects down to those billions of neurons and trillions of synapses.

So I think that the idea that AI is conscious in the sense that most of my colleagues would claim, that unconscious circuits and software are giving rise to consciousness, I think that's false. Now, on the other hand, we do know that there is a correlation between what we call physical systems like brains and conscious experiences. And if you cooled my brain sufficiently, you would find evidence that I'd lost consciousness and so forth, or a stroke or death. So there is a correlation between our access to consciousness and icons like brains in our headset. And so there's the clear question of what is the relationship between consciousness and the icons in our headset? And could AIs be conscious in some deeper sense, not in the sense that circuits create consciousness, but could the circuits be icons in our headset that are pointing to a deeper consciousness?

So for example, just to make clear what I'm saying, right now I'm looking at you on a Zoom screen. So what I'm literally just seeing are pixels on my screen. And some of the pixels I see a green plant, and then I see a wall and I see a bit of a chair behind you. The chair pixels that I'm seeing are giving me no insight into any consciousness at all. They're inanimate as far as I'm concerned. But there are pixels in the middle of the screen corresponding to your face that give me some insight into consciousness, as to whether I'm making sense, whether you disagree or agree. So I'm getting some insight into your consciousness from those pixels, but not from the pixels of the plant. So does that mean that certain pixels are creating consciousness? No. Does it mean that certain pixels are conscious? That the Jenny face pixels are conscious so that the chair pixels are not? That's silly. Pixels are pixels and they're neither conscious nor unconscious. They're just pixels. However, certain pixels provide a portal into the genuine consciousness that is Jenny. And so it's not that the pixels are conscious, but they're providing a portal into consciousness. So they're an interface that has opened a portal into a realm of consciousness behind the interface. Whereas the pixels corresponding to the chair are not opening a portal, at least not a portal that's accessible to Hoffman. …

So there's a different kind of question here. Instead of asking, do AIs, properly programmed, create consciousness from their unconscious circuits and software, which is the standard view, change the question, could we develop technologies that perhaps look like AIs, that would be genuine new portals into consciousness? And I think the answer is yes. I think once we understand our space-time headset well enough, we should be able to reverse engineer it and rejig it to open up new portals. We do have one technology right now to create new portals into consciousness, and that technology is having kids. …

JB: Yes, it seems actually quite similar to something Francis Lucille says. He uses different language because what you call the headset, he calls mind, and he says mind doesn't understand anything. Mind is just a bio-computer. Only awareness understands, and understanding is outside of mind. (Awareness and consciousness being the same in his view.) And that seems to be something similar, I think, to what you're saying. There's a portal from mind into consciousness, but the understanding ultimately resides in consciousness.

> DH: Yes, given those translations between our language, absolutely, I would agree. I think he's using the notion of understanding in a very, very deep sense. There's some sense in which you could say that even ChatGPT understood you because it responded appropriately when you asked it questions. It responded in a way that you felt was appropriate. So there is some lesser sense, perhaps, in which you could say it understood you. And I think that Francis Lucille may be talking about a much deeper sense of understanding than that one.
>
> So as you can see, once again, I'm very, very careful about terms. One of the things we learn in science is to use all of our words very, very carefully and precisely, mathematically precisely, wherever possible, simply because many of the disagreements we have are not disagreements of substance. They're disagreements because we simply misunderstand the language that the other person is using. And so often in science, the standard is 'give me a mathematically precise statement of what you're saying, then there'll be a minimal risk of misunderstanding what you really mean'. And so that's why I always want to be careful about the terms we use. So to make these nuances between the different versions of understanding we might have, or what the different meanings of the word 'real' are. Because that's really one of the lessons we learn from science. The more precise you can be, the more problems you're going to avoid of the uninteresting type of problems. And the more you will see the more interesting and genuine problems that need to be solved. So ultimately, I would love to have the notion of understanding. For example, those two different notions I just mentioned, like the understanding that ChatGPT seems to have, is a difference not nearly as profound as the one I think that Francis Lucille is talking about.
>
> So what I would love as a scientist is to have some way of cashing out the word 'understanding' with mathematical precision that shows it's like a slider. So, here's the lower levels of the mathematical notion of understanding where ChatGPT counts as understanding. When we go to these deeper, more complicated notions of understanding, at some point, no it doesn't, and Francis Lucille is right, it couldn't. I would love to have a mathematically precise notion of that.
>
> Because you can see that what I *don't* want to do is just stop and say 'yes' without pointing out that there's something very, very rich here that we don't understand about understanding itself, about these different shades of it. So my attitude is never assume I know, always question even the very words that we're using and realize that with every question there are vistas of fun and exploration that open up. It is also an antidote to dogmatism. We use these words that are loosely or not well defined. We think we know precisely what we're saying and then we disagree with people who aren't using the words the same way, and so we get all these uninteresting disagreements and uninteresting and pointless dogmatism. My attitude is that it should be more like OK, question everything that I'm saying and all the words I'm using, and question my own understanding of them, and ask 'is there a more precise understanding?'. So it means never assume I know. Always assume I don't know and there's more to probe and that's I think the right spiritual attitude, frankly, as well as the right scientific attitude.
>
> [Living in Not-Knowing podcast, February 2023, Science and Non-duality]

That led me to think more deeply about what we really mean by 'understanding'. I tried to come up with a 'mathematically precise' definition, but failed. So during the Christmas online retreat, I decided to explore the topic with Francis Lucille:

> *JB: I was talking to a cognitive scientist about 'understanding' in the context of the capabilities of AI, and he said we need a precise definition of 'understanding', because AI seems to understand quite a lot. You can ask it a question, it seems to understand, and nowadays it gives you a really*

good answer. But there's a deeper kind of understanding and we need a clear differentiation. So he posed that as a problem. He didn't give any answer to it, but he said it was an avenue worth exploring. So that's what I've been doing, and I wanted to check with you whether my ideas so far are along the right lines and really how to go further.

So when I look at it experientially, when we talk about this deeper kind of understanding what we seem to be referring to is the kind of satisfaction and peace and the dissolution of the question. And that happens in the gap between thoughts, which is what you've been talking about, and in the gap there's just this pure awareness, it's infinite intelligence, it's infinite potential. And then an answer comes in some form of concept – maybe words or maybe images and that sort of flows into mind and then we can communicate that if we need to. And that applies both to spiritual questions but it also applies in my experience to mathematical problems or practical problems.

But then if I look at it theoretically, as you say, mind is just a bio computer and in that respect it's actually quite similar to an AI system. It actually has a much smaller knowledge-base than modern AI systems and it accesses that an awful lot more slowly. So if I ask myself, what's the difference in capability between a human body-mind and AI, then I struggle to find any way in which a human body-mind will always and inevitably be superior. It's superior in some ways at the moment, because AI is still just developing. But both human body-minds and AI systems are made out of consciousness – they don't have any real, separate existence outside of consciousness. And the only possible difference seems to be that the human body mind has some kind of portal into infinite consciousness which AI doesn't, because AI is just limited to its knowledge-base and its set of rules of inference, set of ethics, maybe its method of assessing probability –I don't really know much about how they work. So if there's a portal, what is the nature of that portal, and why does a body-mind have it but insentient objects don't? Is that connected with the gap between thoughts and perceptions? So in other words is there a realm of mind which sentient beings have access to, but insentient objects can't? And theoretically, could that be artificially created using a different kind of technology from what's used in AI at the moment? So that's as far as I've got, and I haven't answered the question that he posed.

FL: You open all kinds of doors, and as you open them, I forget the ones you have opened before, and I have the desire to address it, but then the new door opens! Give me one question.

JB: OK. So is there a portal to consciousness that a human body-mind has, which insentient objects like computers don't have?

FL: Well, you should define first what you mean by a human body-mind. Human body – OK, that's clear. But define human-mind experientially or conceptually. Because remember that my experiential definition of human mind is simply 'the set of all mentations we experience that comprise a phenomenal aspect of our human experience'. So if the question is 'does this set of perceptions have a portal into consciousness?', the answer is 'hell no – quite the opposite, all of these mentations appear in consciousness and they are made out of consciousness'.

JB: Does the same not apply to the computer? All of the things that it writes in reply to my questions to it are ultimately made out of consciousness.

FL: I don't know what the phenomenal perceptions of the computer are.

JB: Well we don't know if it has internal perceptions, but we know what it outputs. It's similar to when if I have an understanding of something and someone asks me a question, then my output might be in words.

FL: The problem with entertaining this question is that we mix first-person experience and third-person experience.

JB: Yes, well that's what I was trying to keep separate by …

FL: There are two different words – one is 'sentience' and another one is 'consciousness', consciousness being the first-person experience and sentience being a quality or a functionality we human beings attribute to some objects and deny or don't attribute to other objects. It's a purely arbitrary definition.

So let's assume we as human beings are facing a computer which passes a Turing test. Then we have to attribute sentience to this computer. Why? Because there is no way that we can establish a distinction between this computer and the human being, so that which goes for the goose goes for the gander: if we attribute sentience to the human being, we attribute sentience to the computer.

Now it is very different when we put into the equation the first-person experience, which we do when, as I said the other day, we can give a human being a set of instructions similar to the instruction in a computer program. This human being is going to execute the program line or command, line of code by line of code. That's a very simple example – when I was in college, we were just programming one line at a time, it was very simple and not very powerful. So this human being can do exactly what the computer does – process the data, give the correct outcome, without understanding what it was about, without having access to meaning.

So in this thought experiment, we have surreptitiously introduced the first-person aspect when I say that this human being who is endowed with consciousness and with understanding, processes this information without understanding the meaning of what he processes. And he acts as a computer which means that the computer can act without understanding. Which shows by the way that understanding belongs to a different realm which is what Roger Penrose called a non-computability. The computer is limited to the realm of computability. He connects this with Gödel's theorems which – if I speak loosely – are based on the fact that the rules don't include the understanding of the rules. The rules of any computational process don't include the understanding of the rules. They don't include their own understanding, just as the set of all sets cannot be a proper subset of itself. The paradoxes that we reach when we go to the infinite is a wave function for the universe. Who is going to collapse it?

JB: Yes. OK, so what you're saying is that your other definition of mind which is as a bio-computer, like all analogies, has its limitations, and when we're talking about this, then we come up against that limitation.

FL: Let me give you another argument against – it's not mine but it's my friend Edward Frenkel who mentioned it. He's a mathematician and is a professor at UC Berkeley and he often talks with people in artificial intelligence in the Peninsula. And when he talked with them about this, he notices that most or perhaps all of the underlying mathematics of these artificial intelligence processes are based basically on one very simple mathematical tool which is a lesser gradient to maximize or minimize to a gradient descent. So out of the full body of mathematics – number theory, geometric algebra, topos theory, category theory and all the mysteries that are out

there, this full body of mathematics created out of this first-person intuition, they use an extremely tiny part in artificial intelligence. So how could this tiny part possibly have the power to recreate and to invent the whole of mathematics? The tiny part was invented by this intelligence or discovered.

It traces back to something Jean Klein used to say often: more cannot come from less, better cannot come from worse, intelligence cannot come out of stupidity, consciousness cannot come out of unconsciousness etc. Nothing can come out of nothing, out of absolute nothingness. It is the same fundamental intuition we have that applies there: quality cannot come out of quantity – we know that when we come from Europe into America!

JB: It's beautiful. So going back to the original problem, I'm getting the sense that we can't actually produce a precise definition of understanding, for the same reason we can't produce a precise definition of consciousness or love or beauty or anything like that.

FL: What we can know, what we can say, however, is that there is no understanding without consciousness. So that which understands is consciousness. Because the reality that perceives, as I define consciousness, is also the reality that understands. So, if you will, understanding is the truth aspect of consciousness. Just as there is a beauty aspect and there is a love aspect, understanding is the truth aspect of consciousness. But it's a first-person experience. You can tap-dance around this, and your friend is asking for a third-person definition of something which has never been experienced by a third person.

JB: Yes, that's right. We're using the word 'understanding' and we're both understanding what we each mean by 'understanding' in this conversation, but we're not needing to define it because the understanding is beyond the conversation.

FL: Yes, because we understand what understanding is because we experience it. Just as we understand what consciousness is because in this moment, we are experiencing it, and in this moment as we are talking to each other we are understanding the words. Perhaps the meaning of a full sentence eludes us, but we nevertheless communicate. If I say the weather is beautiful here this morning, it's sunny, we understand.

Wittgenstein and others have tried to precisely square this circle, which is not possible. To come up with a third-person definition of meaning, or feelings, a third person theory of meaning. And in this case, because meaning is something which seems to be exchanged, which seems to be interpersonal, they come up with a kind of sociological explanation for it, or theory of it. But the meaning of meaning will escape them forever.

[Francis Lucille, 24/12/2022, Online Retreat]

So is there any significant difference between Don's view and Francis's view of AIs, mind, consciousness and understanding? I don't think so. The difference is in the nature of the arguments used, and the form of expression. Don was looking for a definition of 'understanding' in terms of a spectrum, but not necessarily a spectrum with fixed end-points. Just as the totality of conscious agents cannot be described mathematically, but the mathematics points to there being just one universal consciousness, (see page 5 of the paper Science and Belief), so it seems to me that Don's spectrum of understanding points to one understanding, which likewise cannot be defined mathematically, but which, as Francis says, we can define as the 'truth aspect of consciousness'.