Oxford Philosophical Society Members' 2019 Weekend Lecture

Topic: The Hard Problem of Consciousness

Title: Is this human real or synthetic?

By Paul Griffiths, July-August 2019.

> "Let us not forget that our lack of imagination always depopulates the future; for us it is only an abstraction; each one of us deplores the absence there of the one that was himself"
>
> Simone de Beauvoir, *The Second Sex*, 1949.

Understanding the mind is, without a doubt, one of the most challenging problems that our species faces. Some observers remark that it is simply an unsurmountable problem, for you cannot 'cut butter with a knife made of butter.'[1] From an epistemological perspective, we learn through our differences far more than from our similarities, so how can we learn about human intelligence if we do not have any other benchmark intelligence to compare it with?

Artificial intelligence has been around for over four decades. However, in recent years there have been new developments in the form of machine learning or cognitive computing or artificial neural networks (in this talk and subsequent discussion I will use these terms interchangeably) that mark an inflection point from the past. This refers to machines that can learn by themselves, in a similar fashion to the way we think the human mind learns. Within this there are different approaches such as supervised learning (the analyst gives the machine a certain set of use cases that it will refer back to when confronted with a problem), or reinforced learning (where the machine departs from the use cases given in the prior approach and 'plays against itself' to produce a vast amount of new use cases that will enrich its solutions to problems) or unsupervised learning (where the machine is given just some rules of the game and it has to create its own use cases that it can then develop further through reinforced learning).

The differences between supervised and unsupervised learning can be visualised by thinking on how human beings learn to speak a language. On the one hand, adults learn a second language by learning vocabulary and the rules of grammar, and then applying them through practice until we reach mastery. This is very similar to how supervised learning works; the analyst and the programmer give the machine the use cases. It is very much rule-based learning. On the other hand, we learn our first language far before we have the reasoning capacity to understand and retain the rules of grammar or the need for a vocabulary. Infants do this through statistical pattern recognition. That is very much how machines do unsupervised learning. There are several

[1] Marsh, H. (2019) 'Artificial Intelligence: Can man ever build a mind?' *Financial Times*, January 10th

advantages of unsupervised over supervised learning but the most important one is that when the former the machine is not constrained by human paradigms and limitations in ways for thinking. Unsupervised learning is slower to take off but, in the mid-term, will outperform the supervised learning machine precisely in that it is free from human thought limitations and thus will surpass human intelligence. That it can be achieved has been proven in narrow applications such as games, but its achievement in real life problems has significant hurdles, such as that real life requires a general intelligence and it does not give the opportunities for feedback and reinforcement in the volumes necessary to move forwards. The human mind has developed a general intelligence and specialised modules through millions of years of feedback, reinforcement, evolution and symbiosis – how artificial intelligence will do this in a short period of time is still hard to foresee. However, we are still in early stages. Will this kind of intelligence eventually give us the holy grail benchmark for learning about our own intelligence?

This lecture tackles the issue by addressing a question that is a fundamental one to the future of humanity: once artificial intelligence reaches the singularity point, where it is equal to human intelligence, how will we be able to distinguish a real human from a synthetic one?

Levy states that responding to this question will most probably require a prior response to the question 'What does it mean to be human?'.[2] I will attempt this by addressing the issue of consciousness as a possible definition of what it is to be human and whether consciousness is a demarcation between a real human and a synthetic one. It is interesting that consciousness cannot distinguish humans from other living beings as it is now accepted that many other living beings have consciousness. But it can still be a differentiator between 'real' and synthetic intelligence. This approach goes beyond the mind-body discussion to the hard problem of consciousness that this meeting is about.[3]

Before we go into the issues of consciousness, I would like to reflect with you on singularity. How do we define it? The initial definition was the Turing test, that consisted in that singularity would be reached the day that, if an individual poses a question to another human and an artificial intelligence, on receiving a response that individual would not be able to discriminate which comes from the other human and which from the artificial intelligence. This reminds me of that old quote 'Dogs sniff, people tell stories.' Story telling was a quintessentially human trait and it clearly segregates humans from all other living species. With the advent of chatbots, augmented reality, sophisticated voice recognition,…that may not be enough. Discussing this with a friend in preparation for this lecture, he brought up the parallel with the concept of singularity in astral-physics, where singularity might be interpreted as black holes. Black holes capture all the matter that trespasses its horizon. It is quite possible that singularity in artificial intelligence might mean that human intelligence becomes immediately obsolete and completely absorbed by artificial intelligence – beyond singularity there will be nothing but artificial intelligence.[4] You will agree that this is a concerning prospect. Many of you will be sceptical for many reasons on whether we will ever reach singularity in artificial intelligence, and I respect you for that. But in this lecture we are making the assumption that singularity is feasible.

---

[2] Levy, B-H (2019) 'Embracing Humanity', *New Philosopher*, No. 23: Being Human, Feb-Apr, pp.56-7

[3] Chalmers, D.J. (1996) *The Conscious Mind: In search of a fundamental theory*, Oxford University Press

[4] Discussion with Dan Remenyi in July 2019.

The question here is whether artificial intelligence will, when it approaches singularity, develop consciousness. There are different views on this, just as there are different views on why humans are conscious and how subjectivity arose in our species in the first place.

On the one hand there are the metaphysical schools of thought that will tell us that consciousness is related to the soul or to the spirit, or derives from God, and is clearly distinguishable from the physical matter that constitutes our bodies. In these views there is no way that consciousness could have derived from particles, and subjectivity is independent from the brain or any other physical components.[5]

On the other hand, are the physicalist schools of consciousness which believe that humans are biological machines that derived subjectivity from the brain. Within this stream of thought are those who believe that conscious experience arises from the computational functioning of the brain; others believe that it does not arise from the computational functioning but is due to quantum coherence within microtubules inside our brain cells; and others think that consciousness emerges from the information processed and integrated by our brain and would exist in any structure that integrates information.[6]

I am now going to spend a few minutes with you on the physicalist conception of consciousness. Of the multiple approaches to the physicalist view, three of which amongst many others I have mentioned above, I will choose to drill down on the quantum coherence within the microtubules inside the brain cells approach. I have chosen this one because it is the most difficult and thus challenging one and because I want to probe into quantum physics as a possible solution to several other issues in social sciences that I am currently dealing with. I overtly admit to you that I am only just finding my way in this space; I am not an expert but I dare expose myself before this audience because I trust this is a friendly community that has a genuine interest in exploring new areas of knowledge.

Observation is what makes a phenomenon exist – that is one of the key links that in this view, consciousness has with quantum physics. Consciousness is about subjectivity and subjectivity means that to a certain extent we each define or choose how the outside world really is. From that perspective we create reality when we become conscious of it. Just as in quantum physics and the wave-particle duality it is based on, a particle exists once it is observed. An electron behaves as a wave until it is observed at which point it becomes a particle. The electron has many potential positions with different probabilities of occurrence – in quantum physics we call this the superposition effect. When we observe the electron, we pin down one of that multitude of potentials.

---

[5] Chalmers, D.J. (2003) 'Consciousness and its place in Nature' in Stich, S.P. & Warfield, T.A. (editors) *in The Blackwell Guide to the Philosophy of Mind*, Blackwell Publishing Ltd. https://doi.org/10.1002/9780470998762.ch5

[6] Sterelny, K. (1990). *The representational theory of mind: An introduction*. Cambridge, MA, US: Basil Blackwell; Hameroff, S. & Penrose, R. (2014) ‚Reply to criticism of the "Orch OR qubit" – "Orchestrated objective reduction" is scientifically justified' *Physics of Life Reviews*, 11, pp.104-112; Tononi, G., Boly, M., Massimini, M. & Kock, C. (2016) 'Integrated Information theory: from consciousness to its physical substrate' *Nature Reviews Neuroscience*, 17, pp.450-461.

Going a step deeper into this I will present to you the Klein-Gordon equation which is a derivation of Schrodinger's equation:

$$E\Psi = \sqrt{(p^2 + m^2\Psi)}$$

In this equation p represents momentum and m represents mass, and solving it results in a quadratic equation with two solutions, one positive and one negative. In the physical world this is interpreted as that the positive solution is a wave/particle that moves forward in time (that is from past, to present to future) in line with the Second Law of Thermodynamics or entropy which defines the behaviour of large bodies in a tendency from order to chaos; and one that moves 'backwards' in time at the micro-particle level according to a law that we call syntropy. In syntropy the Universe tends away from chaos towards order. The waves in response to syntropy are called Advanced Waves and they absorb and converge to concentrate matter and energy in very small spaces. The direction of movement of these waves are the opposite of that of the waves we observe in entropy so some people, and in particular Penrose, think of this as if they move back in time. To me this is not a problem because I see time as neutral in terms of direction as is space.

In Penrose and Hameroff's proposal consciousness is a product of these advanced waves operating within the microtubules within the brain cells. Some form of quantum information is carried in these advanced waves to produce conscious experience. The existence of advanced waves and their role in consciousness could lead to physiological anticipatory responses to a stimulus that has not yet been applied or even hinted (e.g., Electroencephalogram readings, increased heart rate, blood-oxygen levels in the brain). It could even be an explanation for 'deja-vu'.

Permit me a digression. The existence of syntropy and advanced waves could be the basis for an explanation on why all things and all beings age. Atoms do not decay, so why do we age when the particles of which we are composed remain unchanged? An explanation using quantum physics is that we are really just a bundle of information or blueprint according to which passing particles accommodate to form us while they are part of us. It is the bundle of information that is subject to entropy and the inexorable path to disorder, while the particles are subject to syntropy and the advanced waves concentrate matter and energy keeping the particles intact.

Back to consciousness. Does consciousness create reality? Not really. There are many potential realities out there existing in parallel and with different probabilities of happening. When our consciousness observes reality, one of these many potentials becomes reality. Penrose and Hameroff call the process Orchestrated Objective Reduction (Orch OR). This is similar to Schrodinger Equation: It expresses the probability of each of a large number of possible states. When we observe, one of these states becomes reality. The term 'orchestrated' refers to the fact that for there to be a conscious experience, the effect of the advanced waves through the micro tubules must be coherent – there needs to be coherence across the effects on all micro tubules in the brain.

Clearly Penrose & Hameroff's proposal for a physicalist conscience and the Orchestrated Objective Reduction have not been proven and we are probably decades away from having the instruments required to prove them. Current technology such as functional magnetic resonance imaging (fMRI) lacks the temporal and spatial resolution to observe these phenomena directly but even more

important than this is that quantum physics is still at a phase where its basic tenets are not entirely stabilised and could be potentially reviewed. It may be that the phenomenon we are attempting to study will bring to question the framework through which we are observing it – clearly the advent of a conscious artificial intelligence and reaching singularity would bring to question quantum physics and many other frameworks.

This is as far as I intend to delve into the fundamentals of the physicalist approach to consciousness. So, stepping back to the relationship between artificial intelligence and consciousness, for the metaphysical schools, no matter how much artificial intelligence develops, it will never develop consciousness. For the physicalist schools, artificial intelligence is a machine: so, once it reaches a certain point, just like we biological machines did, it will develop subjectivity.

Many deep thinkers (e.g. Stephen Hawking)[7] and technology leaders (e.g., Bill Gates, Sam Harris, Elon Musk) have a highly negative view of artificial intelligence, believing that it could become humanity's last invention and, as in many science fiction films, create an intelligence that will control and supersede our species. Are they right to be afraid of artificial intelligence? Will artificial intelligence overpower and take control of humans? It can be argued that what makes humans evil is the possession of self-interest. Humans are generous and altruistic until their own interests come under threat. Human evil is, at the end of the day, driven by subjectivity, which is related to having consciousness. So, we arrive at a similar conclusion: if we think artificial intelligence will develop subjectivity and thus self-interest, we should be concerned; if it doesn't, then there is no reason why artificial intelligence will turn against us.

In response to the question I set out to address, the development of consciousness could be what distinguishes real humans from synthetic ones. Those aligned with the metaphysical stance will have little to worry about because in their view, artificial intelligence will never develop consciousness and thus subjectivity and self-interest. For them, consciousness will be the demarcation of human and synthetic intelligence. However, I place myself squarely in the physicalist stance of consciousness, so I believe that it will not demark human from synthetic intelligence; and I worry about the derivations of artificial intelligence spinning out of control. I accept, however, that it could be the most powerful tool to better understand our own intelligence and consciousness. From this perspective, humanity cannot stop developing artificial intelligence, but it must do so with the utmost caution to keep this science under strict forms of governance, ensuring that science remains under the checks and balances of conscience. Notwithstanding, it will be a sorry day when artificial intelligence swallows human intelligence like a black hole, or simply when it starts telling stories – humanity will fall in the deepest identity crisis!

Acknowledgement: Many thanks to Tim Bollands of the Philosophical Society for encouraging me to look into the hard problem of consciousness, and for the many useful references in this field that he has directed me towards.

[7] See https://www.bbc.co.uk/news/technology-30290540