On Minds and Machines

By Daniel Gibson

Until very recently, I viewed the mind as ethereal and intangible. I was a *dualist* – one who believes that the mind's inner workings are fundamentally separate from the body; indeed, separate from the physical world as we know it. Phenomena like self-awareness, emotion, intention and an apparent 'higher-order' intelligence seemed so far removed from the material world of matter, energy, causality and, above all, strict rules governing how everything behaves. My views, I would say, had two related motivations.

Mild anthropocentric narcissism formed the basis of a lot my opinions. Clearly, *homo sapiens* is the dominant species on Earth. This is due to its unrivalled intellectual ability: its capacity to identify patterns, imagine a plethora of possible futures, and envisage abstract concepts are all are evolutionarily advantageous for the hunter-gatherer. From this, and a superficial comparison with other animals and their intellectual abilities, came a sense, for me, that we are unique within biology. This implied that there was, at some point, a discontinuity in our evolutionary path – an exact point where we gained our 'consciousness'.

I applied the same reasoning to artificial intelligence. It seemed preposterous to me that a computer, fundamentally composed of nothing more than transistors and wires, could ever reach the level of sentience. The question of discontinuity arises again: how is it possible that there is some sort of threshold, above which consciousness is achieved? Is it possible that the addition of one more circuit or one more subroutine to a computer could tip it over this threshold? My answer was a stark 'no'. Hence the soul would be required to set humans apart.

I have since been swayed to a more *physicalist* perspective on mind. After clarifying our definition of 'consciousness', I will expose a weakness in the above argument for dualism, as well as presenting some evidence from research in artificial intelligence that contributed to changing my mind. Then I will attempt to explain one way a feeling of consciousness could arise within a purely physical world.

What do we mean by 'consciousness'?

The exact meaning of 'consciousness' is tricky to pin down. Should it refer to a level of intelligence high enough that self-awareness can emerge? Or possibly to the broader idea of an organism with the ability to step 'outside the box', overcoming natural evolutionary instincts and showing signs of empathy, morality, and emotion? Or can an explanation come from hard neuroscience, linking some area of the brain with these consciousness correlates? American philosopher Ned Block distinguishes between two types of consciousness: *phenomenal* and *access*. [1]

Phenomenal consciousness is the purely internal, subjective experience, related to physical sensations like the *taste* of an apple or the *pain* of a toothache; raw emotions like the *longing* for home; and attitudes and intentions, like the *need* to be outdoors. These experiences are called *qualia*. Every human being knows exactly how qualia feel – we can rely on this shared experience to avoid having to describe qualia to each other in great detail.

Access consciousness is just that: accessible. It encompasses any *observable* traits of a conscious mind; capabilities such as logical reasoning, abstraction and generalisation of ideas, self-improvement and displaying self-awareness. Some of these qualities can be tested experimentally, for example using the mirror test, which aims to detect self-awareness in animals.

The only way for someone to grasp how it *feels* to stand at the edge of a thousand-foot cliff is for them to have the experience themselves. But how can we verify that their experience is the same as ours? In the end, observation is all we can do: how do they react and what is their own description of the event? There is an inherent uncertainty in using such reasoning to ascertain whether two people are having the same internal experience. In truth, the only thing each person can know *for certain* is the reality of what is going on inside their own head. Our senses generate a *model* of the world outside, not an exact picture; that model can be wrong. When a conscious mind gains *knowledge* of something physical it has applied its own interpretation to its own perception of reality, making that knowledge a product of the mind rather than of the physical world.

Weaknesses of Dualism

I'd like to briefly counter some of my own past reasons for believing in a non-physical soul.

Regarding the claim of humankind's superiority over other species, I would ask the following question. If we are sentient and have souls, while a species from which we evolved does not, then when and how did things change? If we take for granted Charles Darwin's principle of natural selection, which is a decidedly gradual process, then it becomes difficult to argue that the soul exists in some form since its coming into existence would be a decidedly immediate and non-gradual event. Bear in mind, also, that it is widely accepted in neuroscience that animals have exhibited conscious behaviour. In 2012, a group of neuroscientists signed the Cambridge Declaration of Consciousness, which asserted that "humans are not unique in possessing the neurological substrates that generate consciousness." [2]

I will discuss developments in artificial intelligence using a series of case studies, and try to ascertain whether AI has exhibited any behaviour indicating consciousness.

Artificial Intelligence

During the 1940s and 50s, spurred by developments in computer science and logic, scientists started discussing the possibility of creating an artificial intelligence (AI). In 1955, computer scientist Allen Newell, cognitive psychologist and political scientist Herbert Simon and systems programmer Cliff Shaw began work on the Logic Theorist. This was a program designed to perform automated reasoning for proving mathematical theorems. The Logic Theorist was tested on the 52 theorems in Chapter 2 of *Principia Mathematica* – a pivotal work on the foundations of mathematics written by philosophers Alfred North Whitehead and Bertrand Russell. The program succeeded in proving 38 of the theorems given. The proof of one in particular was considered more elegant than the one given in *Principia* itself; when Russell was shown the proof, he "responded with delight". [3] The question arises: who takes credit for the proofs? The program was doing nothing but following instructions written by its creators, and yet they had no way of predicting what would be produced.

In 1997, computational chess engine Deep Blue, developed by IBM, defeated reigning (human) chess world champion Garry Kasparov in a six-game match. This was a major breakthrough for artificial intelligence. Another milestone came in 2016 when AlphaGo, an AI designed to play the Chinese game of Go, played a five-game match against Lee Sedol (who, at the time, ranked in the top four players worldwide). AlphaGo won four games to one.

It is worth exploring the methods employed by AlphaGo. For a game like Go with essentially limitless possibilities, a brute force method evaluating the results of all possible moves would take much too long. AlphaGo used a much cleverer solution inspired by the workings of animal brains – a *neural network*. Put simply, this is a program with many possible pathways, each representing a gameplay 'decision'. By analysing data on previous human games of Go, as well as playing against itself, the program ascertains which decision pathways lead to success and strengthens them so they

are more likely to be chosen in future. Thus, the system evolves somewhat organically, with strategies 'learnt' rather than being hard-coded.

An AI that self-improves in this way with minimal programming input can demonstrate emergent phenomena that are inexplicable to its programmers. Such a mystery arose when AlphaGo played Lee Sedol. Soon after the start of game two, at move 37, AlphaGo played what seemed to be a poor move – one that went against conventional Go wisdom. Soon enough, Sedol saw how innovative the move was. He said, "It's not a human move. I've never seen a human play this move. So beautiful." [4] Fifty moves later, AlphaGo had won the game. Analysis performed on that game has resulted in entirely new Go strategies. [5] Can such revelations be credited to the creators of AlphaGo? Neither they nor anyone else can explain the pattern of reasoning the neural network uses to arrive at its strategies – it would be inextricably tangled up in the vast array of decision pathways. I would argue the AI should take the credit for the discoveries it made. It seems to have shown independent creative thinking – is this not a sign of intelligence?

Interestingly, we humans have been consistently pessimistic about what AI can achieve next, especially for AI developed to play strategy games. Cognitive scientist Douglas Hofstadter speculated in 1979 that programs able to beat anyone at chess would only emerge as part of a *general intelligence*. [6] We're some way off a general AI (I'm choosing to be nonspecific to avoid being proven a hypocrite!), but Hofstadter was proven wrong by Deep Blue in 1997. Even after passing that milestone, it was considered a fundamentally different and harder problem for computers to beat humans at Go. [7] The common opinion echoed that expressed by the astrophysicist Piet Hut after Deep Blue's success: [8] "It may be a hundred years before a computer beats humans at Go – maybe even longer." [9]

Often, after an AI program succeeds in solving a problem, the solution is declared by critics as 'just a computation' and the intelligence of the program denied. [10] This is known as the 'AI effect' and is related to our history of poor predictions about AI. Douglas Hofstadter concisely exposes the paradox in this line of thinking: "AI is whatever hasn't been done yet." [11]

The tasks for AI presented thus far have all been fairly well-grounded in mathematics. Computers being fundamentally mathematical machines, it is perhaps not surprising that they performed so well. How does AI perform when asked to carry out a more *creative* task? For instance, could an AI ever compose a piece of music that can be appreciated on an emotional level? In the arts, AI has not yet significantly encroached on any human territory like it has in games like chess and Go. However, one project making progress in this area is the Artificial Intelligence Visual Artist (AIVA). AIVA is an electronic composer specialising in classical and symphonic music. AIVA has achieved some success – it is the only AI recognised by SACEM, a musical rights protection association – although it still needs humans to impose an overarching structure on its creations. [12]

It appears we are beginning to observe hints of human intelligence in non-human entities. In the case of animals, intellectual abilities don't seem too surprising if you believe in a continuous process of evolution. AI, however, is more remarkable. I think it is astounding that objects of *our own* creation, machines reducible to electrical currents flowing through transistors, could begin to exhibit human qualities. One could possibly even make the argument that we have observed evidence of *access consciousness* in AI.

The phenomenon of self-awareness has proved more elusive for AI. The problem arises when you consider the purpose of self-awareness: what is the evolutionary advantage of humans being self-aware? One answer is that, to be a part of a functioning society, you have to be cognizant of how other people perceive you. In fact, every member of that society has to be aware that every *other* member has some perception of them. (Some may have an awareness of their own self-awareness, an awareness that other people are self-aware, an awareness that other people are aware that they themselves are self-aware… such spirals of meta-awareness can propagate as far as you want.) When

an organism has an internal concept of itself *as a member of their society*, that allows it to adapt its behaviour to suit its surroundings and 'fit in'. Such self-improvement lies at the core of intelligence.

All of the AI programs we considered were designed for very specific purposes, leading to very narrow 'societal contexts'. Deep Blue, for instance, cared for nothing but chess. It could give you an in-depth lesson on strategy, as well as adapting its own playing style to suit changes to its environment (the opponent) – a kind of very narrow self-awareness. But it will never stop in the middle of a game and announce, "Chess is boring now. Let's play football instead." Such a change of character would require a drastic expansion of its worldview. It would need to learn the meanings of 'play' and 'football', as well as having some concept of itself as a being which could entertain subjective opinions. Implicit in such a statement is a deep wealth of experience and understanding. It is likely, for this reason, that the only AI that will be able to 'step outside the box' like this will be *general AI*: programs developed not to serve any particular purpose but for *all* purposes. Note that, even though we would consider this 'stepping outside the box' ability to be a decidedly *human* characteristic, such a general AI will not necessarily exhibit human behaviour. Our definition of intelligence is, naturally, centred on human ideals; a general artificial intelligence may present to us an entirely novel intelligence.

Progress in AI is slower in some areas than others. But we should not assume that research will, at some point, reach a 'ceiling'. I would propose that AI will continue to encroach on mankind's intellectual territory. If we continue to naively run afoul of the aforementioned 'AI effect', there will come a time when there is precious little left for us to call 'human'. The perceived human intellectual capacity will be a lot like a 'mind of the gaps' (like the 'God of the gaps' argument used to discredit theism). Eventually, all we have left will be our qualia. If AI reaches this point then it will not be enough for us to use qualia to argue for our sentience over theirs; because qualia are unobservable by their very nature, and observation forms the cornerstone of the scientific method.

I hope I have demonstrated that, to some extent, regarding humans as 'special' is not a point of view compatible with Darwinian evolution and, separately, seems to be slowly being disproven by developments in AI. Thus, the question remains: how can we explain our qualia if we aren't allowed the soul or any other such non-falsifiable explanation?

Before I set out the one of the theories of consciousness amongst physicalists, it is worth noting that this problem (the so-called 'hard problem of consciousness') is far from solved. A multitude of theories abound, but a decisive solution is some way off. Do not be surprised if, by the end, you feel as though you have missed something – as there certainly is a lot missing from our present understanding. The point I would make is that theorising like this is much more logically sound (not to mention healthier) than taking for granted an anthropocentric view. Allow your own feeling of 'special-ness' to be challenged.

Emergence and the Irreducible Brain

Emergentism is the philosophical position I will argue for. I think emergentism occupies a comfortable centre-ground within physicalism, in that it does not glorify the consciousness to an effectively dualist level, while also not reducing mental states to a cold set of physical states (reductionism).

Put simply, emergentism postulates that the *phenomenal consciousness* defined earlier is an *emergent property* of the brain. That is, the neurons and their connections that make up the brain have no properties of consciousness on their own, but when arranged into the form of a brain, consciousness emerges as a result. The whole is greater than the sum of its parts, and the properties of the whole cannot be explained in terms of properties of its parts but rather results from the ways in which those parts are arranged and consequently interact. A good analogy involves the *swarm intelligence* exhibited by an ant colony. The individual ants are not themselves intelligent, but by

following simple rules (often relating to laying down and following pheromone trails) they interact to form a colony which displays intelligent characteristics, such as finding the quickest routes to food sources and transferring dead ants to centralised burial mounds.

In his book *Gödel, Escher, Bach: an Eternal Golden Braid*, [6] Douglas Hofstadter attempts to explain how this emergence occurs. He views the brain as existing on multiple different levels. The lowest level involves the firing of single neurons and transmission of raw electrical impulses, often as a result of detection of external stimuli. The next level involves 'clusters' of neurons which take in related sensory inputs and organise them to make basic deductions about the world outside. The level above that involves the outputs from several clusters being organised to make a more complicated deduction. The hierarchy continues, until at some point – the highest level – the interpretations become so complex that meaning arises; some concepts embodied (called 'symbols') become isomorphic with objects in the real world; others are more abstract. It is the interactions between symbols that we perceive as thoughts. It is worth stressing that each level (including the symbol level) is fundamentally reliant on the firing of neurons, nothing more ethereal than that. At present, the exact nature and number of intermediate levels in cognition is not understood.

Consciousness is very much a high-level phenomenon. What we perceive as qualia are the results of the inner workings of all these lower levels (the 'unconscious mind'), but the consciousness has no way of accessing or manipulating the lower levels directly. A computer can neither fiddle around with the transistors that make it up nor perceive the electricity that forms the basis of any calculations it makes. In the same way, we cannot directly influence our own neurons – they follow the laws of physics just like everything else in the universe. Humans are not, however, rigid in this way. Humans are constantly changing the rules by which they think and act. New beliefs are formed, old ones discarded; people's personalities change. This variability and unpredictability is a high-level phenomenon, the result of neurons on a lower level following simple rules – the same rules they have always followed. It is entirely possible for a physical system following simple rules to exhibit complex, unpredictable emergent behaviour in this way; the ant colony is one example. Hofstadter gives an apt illustration of how consciousness is 'stuck' on the highest level: [13] say you are torn between ordering a hamburger or a hotdog. You are conscious of your own indecision and confusion, and might perceive that your brain has somehow ground to a halt. Your neurons, however, are not themselves balking, deciding whether or not to fire. They are running as normal; for instance, activating memories of past sensory experiences with fast foods, maybe also causing rational thoughts about healthiness, cost etc. A wealth of acquired knowledge is accessed in such a situation, most of it unbeknownst to the conscious mind. Once confusion is perceived, the conscious mind may then *feed back* to the lower levels, perhaps communicating feelings of embarrassment and the intention to make a quick decision. The lower levels then react, and a decision is made.

In fact, it is in exactly this way that Hofstadter suggests consciousness arises. He postulates that as soon as a symbol arises representing the self, which is complex enough to reflect on its own physical and mental nature (that is, a human level of self-awareness), that symbol can then perceive, monitor and therefore influence other neural processes. Higher levels inform the lower ones, lower levels feed back to the consciousness, back and forth and back again. Hofstadter calls this tangled hierarchy, in which all levels can interact, a 'strange loop'. [14] The perceived influence of our consciousness on the lower levels of the brain and hence on the actions of the body leads to the perception of intentionality and the appearance of free will. If what we are saying is true, then this control is an illusion. The illusion comes from the fact that our high-level symbols, which the self-symbol interprets as thoughts, are participating in making decisions. Since we cannot see the lower levels, it appears to us like such thoughts are spontaneous and original, but they are mediated by nothing but neurons following rules. The self-symbol has to be content with this incomplete view of the brain, as to have a complete view would require simulation of the entire brain *within* the self-symbol, taking up all the brain's 'CPU time' and leaving us in effectively the same situation. Hofstadter sums this up

nicely: “From this balance between self-knowledge and self-ignorance comes the feeling of free will.” [15]

Where Does This Leave Us?

I’ve argued against the dualist soul, and artificial intelligence is slowly impinging on humankind’s intellectual and creative territory. If you agree that this trend can be extrapolated into the future, a crippling blow is dealt to our feeling of free will. If computers could one day emulate humans, then by simple comparison we must conclude that our brains are, in some way, computational in their nature. How can we possibly make conscious choices if our brains can be reduced to nothing more than an electrical circuit? I believe this is the crux of the matter – and the reason why many people will reject this argument – as the widespread feeling that our species is special and unique hinges on our freedom of choice.

Douglas Hofstadter’s ideas about the irreducibility of the human brain could be used to claw back some pride in humanity. Even if our brains are fundamentally physical in nature, they are certainly complex systems, and many aspects of our behaviour are unpredictable. If it is fundamentally impossible to predict our future, short of allowing the universe to run its course, I would say this unpredictability is *tantamount* to having freedom of choice. That is by no means a rigorous argument, but it is how I think of my own free will, and I find it satisfactory to say “Even if I’m entirely deterministic in nature, and can follow just one future path, my brain is so complex that there is absolutely no way of predicting every aspect of that future path. Therefore, I *may as well* have freedom of choice.”

I don’t think that the physicalist viewpoint necessitates the depressing conclusion that, since everything can be reduced to computation, life loses all its meaning. My own beliefs regarding the mind have changed, but that has had no effect on my own *inner* feeling of free will or consciousness. Whether or not our minds our physical in nature, whether or not free will is an illusion – answering these questions will not change the fact that we *do* feel conscious, experience emotions and perceive our own thoughts. Above all, and without a doubt, it *feels* like we have freedom of choice. In a very real sense, how we feel is all that matters.

References

[1] N. Block, O. Flanagan and G. Guzeldere, The Nature of Consciousness: Philosophical Debates, MIT Press, pp. 375-415.

[2] M. Bekoff, “Animals are conscious and should be treated as such,” 19 September 2012. Online: https://www.newscientist.com/article/mg21528836-200-animals-are-conscious-and-should-be-treated-as-such/.

[3] P. McCorduck, Machines Who Think, Natick: A. K. Peters, Ltd., 2004, p. 167.

[4] M. du Sautoy, The Creativity Code, London: 4th Estate, 2019, p. 35.

[5] du Sautoy, p. 41.

[6] D. Hofstadter, Gödel, Escher, Bach: an Eternal Golden Braid, New York: Basic Books, 1979, p. 678.

[7] B. Bouzy and T. Cazenave, “Computer Go: An AI oriented survey,” *Artificial Intelligence*, vol. 132, no. 1, pp. 39-103, 2001.

[8] du Sautoy, p. 29.

[9] G. Johnson, “To Test a Powerful Computer, Play an Ancient Game,” *The New York Times*, 29 July 1997.

[10] J. Kahn, “It's Alive!,” *Wired*, 3 January 2002.

[11] Hofstadter, p. 601.

[12] B. Kaleagasi, “A New AI Can Write Music as Well as a Human Composer,” 9 March 2017. Online: https://futurism.com/a-new-ai-can-write-music-as-well-as-a-human-composer.

[13] Hofstadter, p. 577.

[14] Hofstadter, p. 709.

[15] Hofstadter, p. 713.