Consciousness is the great unsolved puzzle in Biology. How can we make Artificial Intelligence?

The following work reflects my personal opinion and most of the premises I gave to work towards the solution have, as in all Philosophy, many arguments for as well as against.

From where does consciousness come from?

But what is consciousness actually? Is it the fact that our lives feel like something? Then which were the first animals whose lives felt like something to them?

These questions, as many questions in Philosophy, remain unanswered. In the last centuries, however, Science has provided us with a framework to try giving meaningful answers to some of these great problems.

An example of this advance is the field of evolutionary biology: most of us will agree that consciousness surely did not suddenly irrupt into the universe fully formed; rather the elaborate forms of experience found in us derive from simpler forms in other organisms. We have learnt that the history of life is a history of intermediates, and much about the mind lends itself to a treatment in those terms.

For characteristics such as memory or perception this gradualist attitude makes a lot of sense, but then there is the feel to much that goes on in human lives (watching the sky, eating, being with others): when we take an evolutionist and gradualist perspective this takes us to strange places. How can the fact of life "feeling like something" slowly creep into existence? How can you be half-way to having it "feel like something" to be you? We refer to the most basic kind of subjective experience as sentience.

So, let's go back to the first question: where does sentience comes from? The answer to this question seems to change drastically the way we should proceed to create a real artificial intelligence. Today the two dominant views facing each other to the eyes of the public are Panpsychism and Functionalism.

Panpsychists believe that sentience is a universal and primordial feature of all things, something which pervades all of nature. In their view, even the smallest components of matter have some form of consciousness, they think that you cannot explain sentience in a reductionist way.

Functionalists instead think that mental states are constituted entirely by their functional roles rather than by their strict behavioural consequences. For example, a functionalist would define the mental state of experiencing pain as the functional connection between the inflammation of the nerve in your tooth and your moaning that your tooth hurts. Any system, with the proper functional relationships, can have mental states regardless of the physical material out of which the system is made. With this approach the electrical output of a neural network is similar in relevant ways (e.g. voltage, current, frequency, etc.) to the electrical output of the circuit in a computer, the computer will have the same mental state as the person with a neural network in his brain.

Which approach is more appealing to you?

Let's first consider Panpsychism. Physical science doesn't tell us what matter is, only what it does. The job of physics is to provide us with mathematical models that allow us to predict with great accuracy how matter will behave. In fact, the only thing we know about the intrinsic nature of matter is that some of it – the stuff in brains – involves experience. We now face a theoretical choice. We either suppose that the intrinsic nature of fundamental particles involves experience, or we suppose that it does not. On the former supposition, the nature of macroscopic things is continuous with the nature of microscopic things. The latter supposition leads us to complexity and discontinuity. A scientific imperative called Ockham's razor tell us to take the simplest and most unified theory which is consistent with the data. Everything seems to fit in, right?

Well, most neuroscientists and philosophers of mind tend to dismiss this idea for one main reason: panpsychism is unfalsifiable because there is no empirical test that could decisively confirm or refute panpsychism. We cannot explore consciousness and its relation to neurology through behavioural and anatomical studies in inert matter as we can do with animals and human beings.

If panpsychism were to be true then, because we cannot do any experiment on the theory, we would never know how to create a strong form of Artificial Intelligence, the furthest point we could arrive to is to push deep learning techniques to the edge of what is technically feasible and maybe obtain incredible results which could even change drastically our society, but we will never be able to say that our machine is thinking. Unfortunately for us, the story ends here.

What about functionalism?

With the eclipse of behaviourism and identity theory, functionalism has become the dominant perspective in philosophy of the mind for philosophers and neuroscientists, and there are many reasons for that.

We immediately perceive the main reason for functionalism's modern popularity, its obvious analogy to computer science. From the functionalist perspective, the mind is to the brain as software is to hardware. It is quite appealing because it provides a basis for modelling the mind and brain activity as a kind of computation and it allows the substantial theoretical framework of computer science to be applied to cognitive neuroscience and cognitive psychology. In computer functionalism for example, neural processes implement algorithms in the brain and this implementation is just what the mind is.

Even though the are some problems with it as a theory of mind (one of them is the Chinese Nation experiment: if all the people in China each performed the functional role of a neuron, could we say China has a brain and is therefore conscious?), functionalism mixes beautifully with evolutionary biology: in this view, sentience is brought into being somehow by the evolution of sensing and acting, when thinking about the animals we normally refer to as sentient, we conclude this involves being a complex system with a point of view to the world around it that can exhibit adaptative behaviour. Imitating the brain, from the simplest one to our, appears to be the way to go.

Pioneers in AI already guessed the right idea: in the 40s recent research in neurology had shown that the brain was an electrical network of neurons that fired in all-or-nothing pulses. The close relationship between ideas in cybernetics, information theory and theory of computation suggested that it might be possible to construct an electronic brain. Some scientists in 1943 analysed networks of idealized artificial neurons and showed how they might perform simple logical functions. They were the first to describe what later researchers would call a neural network (a system which, using algorithms, learns to perform tasks by considering examples from complex sets of data)

Neural network or even deep neural network are not Artificial Intelligence though, because even if we can say that they can learn things, the whole process is mechanistic, they cannot modify their behaviour accordingly to what they "learnt" or if they can it is only in a direction decided by researchers (AI learning how to communicate with humans for example: Google, Alexa).

This leads us to the realization that even complex systems such as these modern neural networks cannot display what we call General-Purpose intelligence. Something must be missing and when we analyse the simplest "complex" brains in nature, there are indeed a few things that are missing: one of them is just the complexity of the biological brains, even the simplest ones have billions of neurons, whereas some of the most complex Artificial Networks currently have "only" from tens to hundreds of thousands of neurons (AlphaZero, 2017) Imagine what could such machines do if they had this complexity.

But that's not all, think about how animals learn. Learning in the brain is achieved primarily by forming new synapses. This is a much more powerful form of learning than modifying existing connections as practiced in deep or machine learning. It explains how we learn new things quickly, without affecting previous learning.

While these are core features missing in modern AI, they cannot be the only ones. We need to map simpler brains and try first to imitate those of animals displaying complex adaptative behaviour such as arthropods or molluscs, before "evolving" to brains of vertebrates and finally human-like animals. Some projects in this direction exist (Drosophila Connectome, OpenWorm) but they are not popular enough so far. And yet, this is our best shot at creating a form of Artificial Consciousness, no matter what your answer to the first question was; which brings us to one last big question:

If we were to achieve this task, what would be the implications?

Should these machines be considered as life, hence deserving rights? Could we consider ourselves to be, after all, gods?