Podcast Episode 10: Machine consciousness

As of 1.6.2021

Teaser

Hey everyone! Welcome to the new episode of the Life Extension Podcast – technology & magic, society & business. Will machines become only intelligent, or also conscious? This is an important question, which will determine how humans are going to live together with technology, and if humans will be able to extend humanness into technology. Listen to this episode if you are interested in what scientists and philosophers have to say about consciousness and what that tells us about A.I. We will discuss if consciousness is an illusion, or the result from computational power or brain circuitry. This will lead us to questions of the nature of reality, social relevance of consciousness, as well as morality and motivational inner-drives of A.I.

This episode is about machine consciousness. One of the most fundamental ideas in transhumanist and posthuman thought is that humans will dramatically extend their life spans, first through advances in biomedicine, and afterwards by gradually moving life from the biological to a synthetic platform. This second step requires answering the fundamental question, if machines could become conscious. Otherwise, how could human selves ever emerge from artificial systems. There is a difference between intelligence and consciousness. While there is hardly any doubt left that an artificial general intelligence – meaning an A.I. with the same intelligence as a human - will be possible in time, it is still controversial if intelligent A.I. can also be conscious of itself. The famous Turing test stipulates that a system has human intelligence when it is possible to communicate with it through a closed door without being able to recognize if it is a computer or a human. However the Turing test is not able to test if that system has a conscious self, or is just a zombie.

Let's start to quickly review what scientists and philosophers think of consciousness in general.

The whole discussion about consciousness became a hot topic during the 1980s and 90s due to the development in computer technology. Researchers of A.I. like Malvin Minsky already wrote at that time that computers will certainly be able to develop conscious selves (Minsky 1988). He further claimed that robots will ultimately rule the world, and that we should consider them as part of the human evolution (Minsky 1994). Minsky is very much a precursor of contemporary transhumanist thinking.

For Daniel Dennett, a philosopher, the conscious self is an illusion, and we are all complex zombies. Mind is a function of the brain, which is wired in a highly complex way as a result of neuronal networks having evolved over millions of years. (Dennett 1993). Consciousness is just a user illusion, similar to the screen of a computer, which makes us believe that we are in charge, although all the computation is done behind the screen. Susan Blackmore, a psychologist and researcher of consciousness, claims that consciousness is an illusion created by memes to propagate their own replication. While according to Richard Dawkins's book "The Selfish Gene" (1976) genes are the beneficiaries and drivers of the first

evolutionary process of all organisms, memes as cultural units are beneficiaries and drivers of our large brains (Blackmore 2000, 2003).

This line of thought has been critiqued by philosopher Thomas Nagel. For him consciousness can be completely described by physical processes in the body. Consciousness is real, but only subjectively. In Nagel's thought experiment third party observers can imagine to be a bat by taking its point of view, but can never really know how it's like to be a bat due to different brain circuitry (Nagel 1974).

Bernard Baars, a cognitive psychologist, is the originator of the Global Workplace Theory (GWT). He sees consciousness happening in a sort of workplace, an analogy to the short-term working memory of a computer. Perceptions streaming through the workplace become conscious when they spill over globally to other brain areas and cognitive sub-systems (Baars 2001). The Global Workspace Theory is therefore mainly dependent on a specific architectural feature of the brain.

Neuroscientist Christof Koch and computer scientist Giulio Tononi, on the other hand, are promoting the Integrated Information Theory (IIT). It says that any system will experience consciousness under the conditions that is has sufficient cause-effect relationships built into its circuitry, and that it is sufficiently integrated and complex. This theory differs from the Global Workplace Theory as it claims that consciousness cannot be computed or simulated. Instead, it must be built into the structure of the system. According to Daniel Koch two systems with the same input/output operations could not be distinguished from each other by an outside observer, although one of them is conscious, and the other one an intelligent zombie. The difference is that the conscious system has causal powers built in, and the zombie is more or less wired to produce its output in predictable ways. The philosopher John Searle counts as a precursor to the Integrated Information Theory. For Searle the human mind is a biological phenomenon, not an abstract computation. His well-known Chinese room thought experiment expresses the idea that a computer operates only with meaningless symbols. Meaning is not intrinsic to a computer, but is being interpreted by a third-party observer, such as the programmer or user of the computer.

The difference between those two theories is important in the sense, that according to Global Workplace Theory, machines could experience human consciousness based on computational power, while according to Integrated Information Theory they cannot. According to the latter, human-level consciousness could only be achieved by building hardware in the exact image of the neuronal networks of the brain (Koch 2018, 2019). Therefore both theoretical approaches are being followed nowadays – not only machine consciousness as a result of increased computer power, but also the use of nano robots to measure and copy the entire brain including all the 100 billion or so of our neurons.

In today's discussion on consciousness voices of philosophers have largely been replaced by those of computer scientists. What philosophers had to say about c. has already been said in the 1990s and nothing fundamentally new has been added. Computer scientists are leading the discussion nowadays, simply because A.I. has made major breakthroughs, and development is ongoing at high speed. As scientists are the main players in this field with the ability to shape and participate in experiments, they appear of course as the most visible interpreters of machine consciousness.

Although this turn of discussions from the philosophical to the technological has apparently led to higher immediate relevance, it has not really resulted in an increase in depth. In fact, we still know

almost nothing about consciousness. The so-called hard problem of consciousness, which is about explaining the origin of consciousness, has in no way been solved by any philosopher, psychologist, neuroscientist, or computer scientist. Next to extremely vague hypotheses on the architecture of fully conscious machines there are mainly experiments measuring brain activities and connecting those to specific thoughts, experiences, or actions. These experiments are somehow adding to the illusion that consciousness can be read by an observer, although an externally measured brain activity is certainly not identical with subjectively experienced consciousness and the concept of a self. Furthermore, the same brain is involved in an endless variety of thoughts, emotions, experiences, and actions. Could there really be unique brain activity patterns for each of those, allowing the observer to read or compute them? The dilemma of both the fields of A.I. and consciousness studies seems to be that machines are without doubt becoming more intelligent by the day, but at the same time we still know basically nothing about their potential of becoming conscious.

What should be our take on this? Certainly we should not just rely on computer and A.I. scientists, mainly because they have their own very particular way of seeing the world. Let me suggest not to get confused by the philosophical and technological discussions and to focus instead on psychological and social relevance.

The question if consciousness is an illusion, is actually about the nature of reality. Does the physical world around us consists of forests, lakes, and people, or rather of atoms and molecules, or perhaps just of computer code? All of this is true, but we know that meaning is always constructed and negotiated. A newly born baby knows nothing about its environment, but it is gradually learning to distinguish individual features and to make them meaningful. Units of meaning may come as cognitive schemes, cultural patterns, or memes. This process is also relevant to A.I. We should better assume that A.I.s with the capability to self-program will be able to construct meaning (I mean this in comparison to computers operating on a central program). Pre-condition is of course that they are sufficiently complex. There is no reason to believe that consciousness is a functional property of biology alone. It would be much safer to assume that it is an emergent property of complex systems, independent of what they are made of. Will machines become intelligent enough to pass the Turing test? Probably yes in the not-so-distant future. When that happens, machines will have learned and constructed human meaning from all the ones and zeros of their computer language, simply because they communicate with humans. For that reason, it won't even matter if computers are only simulating meaning. Meaning is meaningless in itself. It gets constructed in relation to other meanings to enable the human or artificial subject to navigate and manipulate its respective life world. Meaning is also context-dependent. That is why chatbots can already be quite good at handling specific questions in service hotlines or choose music based on our listening history.

A remaining question I have about machine consciousness is motivation. Why should an intelligent and conscious system have any interest to do anything? Humans have written tons of literature about why we do things, and we are assisted by armies of scientists from a variety of disciplines, medical practitioners, as well as philosophers, to explain our inner drives. Humans are driven by all kind of things, which make us happy, greedy, fearful, or angry, and which motivate us into action. What kind of inner drives will A.I.s develop? Will they have any at all?

Before finishing this short episode, let's ask once more what is consciousness, and what it is not. First of all, it is not moral. Some computer scientists are mixing up self-awareness and conscious thought with being good (e.g. Gamez 2018). This confusion leads to some transhumanist and IT – minded people to believe that super-intelligent A.I.s would be better humans, avoiding social conflict, crime, and war. This is just an escapist fantasy. Or it is a totalitarian power fantasy hoping to replace diversity with a centralized system. Instead, consciousness is mainly three things: first, it is a function of intelligent systems to observe themselves, build relationships with other intelligent systems, and to navigate and manipulate their shared social worlds. Second, it is an illusion, in the sense that it is constructed as only one particular way of seeing the world. Third it is subjective, as an observer can never understand or rebuild the consciousness of another subject. However, illusion and subjectivity are just not relevant in a social world. Through social negotiation of meaning the existence of consciousness is becoming a social convention. As a result, the entire question of consciousness will never be solved by philosophers and A.I. engineers, but will be absorbed instead by social relations. In the moment that a robot starts to act and communicate in a social context, the question of its consciousness will disappear. We will just assume it is conscious, same as we assume that we are. In the future the technical question of what consciousness is will be replaced by the most important social question of humanity: how to live with robots and other machines. This question is similar to other questions in the past, such as how white people live with black people, rich people with poor people, men with women, or humans with animals, and it may be answered in similar ways. We need to be prepared for conflict.

Bibliography

Baars, Bernard J. (2001, ©1997): In the theater of consciousness. The workspace of the mind. New York, Oxford: Oxford University Press.

Blackmore, Susan (2003): Consciousness in Meme Machines. In *Journal of Consciousness Studies* 10 (4-5).

Blackmore, Susan J. (1999): The meme machine. Oxford: Oxford University Press

Chalmers, David John (2002): The conscious mind. In search of a fundamental theory. Princeton, N.J.: Recording for the Blind & Dyslexic (Philosophy of mind series).

Dawkins, Clinton Richard (1976): The selfish gene. Oxford: Oxford University Press.

Dennett, Daniel Clement (1993): Consciousness explained. London: Penguin.

Gamez, David (2018): Human and machine consciousness. Cambridge: Open Book Publishers.

Koch, Christof (2018): What is consciousness? In *Nature* 557, pp. 8–12.

Koch, Christof (2019): Proust among the machines. Within our lifetimes, computers could approach human-level intelligence. But will they be able to consciously experience the world? In *Scientific American* 321.

Minsky, Marvin (1988): The society of mind. New York: Simon & Schuster

Minsky, Marvin (1994): Will robots inherit the earth? In *Scientific American* 271 (4), pp. 109–113.

Minsky, Marvin (2007): The emotion machine. Common sense thinking, artificial intelligence, and the future of the human mind. New York: Simon & Schuster.

Nagel, Thomas (1974): What is it like to be a bat? In *The Philosophical Review* 83, pp. 435–450.

Searle, John R. (1990): Is the brain's mind a computer program? In *Scientific American* 262 (1), pp. 25–31.

Searle, John R. (2001): Meaning, Mind, and Reality. In *Revue Internationale de Philosophie* 55 (216), pp. 173–179.

Tononi, Guilio; Boly, Melanie; Massimini, Marcello; Koch, Christof (2016): Integrated information theory: from consciousness to its physical substrate. In *Nature Reviews Neuroscience* 17, pp. 450–461.