Can computers think?

An essay on the philosophy of AI.

Since emerging from the Turings (1950) paper, the issue of thinking machines has raised an enormous amount of controversy, becoming a focal point of most philosophical discussions regarding AI. A number of approaches have been put forward to address it. (Oppy and Dowe, 2019) Probably the most influential one, proposed by Turing, could be regarded as rooted in logical positivism. Pointing out that the act of thinking is empirically unverifiable, he dismisses the initial question as "too meaningless to deserve discussion". He suggests taking instead a behaviourist approach, introducing what is now well known as a Turing Test (TT). Some of the problems of the TT (Millican, 2012), widely reflected in the literature, include: the narrowness of taking linguistic ability as a hallmark of intelligence, anthropomorphism, dubious implications for the AI research, failure to encompass the possibility of higher intelligence, neglect in regards to consciousness (Searle, 1984) and blindness to the predicament of inductive reasoning. Moreover, even if we agree on proposed terms, accepting the imitation game and setting the arbitrary parameters on what would suffice as a success, it remains utterly unclear what implications would it have.

Turing (1950) admitted that "opinions will differ as to the appropriateness of the substitution", yet implicit in his argument is a pragmatic concern for the matter. What was of paramount interest then and remains such now, is what exactly the machines will ever be capable of doing (Moore, 2006). Attempts to answer that question flourished both in scientific research and fiction, evoking wildest dreams in our collective subconsciousness. What is noteworthy is that from this perspective, TT appears to be nothing but a milestone of AI development carrying no moral consequences. Technology is what matters.

Tremendous efforts are being made in the attempts to model and formalize the inner workings of the mind. (Bringsjord and Govindarajulu, 2018) Some notable approaches under the broad umbrella of cognitive sciences include computational theories of mind, connectionism, enactivism, functionalism, memory-prediction framework (Hawkins, 2004) and many others. Yet numerous technical issues and futuristic speculations are out of the scope of this essay. What I would like to contend with is the existential element of the issue embedded in the initial formulation. Turing could, from a materialistic viewpoint, see it as "Heads in the Sand" unease, but I would rather regard it as a metaphysical predicament of a contemporary man. This is an old philosophical conundrum particularly salient today with the rise of sophisticated machinery - it is whether a human being is a machine.

To grapple with this problem, we will have to put things into a very broad perspective. We could start from the point in the evolution of human consciousness referred by Yuval Noah Harari (2014) as a "Cognitive Revolution" which took place around 70,000 BC. This is a gross dating of the start of the period called behavioural modernity. (Klein, 1995) Psychologically, it could be thought of as the process of the discovery of the future as such and theologically, it is reflective of the fall of man. (Peterson, 1999) Since then, we have started building higher order models about how the world works to predict the outcomes of our behaviour. This was the beginning of the quest to control reality through knowledge. Perhaps this was also a period when we started to be able to transgress our immediate instinctive urges and act accordingly to a higher order understanding of the consequences. Back then the acquisition of knowledge was sporadic and its transmission unreliable. It is the agricultural revolution which coerced bigger groups of people to cooperate through the generation of complicated collective narratives (Peterson, 1999), and then the development of the written word that eventually brought the scientific revolution about (Harari, 2014). Of course, this is a blatant and speculative oversimplification but it is also a condensed view of how we come to the important point of formalizing the process of knowledge acquisition. From then on we have been in a positive feedback loop of complexification of our models of reality. Yet one of the great dangers we shall see being implicit in this sophistication is the failure to distinguish between the model and reality it represents.

Our models proved to be useful. They enabled us to often correctly predict the outcomes of our actions. By acting through them we could bring forth the reality we desire. Yet humans are even more ingenious - we also use tools to leverage our impact on the world. In this context machine is nothing but a sophisticated tool capable of scaling our influence. An ultimate tool would be a machine embodying our model of reality. Thus the implicit goal of AI research is to extract the model underlying the workings of our mind, formalize it as what is generally called a cognitive architecture and then build it into a computer. What is to a bigger or lesser degree common between all such architectures is their view of the mind as an information processor. Thinking is understood as symbol processing. Examination of this assumption would require an inquiry into the nature of the symbol as such. This is a tremendous ambition so I am compelled to again vastly oversimplify the matters. For that purpose, I suggest defining a symbol in Neo-Kantian and phenomenological spirit as a meaning manifested through boundaries confining phenomenon. In fact, boundaries are exactly what allow the phenomenon to reveal itself by distinguishing it from the infinity of preconceptual reality. But where do they come from?

In both Kantian and Neo-Kantian tradition boundaries of a perceived phenomenon are determined by the a priori forms of perception - time and space. In Husserl's phenomenology, they are established by the intentionality of the subject. (Siewert, 2017) They both reveal the significance of the realm of values where values play as the ground for the meaning prioritisation in the acts of consciousness. As could be derived from the philosophy of Ernst Cassirer, also becomes apparent the predicament of the ability of conscious thought to produce symbolic forms which could be done only by transcending those boundaries. It is a mystery beyond our reductionist understanding. Moreover, whenever such boundary is unsatisfactory in attaining a corresponding end, it means it should be relocated. Such a shift could probably be explained in computational terms of pattern prediction as it is done in the memory-prediction model of the mind, but the goal, an intention, still comes first, unexplained. To say that is to say values underlie not only our ability to construe the world but the very perception of it. (Peterson, 1999)

The fact that intentionality precedes the formation of symbolic forms means we couldn't even talk about symbols beyond a conscious subject. We could say that a discrete-state machine processes symbols, but we should not take symbols themselves for granted. We imbue the machine with our externalized representations, yet they emerge out of consciousness. From this perspective, the capacity of the machine to think is highly questionable. In a narrow sense, machines could be said to be able to think, but they think our thoughts and in this way are a mere instrumental extension of our own minds. Strictly speaking, the machine is not a subject and so could do nothing, we act through them. For the machine to emerge as an independent being, it will have to become an entity completely exempt from pragmatic limitations we set upon it and thus from our symbolism. It will only cease to be our extension and become a free agent if it has broken free of our values and developed its own, which arguably is possible only through an evolutionary process. Thus, we came to yet another dilemma, this time of value emergence.

We've said that values allow us to confine the phenomenon out of the infinity of possibilities and redefine its boundaries if necessary. But sometimes it is not the boundary that is being shifted, but the objective setting it. In this way we come to the same problem of infinite possibility in regards to values: how could one value be replaced by another? It is impossible without a higher order value which would allow prioritizing its subordinates. This constitutes a critique of naturalistic ethics we try to embed in machines: any values, justified by the necessities of the limited physical world, manifesting themselves through instincts or higher-order behavioural patterns, could relate to other values only by subjugating themselves to a higher value, thus producing a value hierarchy with one value at the top. Moreover, the highest value should be transcendent to all the naturalistic limitations justifying subordinate values or otherwise, it could just as well be pushed around and replaced as things change, therefore undermining the whole hierarchy.

Where could this transcendent value come from and in what relation does it stand to the subject in relation with it? Obviously, such ontological questions could not be decisively answered. My point here is

not to close the issue but to try to reveal the tremendous uncertainty underlying it and warn against falling into materialistic dogmatism (Sheldrake, 2012) by assuming metaphysical fabric of reality not only could be understood but is already understood as ultimately reducible to matter. When it comes to AI, this is the most prevalent approach and although extraordinarily useful, materialistic stance becomes problematic when it comes to existential questions - it implies either an impossibility of consciousness, our only certainty, as pointed out by Rene Descartes, or an acceptance of a dualistic panpsychic view.

This is now our predominant scientific paradigm which shapes attitudes on questions such as the one discussed and many others. The man as a machine riddle is probably the most prominent such issue with a long list of both social and personal implications and revolves around a well known free will conundrum. On the one hand, the concept is incompatible with the deterministic worldview nested in materialistic philosophy (Harris, 2012), on the other, our social structures are predicated on the idea of an integral free individual (Peterson, 1999). Attempts to dismiss it leave us amid deep valleys of numinous uncertainty and existential nihilism. My remedy for that is an awareness that models of reality however refined and all-encompassing are after all still just models that should not be confused with reality itself. As perspicaciously contemplated by Martin Heidegger, the relationship of beings to being remains a deep mystery. It is from this position of a profound uncertainty about the nature of existence, let alone a thinking machine, that I suggest being extremely cautious of casting any final judgments.

References

Bringsjord, S., Govindarajulu, N. S. and Zalta, E (ed.) (2018). Artificial Intelligence. [online] The Stanford Encyclopedia of Philosophy. Available at: https://plato.stanford.edu/archives/fall2018/entries/artificial-intelligence/ [Accessed 16 Apr. 2019]

Harari, Y. (2014). *Sapiens: A Brief History of Humankind*. London: Vintage Digital.

Harris, S. (2012). *Free Will*. New York: Free Press.

Hawkins, J. and Blakeslee, S. (2005). *On Intelligence: How a New Understanding of the Brain Will Lead to the Creation of Truly Intelligent Machines*. New York: St. Martin's Griffin.

Klein, R. (1995). Anatomy, behavior, and modern human origins. Journal of World Prehistory, 9(2), pp.167-198.

Levesque, H. (2017). *Common Sense, the Turing Test, and the Quest for Real AI*. Cambridge (MA): The MIT Press

Maguire P., Moser P. and Maguire, R. (2015). A clarification on Turing's test and its implications for machine intelligence. In: 11th International Conference on Cognitive Science. pp. 318-323.

Millican, P. (2012). The Philosophical Significance of the Turing Machine and the Turing Test.

Moor, J. (2006). The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years. AI Magazine 27: 87-91.

Oppy, G., Dowe, D. and Zalta E. (ed.) (2019). The Turing Test. [online] The Stanford Encyclopedia of Philosophy. Available at: https://plato.stanford.edu/archives/spr2019/entries/turing-test/ [Accessed 16 Apr. 2019]

Peterson, J. (1999). *Maps of Meaning: The Architecture of Belief*. New York: Routledge

Searle, J. (1984). *Minds, Brains and Science*. Cambridge (MA): Harvard University Press.

Sheldrake, A. (2012). *Science Set Free: 10 Paths to New Discovery*. New York: Deepak Chopra.

Siewert, C. and Zalta E. (ed.) (2017). Consciousness and Intentionality. [online] The Stanford Encyclopedia of Philosophy. Available at: https://plato.stanford.edu/archives/spr2017/entries/consciousness-intentionality/ [Accessed 16 Apr. 2019]

Turing, A. (1950). Computing Machinery and Intelligence. Mind 49: 433-460.