Review of: "What is it like to be an AI bat?"

Igor Aleksander[1]

1 Imperial College London

Potential competing interests: No potential competing interests to declare.

At last! Here we have authors (Herzog and Herzog) brave enough to take the step of discussing theories of consciousness that include machine consciousness as a field of research. That's the good news. The less wonderful news is that while their article is an excellent journey across the recent work of computer-based laboratories where consciousness is of interest, a coherent definition of what machine consciousness is does not emerge. This is not the authors' fault, and they, in fact, indicate that such a position may not be achievable. It's just that there are many computational approaches that could be said to be about *some aspect* of consciousness but which fail to cohere with one another in a constructive way. The current authors are aware of this, and they ask pertinent definitional questions in their introduction. Sadly, such focusing becomes a little less evident as the paper progresses, as attitudes that are not altogether helpful appear. There is an elephant in the room that conscious systems may *emerge* from future AI development, which is a view often held by the popular media. The truth is that AI development is most likely only to be a useful tool that follows a developing theoretical understanding of what is actually meant by the use of the C-word in a computational context. This is not quite happening as most descriptions run like 'this is what I think that consciousness is, and I will build AI to express it.'

I must confess that I was driven by something like this when developing an 'axiomatic' approach in the modelling of consciousness. For me (Aleksander and Dunmall, 2003), it became important not to model some aspects of consciousness that could be brought in to improve the behaviour of some AI, but to ask the question 'what must a system have so that one can argue that it is *minimally* conscious for functional reasons'. I found aspects of Automata Theory as applied to dynamic neural networks helpful, so my feeling for machine consciousness started drifting towards state machine representations where, through learning, the internal state structure develops to represent the experience of the entity. States are representations of what Damasio identifies as 'feelings' (2013). More of this in a forthcoming publication. In a future version of the current paper, Herzog and Herzog may wish to explore this approach.

Another opportunity for these authors is to look a little more carefully at the historical track of general beliefs about conscious machines. They name 1969 as the first appearance of “Artificial Consciousness” in English scientific literature (sadly, only giving an indirect reference to this event). This may be true, but the sentiment of whether a machine could exhibit something which could be called conscious thought goes back a longer way. In Turing's discussion of whether a machine could 'think', “Computing Machinery and Intelligence” (1950), he refutes *objections* to the idea that computers might 'think'. One of these objections was the oration by Geoffrey Jefferson, “The Mind of Mechanical Man”. Jefferson was a famed neurologist who attacked the notion that a machine might think in his distinguished Royal College of Surgeons oration.

Jefferson argued that “Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that a machine equals a brain— that is, it has not only written it but knows that it had written it.” The highly popular large language systems that hit the headlines at the current time are undoubtedly capable of writing mediocre sonnets, but their linguistic machinations can hardly be described as “because of thoughts and emotions felt.”

Turing labelled this opposition “The Objection From Consciousness,” drawing attention to the solipsism inherent in ‘knowing that I think’ being a fact, but ‘knowing that you think’ becoming a mere matter of politeness. [Here, to keep the argument short, I simply submit that thought is a conscious activity and that consciousness is a space of such activities]. But, historically, the point is that if I were convinced that I was conscious, there is no scientific reason why the above politeness cannot transfer to machines. Indeed, this is a sentiment that arose from a meeting sponsored in 2001 at the Cold Spring Harbour Laboratories (CSHL) by the Swartz Foundation (which normally funds scientific meetings on brain studies). It addressed the question 'Could Machines Be Conscious?'. While there was little agreement on precise definitions of consciousness among the audience of 21, made up of neuroscientists, philosophers, and computer scientists, there was agreement on the following proposition. “There is no known law of nature that forbids the existence of subjective feelings in artefacts designed or evolved by humans. This stimulated many of the contributions mentioned in (http://www.swartzneuro.org/abstracts/2001_summary.asp)”. In the years which followed, this gave rise to some of the research discussed in the Herzog and Herzog paper.

So, well done, Herzog and Herzog, but here’s hoping that the future will give rise to a scientific judgement on whether what makes some machinery effective is or is not something that might be called machine consciousness, not only by the insistence of its creators.

References

Aleksander, I. and Dunmall, B. Axioms and tests for the presence of minimal consciousness in agents. J. Conscious. Stud., vol.10, pp. 7–18. 2003

Damasio and G. Carvalho, “The nature of feelings: evolutionary and neurobiological origins,” Nature Reviews Neuroscience, vol. 14, no. 2, pp. 143-152, 2013.

Turing, A. M. Computing Machinery and Intelligence. Mind, vol. 49, pp. 433-460, 1950