Review of: "Artificial Consciousness: Misconception(s) of a Self-Fulfilling Prophecy Nobody Wants"

Gerald Loeb[1]

[1] University of Southern California

Potential competing interests: No potential competing interests to declare.

The author is correct to point out the relatively primitive state of the deep learning neural networks now being touted as potentially “conscious” compared to the intricate developmental, experiential and computational processes embodied by human and, discomfitingly, other higher vertebrate brains.

Many of the cited properties of consciousness are manifestations of attentiveness, particularly when attention depends on the current importance of and discordance between current and prior experience. A more recent theory of thalamocortical function than the one cited provides a mechanistic basis (Halassa and Sherman, 2019) that is already being added to neural network models of artificial intelligence (George et al., 2020). That seems likely to result in emergent behaviors and associated temporal waves of activity *in silico* that are more similar to those of conscious humans (Loeb, 2023). Whether those behaviors and waves constitute features or bugs or the *sine qua non* of all consciousness is the void at the heart of this whole field.

Recent accomplishments in the field of artificial intelligence have already forced those who would distinguish human intelligence from AI to move last century’s goalposts (Turing test interactions, natural language recognition, strategic game-playing, etc.). Now they are located on turf where robots perform poorly (dexterous manipulation, novel tool use, autonomous vehicles, etc.) but the new goalposts may well fall (eventually) to simulations of robotic infancy (Loeb, 2022). The last, desperate stand may be to require intelligence to employ biological mechanisms or to exhibit the human failings that arise therefrom. This article revisits that retreat for the even more nebulous concept of consciousness. By making consciousness a binary rather than a continuous attribute, the author forecloses different levels of consciousness in non-human animals and machines.

Perhaps the motivation to assert the qualitative superiority of human intelligence and the unique existence of human consciousness is to justify continuing human control over machines and animals (the philosophical basis for Descartes’ defense of vivisection). At least until we foolishly hand over the keys to the machines that we have created, the last stand of human speciesism should hold.

References

George, D., Lazaro-Gredilla, M., Lehrach, W., Dedieu, A., and Zhou, G. (2020). A detailed mathematical theory of thalamic and cortical microcircuits based on inference in a generative vision model. *bioRxiv* https://www.biorxiv.org/content/10.1101/2020.09.09.290601v1.

Halassa, M.M., and Sherman, S.M. (2019). Thalamocortical Circuit Motifs: A General Framework.*Neuron* 103**,** 762-770.

Loeb, G.E. (2022). Developing Intelligent Robots that Grasp Affordance. *Frontiers in Robotics and AI* 9, 951293, doi.org/10.3389/frobt.2022.951293.

Loeb, G.E. (2023). Remembrance of things perceived: Adding thalamocortical function to artificial neural networks. *Frontiers in Integrative Neuroscience* 17**,** 1108271, doi.org/10.3389/fnint.2023.1108271.