Review of: "Towards Modeling Artificial Consciousness"

Steven Bickley[1]

1 Queensland University of Technology

Potential competing interests: No potential competing interests to declare.

First of all, thank you Maksym for publishing your draft pre-print to Qeios. And thank you to the Qeios team for the invited opportunity to publicly review this manuscript.

I can appreciate the work that has gone into developing this paper, particularly in a technical/empirical sense. Overall, I believe that this work has the potential to make an important (and generally interesting) contribution to the field of AI research, and makes links to some interesting references (e.g., Shevchenko 2016, Dehaene, Lau, & Kouider 2017, Graziano 2017). The use of structured mathematical formulae is a notable strength to provide some concrete empirical insights into the modelling of consciousness (i.e., as non-linear connections between segments of different neural networks), but the paper could be improved by providing clearer and more well-supported argumentation, including particularly more clarity around the scope, practical/empirical application, and novelty of the work presented in the paper itself. With revisions, I believe in the potential for this work to be an important addition to the literature.

"Towards Modelling Artificial Consciousness" proposes an approach to modelling (artificial) consciousness for more advanced Artificial Intelligence (AI). The author argues that there are many properties of consciousness identified in the literature that could serve as criteria to proxy/identify the emergence of artificial consciousness in AI that takes into account the more fluid and unpredictable nature of human consciousness. The solution they present (a system of neural networks with non-linear connections between segments of different networks) is interesting and feasible. However, I wonder if we jump too quickly to a solution without building the ground work (linking to what other work has been done previously, being more explicit about what this approach offers that others do not – strengths, weaknesses, etc.). For example, what are some other approaches, why should we adopt this approach over the others (e.g., in terms of relative model strengths/weaknesses).

One major weakness of the current version of this paper is its lack of clarity. For example, while the paper provides an initial statement of “[T]he problem of artificial consciousness attracting much interest…” (p. 1), but doesn’t provide a definition or practical/intuitive examples to help frame the readers’ interpretation/comprehension. While the use of mathematical formulae is a strength of the paper, sometimes the current version of the paper fails to clearly connect these formulas to the overall argument. A graphic/figure would probably help here as well. Some more explicit definitions of e.g., consciousness, information processing systems, attention, global broadcasting, self-monitoring, self-control, subjective perception, emergence, etc., and more concrete examples of how the proposed model could be implemented in practice, and what sorts of benefits/challenges that could emerge from the adoption/application of the proposed model, would help to make the paper more accessible to a wider audience. For example, how the proposed model could be implemented in

practice, what types of applications it might be well-suited for, and how the model might be evaluated or tested.

Some more specific (still general) feedback:

1. The abstract is currently quite broad/general. It would help to signal more explicitly the specific scope and contribution of the paper.
2. The introduction section needs a lot more work to motivate the problem and solution.
3. It would be valuable to connect or interact more with the social and human sciences which have substantial literature around self- and social-consciousness. Whilst I understand this paper is focused more on methods, this interaction could help the author draw more relevant/intuitive examples about e.g., definitions of important concepts, implementation steps, applications of the model, etc. and to help identify potentially relevant empirical settings to validate the model.
4. Aaron Sloman (https://www.cs.bham.ac.uk/~axs/) does some really interesting work on AI and biologically-informed information processing systems that is worth exploring to help enframe the paper's argumentation. For example, "Natural and artificial meta-configured altricial information-processing systems" published along with Jacki Chappell (expert in field of animal cognition), I found to be a really interesting read.
5. In addressing the previous comments this feedback will become less relevant, but we do not really discuss any alternatives here. What about the literature on cognitive architectures more generally? There is likely to be relevant work on consciousness in this literature that may help inform/enframe this paper.

Again, thank you Maksym for publishing your work. I hope some of the comments/feedback help inform the next revision of your paper, and I look forward to seeing how this work of yours develops over time.