Higher-order theories do just fine

Matthias Michel[1,2] & Hakwan Lau[3,4,5,6]

[1] Centre for Philosophy of Natural and Social Science, London School of Economics and Political Science, London, UK.
[2] Consciousness, Cognition & Computation Group, Center for Research in Cognition & Neurosciences, Université Libre de Bruxelles.
[3] Department of Psychology, University of California, Los Angeles
[4] The Brain Research Institute, University of California, Los Angeles
[5] Department of Psychology, The University of Hong Kong
[6] State Key Laboratory for Brain and Cognitive Sciences, The University of Hong Kong

Abstract: Doerig et al. have set several criteria that theories of consciousness need to fulfill. By these criteria, higher-order theories fare better than most existing theories. But they also argue that higher-order theories may not be able to answer both the 'small network argument' and the 'other systems argument'. In response, we focus on the case of the Perceptual Reality Monitoring theory to explain why higher-order theories do just fine.

We applaud Doerig et al.'s effort in providing a systematic analysis on the theoretical landscape of the current science of consciousness. In particular, we agree that there are currently too many theories, relative to our efforts in arbitrating between them. Elsewhere we have speculated how this might have come about (Lau & Michel, 2019; Michel, 2019). If this is partly a structural problem at a sociological level, we worry that sheer logical analysis will not be sufficient to remedy the problem.

That said, we are sympathetic to most of their arguments, and only have a few points to clarify. First, a small terminological note: Doerig et al. grouped our theory under the label Higher-Order *Thought* Theory (HOTT). There are many varieties of higher-order theories, ranging from historical ideas that can be traced to Kant and Locke, to modern versions based on computational neuroscience. Our perceptual reality monitoring (PRM) theory is different from Rosenthal's higher-order *thought* (HOT) view (Brown et al. 2019; Lau, 2019). But in the following, most of our reply applies to both.

The challenge raised by Doerig et al. is the following: either the term 'thought' in Higher-Order Thought is understood in its everyday sense, in which case it is not clear how the theory can be generalized to other systems (the other systems argument). Or a 'higher-order thought' could be defined as any two-stage computation, but in this case even small networks could be conscious (the small network argument).

Let us begin with the small network argument. It is clear that according to both PRM and HOT, very simple networks with just a few nodes forming a hierarchy will not suffice. On HOT the requirement is *not* that we need conscious thoughts; but there has to be something thought-like, capable of *conceptual* representation. Whether the small network argument applies or not then depends on one's theory of concepts (Laurence & Margolis, 1999), but we doubt that small networks can implement something like a language of thought, or have genuine conceptual capacities at the moment (Fodor, 2008; Rosenthal, 2004).

Focusing on PRM, it is also clear that the theory avoids the small network argument. In this case, the relevant higher-order mechanism has to first be able to perform the fairly challenging computational task of distinguishing endogenously generated from externally triggered sensory activities. Second, the same circuit also has to distinguish between meaningful sensory signal and noise. Third, and most importantly, this mechanism has to send output to a general belief-formation and rational decision-making system, for otherwise it will not be performing the relevant functions. That is to say, PRM concerns agents who are capable of predictive sensory coding of a particular kind, with general reasoning capacities. Current AI systems have not achieved this, regardless of whether they are implemented in large or small networks.

This leaves us with the second horn of the dilemma: can the theory be generalized to other systems? If a system implements a perceptual reality monitoring mechanism capable of influencing its rational decision making and beliefs, will that system be conscious? Our stance is: yes in principle (Dehaene et al. 2017), but this *may* be much harder than one thinks. If an AI system shall one day be able to perform 100% like a human being, *and* if the inner workings of the system are known to be identical to our brain/mind at both the algorithmic and computational levels (Marr, 1982), then perhaps many of us will be prepared to accept that the system is conscious.

But not only are we very far from this state of knowledge and technology, there are also reasons to doubt if we ever will be – especially if we recognize that the system has to mimic ours *in real time*. As philosophers have recognized, multiple realizability has probably been overestimated in the past (Polger & Shapiro 2016; Michel, 2018). Hardware requirements set limits to multiple realizability: perceptual reality monitoring functions can't be realized in Swiss cheese. Sometimes there aren't so many different solutions to realize a given function. This is why computational modeling work has been so useful in neuroscience, as artificial networks often spontaneously show resemblance to biological systems (e.g. Banino et al. 2016).

It could be that, at the end of the day, to implement a PRM system efficiently – together with the correct dynamics of a predictive coding sensory system, *and* a general reasoning system to which the PRM system outputs – we will end up having to resort to something so neuromorphic that it would essentially be biological (Godfrey-Smith, 2016). So, although the theory predicts that an AI system could be conscious in principle (Dehaene et al. 2017), this is not a point we insist on. It is not yet clear if our hardware departments can easily come up with future alternative solutions to a metabolically self-regulating nervous system.

Therefore, PRM meets all the criteria set by Doerig et al.

As a final note, we wonder why the unfolding argument doesn't also apply to the thalamocortical loop theory and the NMDA theory. If we put these specific biological substrates in a functionally detached or isolated context, would they continue to be conscious? Likewise, to what extent does the small network argument not apply to adaptive resonance theory, attention schema theory, and the sensorimotor theory too? Can we not create relatively simple artificial neural networks and agents that will achieve these functions at least to some degree?

References

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., … King, H. (2018). Representations in Artificial Agents. *Nature*, *557*, 427–433.

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the Higher-Order Approach to Consciousness. *Trends in Cognitive Sciences*, *23*(9), 754–768.

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness and could machines have it? *Science*, *358*, 486–492.

Fodor, J. (2008). *The Language of Thought Revisited*. Oxford University Press.

Godfrey-smith, P. (2016). Mind, Matter, and Metabolism. *Journal of Philosophy*, *113*(10), 481–506.

Lau, H. (2019) Consciousness, Metacognition, and Perceptual Reality Monitoring. *BioRxiv.* https://psyarxiv.com/ckbyf/

Lau, H. & Michel, M. (2019) A socio-historical take on the meta-problem of consciousness. *Journal of consciousness studies*, 26(9-10), 136-147.

Laurence, S., & Margolis, E. (1999). Concepts and Cognitive Science. In Margolis, E. & Laurence, S. (Ed.), *Concepts: Core Readings* (pp. 3–81). Cambridge, MA: MIT Press.

Marr, D. (1982). *Vision. A computational investigation into the human representation and processing of visual information.* Cambridge, MA: MIT Press.

Michel, M. (2018) Fish and Microchips: On fish pain and multiple realization. *Philosophical Studies*, 176, 2411–2428.

Michel, M. (2019) Consciousness Science Underdetermined: A brief history of endless debates. *Ergo*, 6(28).

Polger, T. & Shapiro, L. (2016) The Multiple Realization Book. Oxford University Press.

Rosenthal, D. (2004). Varieties of higher-order theory. In Rocco J. Gennaro (ed.), *Higher-Order Theories of Consciousness: An Anthology*. John Benjamins.