Review of: "Questioning the Moratorium on Synthetic Phenomenology"

Jonathan Edwards[1]

1 University College London, University of London

Potential competing interests: No potential competing interests to declare.

The point being made in this article is relevant and well motivated, but I give a low rating simply because I do not think the argument is at present enough to justify formal publication. We can all say, and perhaps agree, that Metzinger's case is muddled. We can see he is mixing up categories. What is of interest is the detail of where he may be making false assumptions.

Some specific thoughts:

Cognition is such a muddled word that I try to avoid it. It conflates intelligence/computational function with conscious thought, as the author notes.

We have no good reason to think phenomenal experience has much to do with computation or circuitry, despite the popular stance of people like Edelman or Tononi. The author accepts the premise that ‘systems’ are conscious, but I do not think neuroscience supports that. When we say a man is conscious, we really mean that we infer that, somewhere within a brain within a certain body, representations are manifest in events of the sort we experience ‘here’. Descartes was probably roughly right. The representations are mostly of others, from a point of view. As Hume said, there is little or no representation of self. Certainly human subjects have no access to representations of the ‘systems’ that provide the representations and that Metzinger would seem to regard as the self. There is a ‘sense of a self,’ but the more we know, the more we see how illusory this is likely to be.

As an example, a meticulous hand-painted copy of a Picasso portrait and a video screen display of a photograph may function as indistinguishable representations despite completely different ‘systems’ generating them. Phenomenal experience need not, empirically, have anything to do with antecedent (or consequent) computation.

Metzinger argues against deliberate attempts to synthesise human-like consciousness, or efforts that might run the risk of doing so. That is fair enough. But the counterargument is that nobody much is likely to be trying to build things that experience pain, and we have no idea at all what the risk is. The only real issue is phenomenality with negative valency, and we have absolutely no clue where that occurs. It might be that when we hit a nail with a hammer, there is an experience in the nail of negative valency. Photons might be in deep pain. Mountains might, like Prometheus, be condemned to agony for millennia. Who has any idea? Perhaps there is a certain hubris in assuming that negative phenomenal valency is unique to ‘animal spirits’.

One last thought on ‘systems’. Almost all AI devices will be connected to the internet. What is the ‘system’? The internet? All hardware in northern Milwaukee? Sue's laptop? The central processor? There are no ‘systems’ in reality when complex computational routines are interconnected the way the net does it. So there are no systems to have a sense of self or pain. The real question is whether or not some event in some structure at a particular place - maybe the uploading of a complex set of signals representing ‘a bad situation’ - carries negative valency. But we have no reason to think that if it does, it has anything to do with the antecedent computations. Exposing the same part to sunlight might be much more painful!

Thanks for an entertaining read.