On Consciousness

Paul Kotschy

9 August 2023

Compiled on December 14, 2024

WHAT DOES IT MEAN to be conscious? Can only biological systems be conscious? My preliminary ideas on the nature of consciousness are based in part on Kanai's proposed Information Generation Theory (IGT) of consciousness[1] and on the Global Workspace Theory (GWT) of consciousness by Baars[2] and Dehaene et al.[3]

I consider the essence of consciousness to be *aggregates of coherent interactions between internal sensory representations of past and future external events*. There is some experimental support for this idea of consciousness in biological systems (Kanai[1, p85]). Significantly, aggregates of these interactions between representations allow for non-reflexive behaviours, some of which are strongly associated with consciousness, such as, intention, attention, planning, imagination, curiosity and creativity.

When these internal sensory representations of external events interact inside our own heads, we sometimes think of it as *subjective experience*. To be sure, not all internal representation interactions can be considered subjective experience. For example, after we have learned to focus our eyes, we almost always do so sublimely unconsciously. Conversely, subjective experience is not possible without the internal representations and their interactions.

How then are these internal sensory representations sculpted? They are sculpted over time through a process known as meta-reinforcement deep learning (MRDL) using available short-term and long-term memory. Indeed, as humans we are not born with consciousness, but acquire it over time as we learn about our immediate external environment, and then as we learn to learn. MRDL happens when a learning system is 1. endowed with both short-term—or "working"—memory and long-term memory, and 2. exposed to many interrelated tasks.

It is tantalising that synthetic (AI) and biological (brain) information systems are both equipped with 1. and 2., and that both have been shown to sculpt internal representations using MRDL. And since consciousness is identified as the interactions of these internal representations, it logically follows that consciousness is not limited to biological information systems.

I therefore expect synthetic consciousness to emerge spontaneously at a time when computational neural networking, animatronics and environmental sensing are brought together in non-trivial ways. And that is not *if*, but *when*.

References

[1] R Kanai et al. Meta-learning, social cognition and consciousness in brains and machines. *Neural Networks*, 145:80–89, 2022. Retrieved from `https://doi.org/10.1016/j.neunet.2021.10.004` on 10Jan23.

[2] BJ Baars. Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Progress in Brain Research*, 150:45–53, 2005. Retrieved from `https://doi.org/10.1016/S0079-6123(05)50004-9` on 11Jan23.

[3] S Dehaene et al. A neuronal model of a global workspace in effortful cognitive tasks. *Proceeding on the National Academy of Sciences of the United States of America*, 95(24):14529–14534, 1998. Retrieved from `https://doi.org/10.1073/pnas.95.24.14529` on 11Jan23.