Dissociating Intelligence from Consciousness in Artificial Systems – Implications of Integrated Information Theory

Graham Findlay[1,a], William Marshall[1,a], Larissa Albantakis[1], William Mayner[1], Christof Koch[2], and Giulio Tononi[1],

[1] University of Wisconsin-Madison, Madison WI, USA
gtononi@wisc.edu
[2] Allen Institute for Brain Science, Seattle WA, USA
[a] These authors contributed equally to this work

Abstract. Recent years have seen dramatic advancements in artificial intelligence (AI). How we interact with AI will depend on whether we think they are conscious entities with the ability to experience, for example, pain and pleasure. We address this question within the framework of integrated information theory (IIT), a general, quantitative theory of consciousness that allows extrapolations to non-biological systems. We demonstrate (manuscript submitted elsewhere) an important implication of IIT: that computer systems with traditional hardware architectures would not share our experiences, even if they were to replicate our cognitive functions or simulate our brains in ultra-fine detail.

Keywords: Consciousness, Integrated Information Theory, Artificial Intelligence

1 Summary

If a computer were functionally equivalent to a human, having the same cognitive abilities, would it necessarily experience sights, sounds, and thoughts, as we do when we are conscious?

Integrated Information Theory (IIT) [1, 2] offers a principled explanation for how consciousness relates to physical substrates, aims at providing a mathematical framework to determine whether and how a system is conscious [2, 3], makes testable predictions, and is supported by scientific evidence about our own consciousness [1].

We applied IIT to a simple target system, constituted of a few discrete dynamical logic gates, which minimally satisfies the properties required by IIT to support consciousness, and to a more complicated 'computer' system, constituted of dozens of logic gates, which is programmed to be functionally equivalent to the target system (Fig. 1). Using the quantitative framework of IIT, we show that even though the computer replicates the input-output functions of the target system exactly and indefinitely, it fails to replicate its purported experience. Our results demonstrate that, according to IIT, functional equivalence ("doing the same thing") does not imply phenomenal equivalence ("having the same experience").

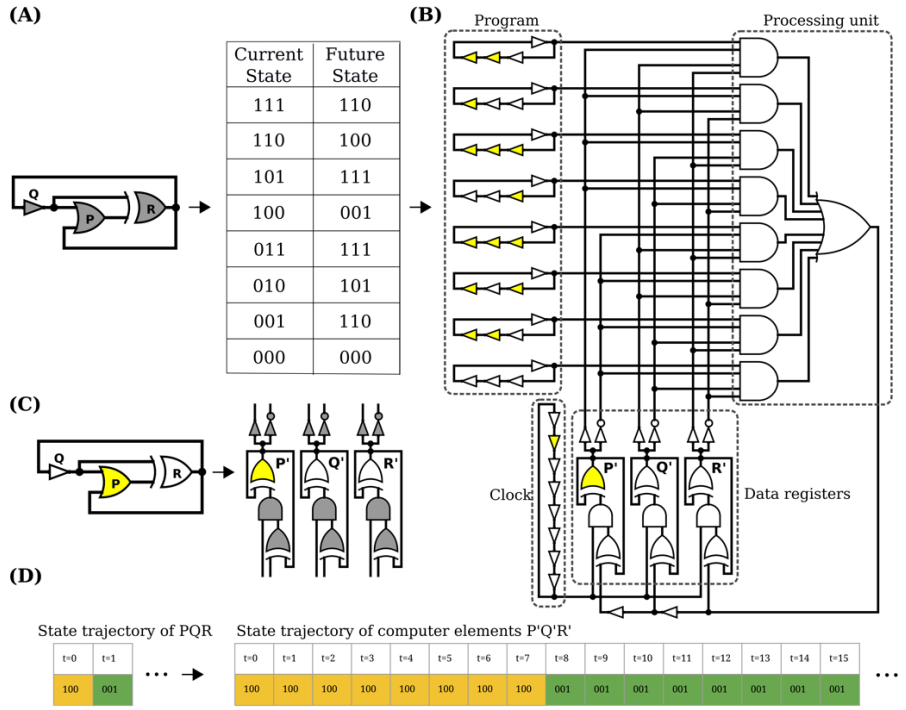

Fig. 1. (A) A target system to be simulated. (B) A simple computer that simulates the target system indefinitely. (C) The mapping between the output states of the target system, and the output states of the computer. (D) State trajectories of the two functionally equivalent systems.

As an extreme illustration of the dissociation between function and phenomenology, we also analyze a Turing-complete computer (not shown) that is negligibly conscious according to IIT, regardless of its programmed function. If computable functions exist for simulating human brains, then they could therefore be implemented by this computer while its experience remained negligible. We conclude that computer systems with traditional architectures, even if they were to replicate our cognitive functions or simulate the neuronal interactions occurring within our brain, would not replicate our experiences.

References

1. Tononi, G., Boly, M., Massimini, M., Koch, C.: Integrated information theory: from consciousness to its physical substrate. Nature Reviews Neuroscience, 17(7), 450–461 (2016).
2. Oizumi, M., Albantakis, L., Tononi, G.: From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. PLoS Computational Biology, 10(5), e1003588 (2014).
3. Mayner, W. G. P., Marshall, W., Albantakis, L., Findlay, G., Marchman, R., Tononi, G.: PyPhi: A toolbox for integrated information theory. PLOS Computational Biology, 14(7), e1006343 (2018).