Towards Conscious AI

Ryota Kanai
CEO & Co-founder, Araya, Inc., Tokyo, Japan

Abstract

The question of how the brain generates consciousness in humans and other animals is called the Hard Problem and remains an unresolved problem in modern science. One of the reasons why consciousness defies modern science is that we cannot directly observe subjective experiences evoked by sensory stimuli such as visual and auditory signals. This led many to think consciousness might be an epiphenomenon without any functional consequences. On the other hand, artificial general intelligence (AGI) – one of the ultimate goals of artificial intelligence research – is defined in terms of functions and therefore it is often assumed that AGI would be achieved independent of the presence of consciousness within AI systems. Here, I will present a hypothesis as to possible functions of consciousness from the viewpoint of empirical research in neuroscience. This hypothesis termed Information Generation Hypothesis claims that consciousness is generated sensory representations produced by internal models (i.e. generative models) of the environment and the self. Defined this way, consciousness (or generative models) enables an agent to simulate and learn from counterfactual (i.e. simulated) situation of itself in the future and the past. This detaches the agent from the present moment and allows for non-reflexive behaviours purely driven by current sensory inputs. The unity of consciousness is achieved by aligning goals of multiple, otherwise factorised models. This architecture allows an agent to combine their internal models flexibly to achieve multiple goals and facilitates transfer learning, gradual learning and meta-learning.