From Attention to Consciousness?

Joscha Bach

Intel Labs, Santa Clara, CA 95054, USA; joscha.bach@intel.com

Artificial Intelligence research offers a unique opportunity for understanding consciousness, via formally stating its necessary and sufficient conditions, so that theories of consciousness become testable via computational models. At the same time, characteristics of learning information processing systems, when contrasted with human performance, give rise to hypotheses of the role of consciousness in the human mind.

Consciousness, informally as ‘a feeling of what it’s like’, can be characterized by its phenomenology (the experiential qualities of features, objects and situations that the conscious agent is attending to, which are the content of consciousness), access consciousness (awareness of the fact that and the mode in which the system attends), and reflexive consciousness (awareness of consciousness as a conscious content).

Within contemporary cognitive science and AI, several, partially converging theories compete in how they describe consciousness, some of which are *Global Workspace theory* (GWT) [1, 2, 3], *Integrated Information Theory* (IIT) [4, 5, 6], Attention Schema Theory [7], the Consciousness Prior [8] and Illusionism [9]. GWT proposes a simple cognitive architecture, within which consciousness is tied to the inner perception of a mental stage, based on an attentional ‘spotlight’ that selectively observes mental agencies, while others are working ‘behind the stages’ (subconsciously). Dehaene’s extensions of GWT anchor it in neuroscience, by functional mechanisms that describe the distribution and integration of integration in the neocortex [3, 10], and emphasize attentional access by prefrontal mechanisms [11]. Conversely, *Integrated Information Theory* starts out from phenomenology, characterizing conscious experience by several properties [6] (which IIT calls “axioms”, but without using them to formally derive all other claims of the theory). Some proponents of IIT maintain that it is anti-functionalist and cannot be simulated computationally, that is, consciousness cannot be recreated using a different underlying causal structure. (But arguably, physicalism is itself a functionalist stance; IIT may have to decide at some point between rejecting physicalism or accepting functionalism.) At the core of IIT stands the notion that consciousness is identical to a system having properties that integrate information (which IIT aims to characterize in ongoing work, using a parameter called Φ [5], which represents a functionalist measure of information integration). A major criticism of IIT is its failure to demonstrate necessary and sufficient conditions for the observable aspects of consciousness [12], which it captures phenomenologically so well. *Illusionism* appears to be a polar opposite of IIT, in rejecting and deconstructing the phenomenology of consciousness as illusory, eg. by pointing out that the experienced order of events is constructed after the fact and does not have to represent the actual order in which events were perceived. Illusionism might be nominatively overstating its case, since the perceived contents of consciousness are indeed represented in some form in the brain. Contents and observer are not real, but they are also not illusory: they are *virtual*. Consciousness might not be understood as a physical property [13], but a property of a *simulated* system (we could call this stance *“virtualism”*).

According to Graziano’s *Attention Schema Theory*, consciousness is a control model of attention itself, existing in parallel to the self model and the body schema. I agree that the notion of being the locus and agent of attention is central to the experience and functional characteristics of consciousness. Consciousness is a particular kind of representation, a ‘multi-media’ narrative of what is being attended to. However, I would subsume consciousness largely as part of the self model, and also point out that the control of attention does not have to be volitional, and at least in part lies outside of the conscious domain. We might describe the role of consciousness as similar to the conductor of an orchestra, in selecting out features of different, autonomously acting instruments, with the goal of harmonizing their behavior, and submitting it to globally coherent volitional control [14]. Bengio's notion of a *Consciousness Prior* is compatible with consciousness as attentional control (or as a conductor) and describes it as a low dimensional function that regularizes the model state to obtain a local minimum of its energy function (i.e., to achieve the most efficient, globally consistent, predictive representation of the agent and its environment). An important aspect of consciousness is however the generation of a stream of consciousness, which requires the creation of partial binding states of working memory, and a protocol memory that allows to retrieve and recall them later on, which Bengio's contribution does not discuss.

What all presented theories have in common are the central roles of consciousness within attention and integration of mental representations into a coherent interpretation of reality. They also agree in their rejection of conscious phenomenology as a causally irrelevant epiphenomenon, and in the impossibility of philosophical zombies [15] (agents that show all features of intelligent human behavior, while lacking conscious experience). However, the theories differ in the proposed ontology of consciousness, and the necessary and sufficient conditions that would characterize it. IIT takes consciousness to be identical to the integration of information, GWT, and the Consciousness Prior focus on the mechanism that would achieve the integration, while the Attention Schema and Illusionism point at the representations that result from the control of information integration. Existing notions of attention within AI, especially in the context of the *Transformer* algorithm [16] are very different from the integrated, centralized attention in GWT or Attention Schema Theory. Here, attention is set of distributed, local mechanism to identify statistically relevant features in representations that are to be singled out for learning. We can use the convergent aspects of these theories to derive a characterization of consciousness as dynamic representations of control models of attention, which allows us to ask the question whether large scale learning systems integrating over all sorts of data and modalities will require such control mechanisms to achieve coherence, and if these mechanisms can emerge over existing global reward maximization functions, or whether they will require specific priors built into AI systems.

References

1. Baars B. (1988). A cognitive theory of consciousness New York, NY: Cambridge University Press.
2. Baars B. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. Progress in Brain Research 150: 4–53.
3. Dehaene S, & Naccache L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. In Dahaene (ed.), The Cognitive Neuroscience of Consciousness (pp. 1–37). Cambridge, MA: MIT Press.
4. Tononi G. (2012). Integrated information theory of consciousness: an updated account. Arch. of Ital. Biol. 150(4): 293–329.
5. Tononi G., Sporns O. (2003). Measuring integrated information. BMC Neuroscience 4(31)
6. Oizumi M., Albantakis L., Tononi G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. Computational Biology 5(10): 1–25.
7. Graziano, M. S. A., Webb, T. W. (2014). A Mechanistic Theory of Consciousness. International Journal on Machine Consciousness.
8. Bengio, Y. (2017). The Consciousness Prior. arXiv:1709.08568.
9. Frankish, K. (2016). Illusionism as a Theory of Consciousness. Journal of Consciousness Studies 23 (11-12):11-39 (2016).
10. Dehaene, Stanislas (2015). Consciousness and the Brain. Viking.
11. Del Cul, A., Dehaene, S., Reyes, P., Bravo, E., Slachevsky, A. (2009): Causal role of prefrontal cortex in the threshold for access to consciousness. Brain (2009) 132 (9): 2531-2540.
12. Cerullo, A.M. (2015). The Problem with Phi: A Critique of Integrated Information Theory. PLoS Comput Biology; 11(9): e1004286.
13. Tegmark, M. (2014). Consciousness as a State of Matter. arxiv.org/abs/1401.1219.
14. Bach, J. (2018). The Cortical Conductor Theory: Towards Addressing Consciousness in AI Models BICA 2018.
15. Chalmers, D. (1996): The Conscious Mind, Oxford University Press, New York.
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2017). Attention is all you need. Neural Information Processing Systems, 2017.