Olivia Y. Lee
SYMSYS 202 – Final Paper

Consciousness, Phenomenality, and the Representational Layer

Many high-level theories of consciousness, such as first-order representationalism (Lycan, 2019), higher-order thought (HOT) theory (Rosenthal, 1997), and Global Workspace Theory (Dennett, 2001), involve the generation of sensory or cognitive representational states. However, the relations between phenomenal consciousness and representational states remain an open question. The term "representational layer" refers to the layer of abstraction at which representational states exist. The representational layer overlays our raw perceptions and direct sensory interactions with the world, and at this level of abstraction, mental states gain more complex representational content beyond the sensory states obtained from directly interfacing with the environment. This paper contends that metacognition on top of the representational layer beyond mere possession of representational states, as proposed by HOT and Global Workspace Theory, is critical for consciousness. Based on the modular nature of perceptual processing, this paper distinguishes between lower-level sensory states used only as intermediates in perceptual processing, and higher-level sensory states (outputs of perceptual processing) that are representational states upon which metacognitive processes occur, which is how phenomenality arises. Section I begins with a comparison of three high-level theories of consciousness and their hypotheses about the treatment (or lack thereof) of representations that give rise to consciousness. Section II continues with a discussion on HOT theory and Global Workspace Theory and investigates the question of how phenomenality arises in the causal order of representations. Finally, Section III concludes by considering the implications of the nature of the representational layer on the functions of consciousness and the downstream behaviors that arise from it.

I. Comparing High-Level Theories of Consciousness

This section analyzes distinctive treatments of representations by different high-level theories of consciousness. The modular view of Global Workspace Theory will be regularly referred to across all three theories for visualization and as a basis for comparison. The first distinction made here is between first-order representationalism and HOT / Global Workspace Theory in addressing what (if anything) is required beyond the existence of representations. First-order representationalism contends that simply possessing representations is sufficient for consciousness. Tye (1995) puts forward that phenomenal character of first-order representations is Poised (readily available as input to the creature's belief or desire system), Abstract, Nonconceptual (subjects need not have concepts enabling to have thoughts about all the features represented in the experience), Intentional Content. Interpreting this theory with the modular view of the Global Workspace Theory, first-order representationalism contends that the broadcast of any representations in the respective modules of the brain into the global workspace are sufficient for consciousness, regardless of whether those representations receive sufficient levels of feedback from other modules in the brain. On the other hand, both HOT and Global Workspace Theory contend that processes of metacognition *on top of* representations, beyond merely having representations, are necessary for consciousness. In HOT theory, metacognition is instantiated in the formation of higher-order thoughts about first-order representations, which makes those first-order states conscious. In Global Workspace Theory, metacognition on first-order representation

arises through the feedback loops between modules in the brain after those first-order representations are broadcast into the global workspace.

In weighing two views relating representational states to consciousness – merely possessing first-order representations versus performing some form of metacognition on first-order representations – one can compare human mental processing to that in animals or other living things. The canonical dog problem (Byrne, 1997) is often cited to question the plausibility of animals like dogs thinking about their own mental states, and similar discussions about the behavior of fish, crabs, and even plants have been raised (Andrews, 2014). To address these arguments, a key distinction needs to be made between using sensory states for downstream processing and *metacognition* on representational states, or the outputs of perceptual processing. The former has a direct role in the causal order that gives rise to downstream behavior, while the latter has a much more indirect role (see Section IV for more discussion on this). While a dog is certainly capable of responding to a ball thrown in the distance by running after it, it is much less clear whether it can interpret the "richness" of its experience, the affective quality of its chasing after the ball, and other qualitative aspects of experience that humans may be familiar with. The ability of an agent to take in sensory states as input and act or modify its behavior in response to those stimuli certainly is a form of processing of representational states (often referred to as associative or reinforcement learning), but perhaps not the sort of metacognitive processing we are interested in when discussing phenomenal consciousness.

Given that animals and other living things may possess and process representational states while the phenomenal consciousness of these agents remain questionable, this suggests that the first-order representationalist's view of consciousness – that merely possessing first-order representations is sufficient for consciousness – may not be tenable. The intuition persists that some form of higher-level processing is critical for consciousness, and such higher-level processing is instantiated differently in HOT theory from Global Workspace Theory. According to Global Workspace Theory, the richness or intensity of the interaction between modules, as measured by the degree of feedback between modules derived from a certain representation broadcast into the global workspace, determines whether a certain mental representation becomes conscious. This amount of feedback, or level of metacognition, could be quantified, similar to the *phi* value in Integrated Information Theory (Tononi, 2008). In contrast, one way of interpreting HOT theory through the lens of the modular view of Global Workspace Theory is the suggestion of an "attentional spotlight" in the brain that brings first-order representational states "into focus", enabling the agent to become conscious of the first-order state by developing higher-order thoughts about it. Alternatively, one could interpret HOT theory as being equivalent to having a dedicated attention module that samples from the representations broadcast onto the global workspace and decides what to make conscious and filters out relatively less important information.

The plausibility of an attention "spotlight" or proto-agent attention module that makes representational states conscious is debatable, at least seemingly less so than the view of feedback between modules in the brain put forward by Dennett (2001). At the risk of an overly literal interpretation of the attentional spotlight, it is worth questioning what (or who?) is operating this spotlight and deciding which representational states to bring into focus. A similar argument can be made for proto-agent attention modules in deciding from a sample of representational states which are sufficiently important or relevant. If the agent themselves is performing such decision-making,

the argument becomes inherently circular as evaluating a representational state before deciding if it should become conscious implicitly assumes being conscious of that representational state. However, the alternative of having a Cartesian theater-esque homunculus agent control this attention spotlight or module consequently runs up against traditional arguments against Cartesian dualist views of consciousness. Some proponents of HOT theory, grounded in the context of the global workspace, propose that attention is a property of the connections between modules, where the attention level of each connection is variable on a continuous spectrum. However, when considering the instantiation of such a property in a neurobiological or even artificial system, it is hard to formalize how this property is determined, specifically at which point on the attention spectrum each connection lies. On the other hand, the assertion that representational states become conscious when feedback between modules derived from that representational state exceeds a certain threshold is consistent with the current connectionist interpretations of neurobiology.

Overall, by distinguishing first-order representationalism from HOT theory or Global Workspace Theory, the intuition holds that metacognition on top of the representational layer, beyond mere possession of representational states, is crucial for representations to become conscious. Further comparison of HOT theory in comparison to Global Workspace Theory in how such metacognitive processes are instantiated reveals certain difficulties in viewing metacognition on representational states as higher-order thoughts that make the corresponding first-order representations conscious. In contrast, Global Workspace Theory determines the consciousness of a representational state output by a certain module based on the amount of feedback from other modules generated from that state, where feedback loops between modules of the brain (rather than higher-order thoughts) are the mechanism of metacognition. This proposal is not only consistent with current neurobiological theories, but also can be instantiated or realized in many other systems, both biological and artificial, which further strengthens its plausibility.

II. Phenomenality in the Representational Chain

With the notion that some level of metacognition on top of the representational layer is essential for consciousness, a natural next question to consider is at which stage phenomenality is introduced in the representational chain of metacognitive processes, whether that be in the chain of higher-order thoughts as in HOT theory or in the interconnected network of modules in Global Workspace Theory. A concrete example is when we see an object, for instance a red apple, in the world, we are conscious of the apple and the high-level features that are apparent to us like its redness; however, we are not conscious of its edges, shape from shading, and other lower-level visual features. It is hard to articulate (in a way that is different from the difficulty in articulating redness) what it is like to experience this edge or shape from shading at a particular location on the apple, and in some sense, we do not have representations for edges or shading in a way that we do for color and shape. It seems that in moving from lower-level features (like edges, shades) to higher-level features (like colors, shapes), phenomenality is introduced somewhere in moving up towards more complex visual features, resulting in only higher-level features being phenomenally represented. An open question, therefore, is at what point phenomenality is introduced in this spectrum of complexity, and how is it introduced. Why does our representational chain, and correspondingly phenomenality, stop at high-level visual features, and not at lower-level features?

According to HOT theory, a first-order mental state becomes conscious when one has a HOT about it. With this view, having a HOT about the perceptual representation of an apple thus allows us to be conscious of seeing the apple, and undergo a phenomenal experience corresponding to seeing the apple. However, phenomenality stops at that level of the perceptual representation of the apple, and not its lower-level visual features, such as its edges or shading. This may seem problematic for HOT theory, for we must be having *some* thought about the apple's edges, shape from shading, and other lower-level features that somehow combine to give rise to the overall high-level visual experience of the apple. In other words, we should have first-order representation of these lower-level visual features that we then become conscious of by having HOTs about them. However, this line of reasoning is fallacious for reasons similar to the distinction between mental processes in animals or plants and human beings: there is a key difference between the use of sensory states (or visual features) for downstream processing – such as compositional processing which combines lower-level visual features to form a higher-level visual experience – and metacognition on sensory states that are outputs of perceptual processing. By the argument of the HOT theorist, the absence of phenomenality at the level of lower-level visual features can be explained by the modular nature of perceptual processing. It is not possible to have HOTs (whether they are causal effects of lower-level mental states) or other metacognitive processes about intermediate steps (e.g., how an object's edges, shapes from shading, etc. combine to give rise to the visual percept of the object), but only about the outputs of the modules, i.e., the visual percept of the object itself. The outputs of the perceptual modules hence determine the level of conscious representation. While we may process the low-level visual features of an object, we do not form representations of them as such features are intermediates in the process of perceiving the object.

Like HOT theory, Global Workspace Theory relies on a similar notion of modular perceptual processing. Each of the modules in the brain (e.g., visual processing, audio processing, memory retrieval, etc.) process sensory or perceptual states before broadcasting the output to the global workspace. The contents of the states being processed within the modules do not have the potential to become conscious until they are broadcast into the global workspace and can generate feedback loops from other modules. Applying this view to the experience seeing a red apple, the low-level visual features like edges and shading processed within the visual perception module are structurally combined in a way that when broadcast into the global workspace, the composed percept of the apple is a representation that has the potential to receive feedback from other modules. For instance, the memory retrieval module may allow an agent to identify the object as an apple, or a memory associated with apple picking, for example. Additionally, the processing of low-level features need not be entirely contained within a single perceptual module and could be jointly processed by multiple perceptual modules. This could explain cross-modal illusions such as the McGurk effect (McGurk & MacDonald, 1976) or the double flash illusion (Shams et al., 2002). In such cases, multimodal processing involves "backchanneling" or passing of low-level percepts between the visual and audition processing modules, whose intermediate sensory states are jointly processed. However, these low-level features are not representationally rich enough to elicit feedback loops from other modules, hence the intermediate mental states in the multimodal processing leading to such illusions are not conscious. Rather, the combined representational state that is broadcast to the global workspace as the output of multimodal processing, whether that is the experience of hearing a certain sound given a certain visual (as in the McGurk effect) or seeing a certain number of flashes given a certain sound (as in the double flash illusion), is conscious due to the feedback loops that are generated from the broadcast of this representational state.

Overall, both HOT and Global Workspace Theory rely on the notion of modular perceptual processing, which allows us to distinguish between lower-level sensory states that are only used for downstream processing, and higher-level sensory states that possess representational content and are the outputs of perceptual processing. Higher-level representational states that are outputs from perceptual processing modules allows for metacognition on these representation states, whether through the formation of HOTs or generating feedback loops between modules in the brain. This enables such representations to become conscious or have an associated phenomenality.

III. Implications for the Functions of Consciousness

With the notions of the representational layer and phenomenality of higher-level mental states in mind, it is worth contemplating the implications for the functions of consciousness. A key question is to understand where phenomenality lies in the causal order of downstream behavior, which may shed light on whether consciousness is evolutionarily advantageous and naturally selected for. On one hand, there is the argument that consciousness has a function through which it enhances evolutionary fitness (Seth, 2009). Consciousness is fundamentally linked to downstream behaviors; the phenomenal quality of pain is a significant motivator of pain-avoiding behaviors, for instance, and such behaviors may be naturally selected for. On the other hand, some argue that consciousness itself has no function, but is a byproduct of other observable properties of the brain that do have an evolutionary function (Robinson et al., 2015). The former view suggests that consciousness motivates the selection of evolutionarily advantageous actions, while the latter view suggests that consciousness merely correlates with neural activity that was naturally selected for. Both consciousness and evolution are complex processes and the answer to this question is unclear. Nonetheless, this section seeks to explore what earlier discussions about the representational layer suggest about the emergence of consciousness, as well as some of the downstream effects of consciousness that may suggest it has evolutionary functions.

One evolutionarily advantageous behavior is the ability of an agent to imagine states that are not true of the present but, through a particular selection of actions, have the potential to be true in the future. An agent's predictive ability to imagine possible scenarios allows it to plan for potential future states and adapt its current behavior to move towards desired goals. This ability is highly nontrivial and is premised on the agent's ability to create a unified and coherent representation of past information about the world that it has previously processed (Crick & Koch, 1998). Such an ability has not only been observed to be beneficial in biological agents, but also artificial agents. For instance, self-supervised reinforcement learning agents developed with the Plan2Explore algorithm (Sekar et al., 2020) demonstrated generalized task adaptation due to their ability to leverage a world model to predict future model states in latent space, the computational equivalent of a representational layer. Therefore, a representational layer containing learned sensory and cognitive representations, upon which prediction or imagination of future states is based, is crucial for developing of this evolutionarily beneficial predictive ability.

As discussed in Section I, metacognition (including prediction) on top of the representational layer seems to be key to the emergence of consciousness. Based on this line of reasoning, consciousness seems to be correlated with neural processes, like imagination and prediction, that are evolutionarily advantageous. The representational layer, a layer of abstraction

above direct sensory perception, creates a mediation between the agent and the real world. This allows agents to imagine alternative possibilities generated from the current state of the world, reason about errors in its perceptual systems, and plan to take actions that align with the intended trajectories, without thinking that those possibilities are indeed happening in the present. Hence, consciousness is often associated with flexibility and adaptation, as the same metacognitive processes on top of representational states from which consciousness arises also facilitate agents choosing actions that are evolutionarily advantageous to adapt to their changing environments. Therefore, the emergence of consciousness from metacognition on the representational layer, which enables evolutionarily advantageous neural processes like imagination and prediction, suggests that consciousness is a byproduct of neural processes that were naturally selected for.

While this may be said, a crucial question pertains to the motivations behind an agent's decisions about which predicted goals or future states they should pursue, which consequently affects the actions they choose to take. It would be counterintuitive to deny the role of phenomenality in such decision-making, as it is a key component in how we decide which experiences are pleasurable and worth seeking or which are painful and worth avoiding. Therefore, while the argument above suggests that consciousness is a byproduct of neural processes that facilitate the selection of evolutionarily advantageous behaviors, it seems that the selection of those behaviors is, at least in part, dictated by corresponding phenomenal experiences that determine which outcomes are desirable. As a concrete example, one would withdraw one's hand if touching a hot stove because it is painful; it seems somewhat backward to say that the feeling of pain emerges from the neural processes that result in the withdrawal of one's hand. It can be argued that the hand is often withdrawn *before* one become conscious of the felt quality of pain, so it was the somatic reflexive action of one's nervous system that kicked in leading to the withdrawal of one's hand, and the qualitative experience of pain is a byproduct. Nonetheless, it is difficult to disregard the impact that felt quality of pain has on one's future behaviors in being especially cautious around hot stoves. Based on the argument that phenomenal experience is fundamentally dispositional in nature (Dennett, 1998), where the felt quality of a state boils down to its dispositional properties to cause further effects in the brain, it is difficult to divorce the phenomenal quality of an experience from the downstream behaviors that are induced. Consciousness and the phenomenal quality of experiences therefore seem to serve an important function, namely in determining which states are desirable, that are crucial in eliciting behaviors that are evolutionarily advantageous.

In conclusion, this paper contends that metacognition on top of the representational layer, the representational states that overlays our direct perceptions of world, is critical for consciousness. Furthermore, the modular nature of perception results in phenomenality arising from metacognition on higher-level representations output from perceptual processing. Together, these assertions imply that consciousness both emerges as a byproduct of neural activity that leads to selection of evolutionarily advantageous actions, as well as motivates said adaptive behavior. Given the complex history of consciousness and evolution, this may suggest a dynamic feedback loop in which consciousness facilitates evolutionarily advantageous behavior and vice versa. Therefore, understanding the evolutionary adaptive value of consciousness ultimately boils down to a deeper understanding of the relationships between brain, behavior, and consciousness.

Bibiliography

Andrews, K. (2014), *The Animal Mind.* ch. 4, “Consciousness”

Byrne, A. (1997). ‘Some like it HOT: consciousness and higher-order thoughts,’ *Philosophical Studies*, 86: 103–129.

Crick, F., and Koch, C. (1998). Consciousness and neuroscience. *Cereb. Cortex.* 8, 97–107.

Lycan, W. (2019). "Representational Theories of Consciousness", *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition).

Dennett, D. C. (1988). “Quining qualia”. In Anthony J. Marcel & E. Bisiach (eds.), *Consciousness in Contemporary Science*. Oxford University Press.

Dennett, D. C. (2001). Are we explaining consciousness yet? *Cognition*, 79 (2001), pp. 221-237

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. Nature 264:746–748. doi: 10.1038/264746a0.

Sekar, R., Rybkin, O., Daniilidis, K., Abbeel, P., Hafner, D., Pathak, D. (2020). *Planning to Explore via Self-Supervised World Models.* ICML 2020. arXiv:2005.05960

Robinson, Z., Maley, C. J., and Piccinini, G. (2015). Is consciousness a spandrel?. *J. Am. Philos. Assoc.* 1, 365–383. doi: 10.1017/apa.2014.10

Rosenthal, D. M. (1997). A Theory of Consciousness. In Ned Block, Owen J. Flanagan & Guven Guzeldere (eds.), *The Nature of Consciousness*. MIT Press.

Seth, A. K. (2009). “Functions of consciousness,” in *Encyclopedia of Consciousness*, ed W. P. Banks (Amsterdam: Elsevier/Academic Press), 279–293.

Shams L., Kamitani Y., & Shimojo S. (2000). What you see is what you hear. Nature 408:788–788. doi: 10.1038/35048669.

Tononi, G. (2008). “Consciousness as integrated information: a provisional manifesto.” Biol Bull. 215: 216–242.

Tye, M. (1995). *Ten problems of consciousness: A representational theory of the phenomenal mind*. The MIT Press.