Jessica Ng
18 November 2022
Department of Cognitive Science, UC Berkeley

The Global Workspace Theory and its Implications for Machine Consciousness

Part I:

In Chapter 4 of *A Cognitive Theory of Consciousness* (1998), Barnard J. Baars discusses the purpose of consciousness in the human brain. In the introduction, he first defines conscious events as anything that one can clearly report with detail and accuracy. Because we can define conscious brain events, we can conduct numerous experiments regarding the reasons for why unconscious and conscious stimuli activate particular parts of the brain and not others, as well as explore the effects of brain damage on waking consciousness. To conduct conscious cognition experiments such as these, it is necessary to consider consciousness a controlled variable.

Baars then introduces the Global Workspace Theory (GWT): a cognitive architecture that explicitly details the purpose of consciousness. Essentially, the brain is considered an enormous set of specialized processors. Examples of such processors are the sensory system and hippocampal system. The GWT maintains that our consciousness is what allows for global access between specific brain operations that otherwise would function separately. Essentially, conscious cognition acts as a central gateway for information to be distributed and exchanged. Consciousness can help coordinate, integrate, and control the entire brain system. As a result, specialized networks are capable of broadcasting information to the entire system.

The theater metaphor provides a visual representation of the GWT. Conscious experience is compared to the bright spot on the floor of the stage of a theater. The stage resembles immediate and working memory. Everything else in the theater is dark and unconscious. Attention is compared to the spotlight which decides where the spot lands on the stage. When a specific brain system is activated in response to stimuli, we become conscious of this content. It is then accessible to various other networks in the brain, which are resembled by the darkened audience in the theater that can perceive the spotlight of consciousness.

It is important to distinguish between the two different methods of control of our selective attention system (the “spotlight” in the theater metaphor that directs our conscious experience). Voluntary attention is enabled by the frontal executive cortex. Involuntary automatic attention is determined by a variety of areas such as the pain systems, insular cortex, brain stem, and emotional centers. These special regions determine which stimuli are significant enough to break through the attention barrier and become a part of our conscious experience, and which are discarded.

As evidence for the GWT, we can examine sensory consciousness. The ventral visual stream of the visual cortex identifies visual features that become conscious. For instance, when we look at another person’s face, the ventral visual stream is activated. However, in order to remember someone’s face, our brains must consult the hippocampal system. To emotionally respond to viewing the face of someone whom we have not seen in a long time, the amygdala has to be activated. Because we can both recollect a human face and respond emotionally to it, the ventral visual stream must influence regions that are not needed to support conscious contents directly. Hence, this demonstrates that our consciousness allows for our brain to share information across multiple specialized processors, which in this case are the hippocampus, amygdala, and visual cortex.

Baars also draws upon an electroencephalogram study conducted by John et al. (2000). When consciousness is lost, gamma power decreases and the lower frequency band power increases. In addition, they witnessed a decrease in coherence between the homologous areas of the two brain hemispheres, as well as a drop in coherence between anterior and posterior hemisphere regions. This remains consistent with the GWT notion that integration of multiple brain functions and networks is not possible without consciousness. Finally, Baars cites a study by Portas et al. (2002) that further reinforces the GWT. The study found that only the primary auditory cortex, not the entirety of the auditory cortex, is activated in response to an auditory stimulus when a person is in deep sleep. Likewise, a study by Laureys et al. (2000, 2002) found that after a serious brain injury, normally loud or painful sounds only activate the primary sensory cortices of a patient in a vegetative state. These studies provide evidence for the idea that consciousness is required to mobilize widespread areas of the cortex.

Many regions of the cortex work together to produce an inner dialogue, imagery, and emotional facial recognition. To produce outer speech, our brain must activate the left hemisphere. Interestingly enough, the same cortical areas involved in outer speech are also involved in producing inner speech. Similarly, the visual cortex is involved in the production of mental imagery. Also, somatosensory imagery can generate feelings of fear, pain, pleasure, and more. This is because imagining internal sensations can reflect emotional processes by interacting with other parts of the brain through global distribution. In order to recognize a fearful facial expression versus an angry facial expression, for example, the brain has to consult the amygdala. Therefore, it appears that many cortical regions work together to process conscious experiences, all thanks to the presence of a consciousness that integrates these networks.

The Global Workspace Theory also outlines the idea of “contexts”. In the frame of reference of the GWT, a context is an unconscious brain event that shapes conscious activity. For instance, imagine you disembark a boat that has been sailing on choppy waters. As you step onto solid land, you may stagger or sway because the horizon seems as though it is wobbling. This is an example of some sort of unconscious activity affecting your visual perception and conscious experience. Another example of the use of contexts would be flying on an airplane during the nighttime. Even though it is dark outside and the passengers have no knowledge of the direction of the plane, they are able to see the cabin tilting as the plane lands. The GWT describes that in order to produce a conscious event, there must exist an interaction between sensory contents and contextual systems.

Conscious sensory input caused more intense and widely distributed brain activity, especially in the parietal and prefrontal cortex, than identical unconscious input. Therefore, consciousness seems to be necessary for widespread activation of large areas of the brain. Baars comes to the conclusion that the role of consciousness is to enable access between multiple specialized brain networks. Essentially, consciousness functions as a workspace that facilitates coordination and control of the entire brain system.

Part II:

Thirty years after Baars introduced the Global Workspace Theory to the scientific world, researchers are attempting to apply this cognitive architecture to other fields of study. In an article titled *Deep learning and the Global Workspace Theory*, Rufin VanRullen and Ryota Kanai propose the creation of a global workspace in an artificially intelligent machine in order to expand the capabilities of AI in the future.

The authors of this article first define deep learning, which is a machine-learning technique in which artificial neural networks “learn from” or are “trained” by complex algorithms and large volumes of data. These synthetic networks have many hidden layers between the input and output layer of a deep

learning system. So far, deep learning has enabled computers to emulate cognitive functions and perform tasks in ways never seen before. VanRullen and Kanai intend to build upon the deep learning framework by taking into consideration the GWT and the neuroscientific version, the Global Neuronal Workspace (GNW). The GNW states that we experience consciousness when multiple brain stems have access to incoming information. This global broadcast of information is made possible by a neuronal network of axons that are distributed in the parietotemporal, prefrontal, and cingulate cortices.

Next, the authors identify the components necessary to implement a global workspace in an artificial intelligence system. They note that this suggested "roadmap" to a deep-learning global workspace will likely need revising, as their roadmap is more of a general theoretical proposal rather than a precise process.

The first step in this theoretical roadmap is to integrate multiple specialized modules. These modules could be neural networks that were pre-trained for sensory perception, natural language processing, motor control, long-term memory storage, or any functions along those lines. Each module would have its own latent space, which is a layer that internally encodes and represents data that was externally observed. Connecting these specialized modules could be considered a workspace because multiple networks have access to a singular input.

The second step would be the implementation of a global latent workspace (GLW), which is an intermediate latent space that is shared between specialized modules that can perform neural translation unsupervised. The GLW would be trained, with the objective of obtaining cycle consistency, to effectively translate between the latent space representations of two modules. The goal of cycle consistency training is for the GLW to transcribe between a pair of modules (even when matched data is not available), and then translate once again to return the original input.

Additionally, a global latent workspace requires some sort of attention system. This can be achieved through a key-query matching process. One network layer emits a query, and another layer produces a matching key. A match means that the input is brought into the GLW; a mismatch means the information is discarded. This is similar to how, from a cognitive scientist's perspective, attention determines what information passes into our consciousness.

Once a module connects to the GLW as a result of the key-query matching attention system, its specific latent space gets copied into the workspace. This internal copy functions as a bidirectional connective interface between GLW and the matching module. These internal copies are now accessible to all other modules in the global environment.

One major advantage of a GLW architecture is that the deep-learning machine can combine multiple modules to execute completely novel tasks. The entire system has the ability to perform more general functions from the abilities of specialized modules. This is all due to transfer learning, which is a term for when one module's abilities are mobilized and deployed onto another module's representational latent space. As a result, the deep-learning machine is incredibly flexible; it can generalize previously learned modules and adapt to novel environments. Furthermore, this adaptable mental composition system may be able to produce counterfactual reasoning. The deep-learning network could ask "what if" questions regarding the environment's response to its own actions, which then can be used to predict possibilities for future events and states of the world. These combined benefits of a GLW could be a major stride towards strong artificial intelligence since functions such as imagination, planning, and reasoning about possible future states are at the core of high-level cognition.

At the end of the article, Rufin VanRullen and Ryota Kanai acknowledge that their proposed "roadmap" towards the GLW is more of a general theoretical framework rather than a concrete

step-by-step process for true implementation. Actual implementation demands quite a lot of trial and error as well as considerable computational resources. The article was mainly written to prove that taking inspiration from cognitive architectures and applying them to a deep-learning context could give rise to significant functional advantages.

Part III:

Bernard Baars' original Global Workspace Theory explains that consciousness essentially provides a workspace that connects many specialized brain functions, allowing them to interact to communicate when they would otherwise operate independently. Rufin VanRullen and Ryota Kanai propose the implementation of a Global Latent Workspace in a deep-learning system that mirrors Baars' theory of consciousness. Naturally, this brings up the following question: if we successfully equip an artificial network with a GLW, would it be considered conscious?

First and foremost, it is essential to take a deeper look into the definition of consciousness. In an article titled *On a confusion about a function of consciousness*, Ned Block distinguishes between two different aspects of consciousness (Block, 1995, p. 229). He defines access conscious content as any information that is broadcast in the global workspace. Access consciousness is therefore functional; the information in the workspace can be used to guide reasoning and rational thinking, which controls actions and speech. However, Block reasoned that there must be another state of consciousness working together with access consciousness. When we see a complicated visual scene, we experience an abundance of content, much of which we cannot report. Block's idea of a phenomenal consciousness covers the subjective experience of perceptions, thoughts, desires, emotions, sensations, and more. Overall, he claims that there exists mental information that can be experienced, but not accessed.

In Baars' introduction, he mentions that he regards a conscious event as anything that a person can report with accuracy and detail. Thus, his global workspace model completely neglects the idea of the phenomenal consciousness. Despite this, an artificially intelligent system equipped with a global latent workspace does seem to meet the criteria for access consciousness. The internal copies of each module's specific latent space could be considered access conscious contents, since this information is being shared across an entire system. Both human brains and the proposed AI machine have global access to broadcasted information thanks to a global workspace. Thus, it seems as though VanRullen and Kanai's GLW network could be considered partly conscious according to Block's definition of access consciousness, as content information is being distributed across a workspace.

VanRullen and Kanai also highlighted the potential high-level functions of a GLW, one of which is counterfactual reasoning that could lead to imagining future states of the world and planning reactions to these states. Rational thinking that informs speech and action is an indication of access consciousness, according to Block. Thus, this provides more evidence for the notion that a network with a GLW could be considered at least partially conscious.

However, whether phenomenal consciousness exists in VanRullen and Kanai's proposed deep-learning system is an entirely different question. Phenomenally conscious content is information that is not reported, but experienced. This type of content does not seem to exist within a machine with a GLW. Modular information (in the form of a representational latent space) is either made into an internal copy, depending on whether it passes the attention barrier, or not. Internal copies are globally accessible to other modules, whereas the latent content that was discarded is not. Either the unsupervised neural connection is created, or not. There seems to be no way in which a network can do the machine-equivalent of "experience" a sensation but not neurally translate it onto its workspace.

Phenomenal conscious contents in the human brain can be described as non-functional but experienced, yet GLW contents are either functional or completely discarded. Therefore, phenomenal consciousness will not be achieved by an artificial network equipped with a GLW.

Additionally, it is difficult to come to a concrete conclusion about whether future AI technologies will be considered conscious because the philosophy surrounding consciousness is constantly changing. Ned Block's idea of access versus phenomenal consciousness is sometimes criticized, and a few have suggested modifications that they deem necessary. Other scholars have proposed entirely different definitions for consciousness. For instance, in the textbook *Robotics, AI, and Humanity*, there is a chapter titled *What Is Consciousness, and Could Machines Have it?* Dehaene and colleagues (2021) suggest that there exists a self-monitoring aspect that constitutes the second dimension of conscious computation besides global availability (p. 44). Humans can acquire and process information about themselves. We know the position of our limbs at any given time, we know whether we made an error, and we can figure out if we know something or not. This is also called introspection or meta-cognition, according to psychologists. Notably, a deep-learning machine that possesses a global workspace lacks a self-monitoring system, meaning it would not be considered fully conscious. Therefore, it seems as though a GLW AI system would most likely be considered only partially conscious, if a GLW were to ever be implemented into a deep-learning system.

Reference list

Baars, B. J. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Progress in brain research*, *150*, 45-53.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and brain sciences*, *18*(2), 227-247.

Dehaene, S., Lau, H., & Kouider, S. (2021). What is consciousness, and could machines have it?. *Robotics, AI, and Humanity*, 43-56.

VanRullen, R., & Kanai, R. (2021). Deep learning and the global workspace theory. *Trends in Neurosciences*, *44*(9), 692-704.