Műszaki Tudományos Közlemények vol. 11. (2019) 175–178.
https://doi.org/10.33894/mtk-2019.11.39
Hungarian: https://doi.org/10.33895/mtk-2019.11.39
https://eda.eme.ro/handle/10598/31236

THEORETICAL FRAMEWORK OF ARTIFICIAL CONSCIOUSNESS

Ferenc SZIGETI, jr.

Budapest University of Technology and Economics, Faculty of Mechanical Engineering, Budapest, Hungary, szigetif.31@gmail.com

Abstract

Human consciousness is our most perplexing quality, still, an adequate description of it's workings have not yet appeared. One of the most promising ways to solve this issue is to model consciousness with artificial intelligence (AI). This paper makes an attempt to do that on a theoretical level with the methods of philosophy. First I will review the relevant papers concerning human consciousness. Then considering the state of the art of AI, I will arrive at a model of artificial consciousness.

Keywords: *artificial intelligence, philosophy, consciousness, cognitive science.*

1. The problem of consciousness

Currently, functionalism is the most widely accepted and adapted theory of mind in the philosophy of mind. It is also compatible with the scientific approach as a physicalist theory. It's main idea is that mental functions or states are constituted solely by their functional roles and can be implemented in any physical system regardless of their origin (e.g. biological or artificial). The only weakness of the theory is that it cannot explain phenomenal consciousness. With that in mind, we will not deal with problems concerning phenomenal consciousness[1].

There are many possible formulations concerning the function of consciousness, but there is no consensus on their validity. One can approach this formulation from thinking, creativity, problem solving, attention, defining goals or following them. These concepts are always strongly correlated with consciousness, but still in their very nature they are as elusive as the concept of consciousness itself. A definition based on such a concept will typically just specify a relation to a cognitive module, which usually is likewise hard to define or describe. Neither of them can point to a certain physically well definable or describable phenomenon, based on which an artificial model can be implemented.

Much work has been done on the human brain and mind. One can say that we are not far from the idea of engineering a brain from neuron to neuron. However it seems that this solution would be immensely difficult. The replication of such complexity is almost unthinkable. This is precisely the reason why we need simplified models of consciousness. Such models which can reduce complexity but in doing so, keep the essential functions of consciousness intact.

2. What are the functions of consciousness?

In cognitive science, besides the mapping of the mind, consciousness is also studied. Based on the experimental results we can infer to the functions of consciousness. These findings are summarized and translated into functions in the global workspace theory [1]:

- the intentional control of action;
- durable and explicit information maintenance;
- the ability to plan new tasks through combining mental operations in novel ways;
- control of attention.

The model of consciousness presented in this paper will focus on these four functions. In the next section we will briefly review some of the

[1] It is the „hard problem of consciousness". There is no satisfactory explanation in physicalism, why is there a certain subjective experience to every kind of perceptual or thinking process.

cognitive literature, so that we can have a better understanding on the workings of the model.

3. Review of the neural processes relating to consciousness

In this section every paragraph summarizes a cognitive model. Some elements of each presented model will be used in constructing our own.

Global neuronal workspace theory is an accurate model of human consciousness [2]. It describes consciousness as the global availability and amplification of information. Long-distance reciprocal connections of pyramidal neurons[2] enable this kind of connectivity between certain cognitive modules. These modules are responsible for higher cognitive functions and in such connected manner they constitute the workspace. Consciousness processing is realized through global availability of information in the workspace.

The mechanisms of attention can be understood through four main processes [3]. In the working memory the selected information is processed but only temporarly. It has limited capacity (3-4 items at a time), so it cannot operate on every piece of incoming information. Competitive selection determines which information channel will be able to access the working memory. Top-down sensitivity control regulates the relative signal strength of information channels. This is important, because comperitive selection bases its decisions upon the relative signal strength of information channels. Bottom-up salience filters automatically enhance the signal strength of information channels with infrequent stimuli or stimuli of learned importance (e.g. we are inclined to notice a red flower in green grass). According to the model, the first three process (working memory, competitive selection, top-down sensitivity control) form a loop, which allows intentional and conscious control of attention.

The working memory can be described through four processes: attention, perceptual representations (stimuli), long term memory representations (memories) and prospection. It is not a single cognitive module, but a distributed system, always organised to attend to some selected information. Limited[3] in capacity (3-4 items). Operates through activation of the long term memory [4].

Though there is no common consent based on the experimental results, it is safe to say that animals[4] also possess a working memory. It seems that the underlying mechanisms are similar, but operate in a more limited fashion in their case (e.g. less capacity, easily distractable attention, applicable only for real time tasks) [5]. According to the attention schema theory, consciousness or intentional control of attention is governed by a simplified attentional model in the mind. There is a simplified model of body in our minds to help with the intergration of various perceptual information into a concrete understanding of the surroundings and the body in it. It is only appropriate to think that there is one to help with the control of attention. It offers also an explanation of the nature of perceptual awareness. The mind concludes, based on this simplified schema, that it has a certain ability which can govern attention without any physical process (this simplified schema is of course free of understanding any underlying physical process, that is the simplification in it). It is aware of things, but cannot grasp how it is possible [6].

Review of the literature concluded, we can move on now with a better understanding of the theories of some highly related cognitive abilities.

4. Simplified model of human consciousness

In this section the results of the ongoing research will be discussed. It is a cognitive model, based on which a neural model and then an implementational model of artificial intelligence is possible. In the following subsections the workings of all four functions of consciousness will be explained in the frames of the model.

4.1. How does the human brain work?

The incoming stimuli is processed by a system of neurons with the energy provided by the body. Through the processing of information some action will be implemented. Depending on how much they work, neurons get some energy. They want more energy so they always want to work more, it means they prefer more active connections and with it always more chance to process information.

4.2. How does voluntary control of attention works?

Attention as discussed above, is not a module, but a process or phenomenon. To put it simply,

[2] This kind of neuron is equipped with a long axon and a profusely branching dendrite. It enables long distance connectivity between brain areas.

[3] One popular idea is that the limits arise from difficulty in keeping multiple active representations segregated from one another in neural activity with minimal interference [4].

[4] There are of course huge differences between animals. The experiments were conducted on doplhins, seals, primates, rodents and corvids.

we can interpret it as the activity of neurons. Attention is focused to the area of brain which is the most active. Many brain areas can be active simultaneously, but only the most active few can access the working memory.

Working memory is no more than the groups of neurons currently carrying the information which is about to be processed. It seems that the definitions of working memory and attention are circular. It is true in a way, as we will see, our goal is establishing some kind of a loop. So we will treat working memory and attention as the same process. In addition to that, as long term memory activated is the working memory according to some, it is also equal to the other two processes.

Basically all three of them are activated neurons. It is no great surprise as everything in the mind is neurons, but the idea is that we shouldn't treat them as different systems. We have achieved here already some simplification which was the whole prupose. The question though is not yet answered, so we should move on.

Attention is always directed towards something, it can be a stimuli (internal or external) or a memory. The first is some current perceptual information, the second was once a stimuli, but it was processed many times and kept around for future use. The voluntary control of attention is a loop of two or more pieces of information (some of them should be memories). Always one of them is in the center of attention (its activity is higher than of any other information). The loop is successful when the chain of information can reliably amplify itself and achieve highest activation for every piece, one after another.

Presumably, this amplification process is possible when the corresponding groups of neurons are physically connected, so that electric current can run through the loop. Following this idea, those areas of brain which possess a better connectivity to each other may be better in controlling attention. These areas are precisely the ones known collectively as workspace, modules of higher cognitive functions.

With this idea, changing the focus of attention is changing part or all of the information chain to another. In this way we can run through our memories or think through something.

As stated earlier not one, but many such loops independently operate with various levels of activation. It is only logical to infer that when two or more loops have some part in common, there is a better chance that this part will reach the required activation level.

Of course this loop of amplification can be overridden at any time by the influence of bottom-up salient filters. It enhances the activation of some stimuli and so attention is directed towards that stimuli. We are easily distractable, it is really hard to achieve deep concentration and only with a surrounding absent salient stimuli.

A question might arise from all this: if higher animals (e. g. primates) possess similar working memory, what is the difference between us and them? Up to this point there are only quantitative differences according to the model. Their brain might be less complex, they are more easily distracted or fewer of their brain areas can be active simultaneously.

Apart from complexity there is another distinct ability to the human brain. The idea comes from the attention schema theory. This schema is also a kind of memory or at least we can interpret it so. According to this theory voluntary control of attention is enhanced with the use of the attention schema. We can incorporate this into our model. The schema is organised along the pyramidal neurons, it is a system of memories which contains strong connections between certain cognitive modules (workspace). It serves as a kind of motorway for attention, with it, we can reach reliably any part of our workspace.

As the attention schema theory also contains the notion of awareness, it is appropriate to propose its place in this model too. Awareness is the subjective experience which is associated with the most activated information. Many loops are active at once, but only a few are active enough to be in our awareness. This is the simplification of the schema, it is its purpose to keep control over the limited capacities of the body and with it, the mind.

We should not imagine this attention schema as something which is „above" all brain processes or the „real cause" of conscious processing. It is a system, which enables effective control over limited resources. It means, that it controls the activation of neurons, so that activation would be high where needed. Required energy will flow to brain areas which need it most, information processing would commence where it should. Incoming stimuli will only disturb the concentrated information processing when it is of crucial importance to survival. This system enables higher thinking, only by effectively controlling information processing.

How could such a schema form? By chance?

It is plain that the schema is beneficial for survival, so evolutionary origin seems to be a logical conclusion. In other words, it was formed by coincidence, but it is no coincidence that it stayed.

A possible background of its origin is offered within the framework of the model. The brain's basic processing is pattern recognition/discrimination. It compares a stimuli to a memory (past stimuli). By chance or by error, memories might be compared to each other in some cases. The results of these deficient comparsions have been stored, if they occured frequently enough. With time more and more such memory should enable supervised learning.

In certain situtations decision making is enhanced with supervised learning. In this case, instead of pure reaction, we have design. The agent will be able to simulate, plan actions based on former similar situations. The agent will not remain in the present moment, always slave to the incoming stimuli.

4.3. Voluntary control over actions

The model is complete, from this point on, we need only to apply it to the three other functions.

Controlling of actions is possible through certain group of neurons, which then send the signal to the actuators. The workspace is connected with these areas, so the activated loop can also connect, and activate them.

4.4. Durable and explicit information maintenance

Connections between neurons are reinforced when used more. The model enables persistent usage of connections with loops. Memories can be formed with this process.

The mechanisms of forgetting can be explained by the behavior of individual neurons. They want more energy, so they prefer connections with more activity. They weaken or cut rarely used connections, so memories which were connected to it or constituted by it, will become inaccessible or will fade.

4.5. Planning new actions by combining mental resources in novel ways

This is creativity and higher thinking. According to our model, any memory can be accessible in the case of proper connectivity to the workspace. When a memory can be accessed, then it can be combined with another. Better connectivity means better attentional focus, because with time the amplification fades, when it travels for less time it can be reinforced more often. Of course for thinking, loops of memories are required because they contain the mental resources.

5. Model of artificial consciousness

In conclusion, we will briefly summarize the model. Consciousness processing is a kind of loop, which can control itself with some limitations.

There are many loops active at a time and they compete to be more active, because then the constituent neurons receive more energy, which is their evolutionary goal.

Loops of memories enable the formulation of more complex memories which are the basis for higher cognitive functions and processes. Some of these functions are already encoded in our DNA, one such example might be the attention schema. This schema enables the effective control of attention.

The proposed model should be realizable without any difficulty. However, first it should be translated into a neural model and then into an implementation model. Through this process, it can be formulated from a theoretical to a practical model of consciousness. These are the future goals of this research.

Acknowledgementes

This work was supported by the Hungarian Ministry of Human Capacities with scholarship within the New National Excellence Program (Új Nemzeti Kiválóság Program). I thank Á. Gyarmathy and M.Héder for their helpful critisism of the earlier drafts.

References

[1] Dehaene S., Naccache L.: *Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework*. Cognition 79/1-2. (2001) 1–37.

[2] Dehaene S., Changeux J. P., Naccache L.: *The global neuronal workspace model of conscious access: from neuronal architectures to clinical applications.* Characterizing consciousness: From cognition to the clinic?. Springer, Berlin, Heidelberg, 2011. 55–84.

[3] Knudsen E. I.: *Fundamental components of attention.* Annu. Rev. Neurosci. 30 (2007): 57–78.

[4] Eriksson J., et al.: *Neurocognitive architecture of working memory*. Neuron 88.1 (2015): 33–46.

[5] Carruthers P.: *Evolution of working memory*. Proceedings of the National Academy of Sciences 110. Supplement 2 (2013): 10371–10378.

[6] Graziano M. S. A., Webb T. W.: *The attention schema theory: a mechanistic account of subjective awareness*. Frontiers in psychology 6. (2015) 500.