Open Peer Review on Qeios

Towards Modeling Artificial Consciousness

Maksym Vakulenko

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

A new synergetic approach to consciousness modeling is proposed, which takes into account recent advancements in neuroscience, information technologies, and philosophy.

M. O. Vakulenko

Institute of Problems of Artificial Intelligence, Kyjiv (Kiev), Ukraine

Key words: artificial intelligence, artificial consciousness, neural networks, numerical simulation, artificial personality, synergic processes, phase portrait, strange attractor.

1. Introduction

The problem of engineering artificial consciousness (AC) has been attracting much interest in the context of the creation of an artificial personality (AP) with artificial intelligence (AI). In particular, Seth, Baars, and Edelman (2004) identified 17 widely recognized properties of consciousness in mammals that may be regarded in a broader sense such as to serve as criteria indicating the rise of AC of an AP. It was claimed in (Shevchenko 2016) that AC is the main block of a system possessing AI that is knowledgeable about itself and its environment, can control the processes in the system, and define its purpose. Dehaene, Lau, and Kouider (2017) proposed to distinguish two aspects of conscious information processes in the brain: the selection of information for **global broadcasting** needed for computation and report (C1, or consciousness in the first sense), and **self-monitoring** of this information processing (C2, or consciousness in the second sense). Graziano proposes an **attention schema theory** of consciousness, according to which the attended process is carried out in more depth, and the corresponding signal is broadcasted to other processes and is more likely to result in a behavioral decision (2017, 4). In this sense, Graziano's attention schema describes the C1 of Dehaene, Lau, and Kouider. In addition, Graziano and Webb claim that the attention schema is a bridge between information about the self and information about various aspects of the internal or external world to which attention is directed (2017, 204). They list three adaptive uses of the attention schema: (i) the control function; (ii) the integration of information obtained from

different information domains; (iii) promotion of social perception (Graziano and Webb 2017, 205–208).

However, the available description is not formalized to the extent applicable for machine use and modeling. In this article, we will put forward a more formal representation of consciousness allowing for numerical simulation and technical experiments.

2. Method and model

The proposed model is based on the results of (Seth, Baars, and Edelman 2004; Shevchenko 2016; Dehaene, Lau, and Kouider 2017; Graziano 2017; Graziano and Webb 2017). Based on these, it was assumed by Shevchenko et al. (2022, 30) that the emergence of consciousness is conditioned by two interrelated processes: the action of the attention mechanism (giving rise to the “computation and report” consciousness), and correlation of information flows in the system (that determines the “subjective perception” and “self-control” consciousness). These aspects of consciousness, in turn, are manifested in the well-known features of consciousness described in (Seth, Baars, and Edelman 2004, 121–122): widespread brain activity, wide range of contents, informativeness, internal consistency, sensory binding, focus-fringe structure, facilitation of learning, stability of contents, allocentricity, and conscious knowing and decision making.

Then the AC modeling should include two sides of the same process: (i) the modeling of an **attention schema** as a mechanism of information selection and broadcasting (C1); (ii) the modeling of the mechanism of **information flows correlation** (C2).

We regard the attention schema that generates C1 as a comparative mechanism of the global neural network, which estimates the importance of processes in the AP. Graziano suggests that the main node responsible for the attention schema of the brain is located in the temporoparietal junction (2017, 6). The scoring system is based on the built-in values of this AP, in particular on the moral and ethical principles implanted in the AC unit. The system allocates more resources to a more important process, including the provision of additional blocks and rebuilding the architecture to successfully solve the most necessary or immediate task. The rise of C2 is conditioned by the presence of a nonlinear system with feedback and external excitation.

The fact that the attention schema “possesses a non-physical, subjective awareness” (Graziano 2017, 1) and “describes something impossible and physically incoherent, a caricature of attention” (Graziano and Webb 2017, 193) make it possible to assert that consciousness is an emergent phenomenon that appears due to specific structures and links of the brain. It is important that this phenomenon is metastable (Friston 1997). Therefore, the model of consciousness can be formulated within the framework of synergetics (Hacken 2004). The features of this model are as follows:

- Consciousness is described by a decaying strange attractor, which exists for 3-4 periods;
- Consciousness extends to those physical objects that support the strange attractor;
- A personality is represented by a phase portrait (see Shevchenko et al. 2022, 30).

The decaying character of a strange attractor represents the properties of adaptive and fleeting nature and limited

capacity and seriality of consciousness that were described in (Seth, Baars, and Edelman 2004, 121).

A possible mechanism for modeling consciousness is a system of neural networks in which there is a nonlinear connection between neurons that are involved in different processes – that is, between segments of different networks. So, we write down the equations of two or more neural networks that connect hidden states and neurons, then apply a nonlinear connection to the input neurons (for example, consider them as points of a plane wave). The resulting equations, which contain the necessary nonlinear relations, can be used to construct phase portraits representing an AP (Shevchenko et al. 2022, 31).

Suppose that there run N simultaneous processes in a system of neural networks (either natural or artificial), each of which, according to the general theory of neural networks (Zhang, Lipton, Li, and Smola 2020), is described by the general equation:

$$Z_{\{i\}}^{[l]} = W_{\{i\}}^{[l]} X_{\{i\}}^{[l-1]} + b_{\{i\}}^{[l]}, \qquad (2.1)$$

where $X_{\{i\}}^{[l]}$ and $Z_{\{i\}}^{[l]}$ are vectors of input and output signals on the l-th layer of the network in the i-th process, $i = 1, 2, \ldots, N$, $l = 1, 2, \ldots, n$; $W_{\{i\}}^{[l]}$ and $b_{\{i\}}^{[l]}$ are the corresponding weight matrices and biases, respectively, which are trained parameters of the neural network that determine its properties. To be suitable for numerical simulation, the physical quantities $Z_{\{i\}}^{[l]}$, $X_{\{i\}}^{[l-1]}$, and $b_{\{i\}}^{[l]}$ should be normalized to their maximum values. The output vector of each layer must be filtered using a nonlinear activation function:

$$X_{\{i\}}^{[l]} = f_{act\{i\}}^{[l-1]}(Z_{\{i\}}^{[l-1]}). \qquad (2.2)$$

Depending on the problem to be solved, such a function may be a sigmoid [Hinton et al. 2012], ReLU or GELU [Hendrycks, Gimpel 2016], softmax [Goodfellow, Bengio, Courville 2016, 180–184], etc. In fact, the operation of the neural network (both natural and artificial) is described by the matrix multiplication procedure yielding weight coefficients for input signals, with the subsequent application of nonlinearity.

The result of the action of the neural network system, or its prediction, is expressed by the formula:

$$Y_{\{i\}} = f_{act\{i\}}^{[n_l]}\left(W_{\{i\}}^{[n_l]} f_{act\{i\}}^{[n_{l-1}]}\left(W_{\{i\}}^{[n_{l-1}]} f_{act\{i\}}^{[n_{l-2}]}\left(\ldots f_{act\{i\}}^{[1]}\left(W_{\{i\}}^{[1]} X_{\{i\}}^{[0]} + b_{\{i\}}^{[1]}\right)\right) + b_{\{i\}}^{[n_{l-1}]}\right) + b_{\{i\}}^{[n_l]}\right) \qquad (2.3)$$

where $Y_{\{i\}}$ is the prediction (output) of the system in the i-th process, $X_{\{i\}}^{[0]}$ is the input signal (data input).

It should be noted that the consciousness of living beings does not arise immediately after birth, but develops over time (Dehaene, Lau, and Kouider 2017, 5). This suggests that consciousness is an attribute of quite complex and already trained neural networks. Thus, modeling of consciousness presupposes the presence of trained values of $W_{\{i\}}^{[l]}$ and $b_{\{i\}}^{[l]}$, and therefore does not require consideration of the backward gradient propagation process that occurs during network training.

Since the transmission of voltage from neuron to neuron occurs in a natural neural network, an electrical analogy may be used. In this analogy, the process of brain activity or neural network calculations can be represented by a block diagram of the propagation of a set of electrical signals in a system of serially connected multiport devices with start relays after each

that are activated if the output signal exceeds a threshold.

In this case, information from one neuron to another is transmitted over time

$$\Delta t^{[l]}_{\{i\}jk} = R^{[l]}_{\{i\}jk} C^{[l]}_{\{i\}jk}, \qquad (2.4)$$

where $R^{[l]}_{\{i\}jk}$ and $C^{[l]}_{\{i\}jk}$ are the total electrical resistance and electrical capacity of the elements of the matrix $W_{\{i\}}^{[l]}$, respectively, which are trained parameters of the network.

After passing one neural network layer, the increase in the vector of the electrical signal is:

$$\Delta X^{[l]}_{\{i\}} = Z_{\{i\}}{}^{[l]} - X_{\{i\}}{}^{[l]}, \qquad (2.5)$$

Given this, the rate of change of the electrical signal vector is

$$\frac{\Delta X^{[l]}_{\{i\}}}{\Delta T} = [E - \left(W_{\{i\}}{}^{[l]}\right)^{-1}]X_{\{i\}}{}^{[l]} + \left(W_{\{i\}}{}^{[l]}\right)^{-1} b_{\{i\}}{}^{[l]}, \qquad (2.6)$$

where E is the identity matrix. Since the objects in (2.6) have, in general, different dimensions, the arithmetic operations on them may involve broadcasting.

The quantity in the left-hand side of (2.6) will be considered as a generalized time derivative of the input vector. Its discrete character complicates the analytical calculations based on (2.6), as smoothing and extrapolation procedures are required. However, this function is fully acceptable for a numerical experiment.

Thus, for numerical simulation we will use (2.1), (2.2), and (2.6). These equations already have a nonlinearity – the activation function $f_{act\{i\}}{}^{[l-1]}(Z_{\{i\}}{}^{[l-1]})$, which is applied to each layer of the network. But to model the global dependence of the processes that take place in the neural network, it is necessary to impose a nonlinear restrain on the input signals of different processes in the system.

The scheme of attention, which is responsible for C1, identifies the most important process that is the most needed or most urgent for survival or a current task. Let $i = 1$ for this process. This process, on the one hand, acquires priority in determining the following actions of the system (starting the manipulator, starting some other process, redistributing memory, etc.). On the other hand, the input signal of this process is shared with other processes in the system and becomes interconnected with them, which is a manifestation of C2.

Certainly, the degree of such sharing has its limits. The minimum degree is the transfer of one transformed component of the input signal of the selected process to some neuron of each other process. The maximum sharing degree is the transmission of all the selected signal components to all other processes in the system, which will indicate their complete suppression. Obviously, in the minds of living beings there are mostly some intermediate stages of sharing, where certain components of the selected signal link the system into a single whole.

Assume that $X_{\{i\}}^{[0]} = \left(x_{1_i}^{[0]}, x_{2_i}^{[0]}, \ldots, x_{n_{xi}}^{[0]}\right)$ is the input signal of the i-th process, the components of which are normalized relative to their maximum value. Therefore, these components are dimensionless and their absolute values do not exceed unity, which allows us to consider them as values of a sinusoidal function. The wave of the globalization signal, which combines different processes, is represented as a sinusoidal dependence of the component signals of the system processes, where the phase of the wave corresponds to process number i.

Then we have for the components of the input signals:

$$x_{ki}^{[0]} = \sin[\arcsin\left(x_{ki}^{[0]} - 2\pi(i-1)/N\right)], \qquad (2.7)$$

where $k_i = 1, 2, \ldots, n_{xi}$. If the correlation is conditioned by only one component, then $k_i = 1$.

The formula (2.7) defines the nonlinear relationship between different processes due to the action of consciousness in the second sense (C2) giving rise to meta-cognition (cf. Dehaene, Lau, and Kouider 2017, 5), which is the basis for considering these processes as synergistic.

3. Discussion

As the development of consciousness strongly correlates with that of their reporting capacity (Dehaene, Lau, and Kouider 2017, 5), we expect that the language-competent systems able to express their states and wishes will be more prone to acquire sophisticated consciousness. We see here promising perspectives in the use of machines with semantic modules able to deeply understand language based on deep semantics grounding on semantic fields (Vakulenko 2021; Vakulenko 2022a; Vakulenko 2022b) aiming at semantic landscapes of languages.

The promising follow-up of this research may be experimenting with nonlinearly interconnected real artificial neural networks in a computer or a set of joined computers.

5. Conclusion

So, in this article, we proposed a synergetic approach to modeling artificial consciousness, that relies on recent advancements in neuroscience, information technologies, and philosophy. The method requires numerical calculations and experimental study.

References

- Dehaene, S., Lau, H., and Kouider, S. 2017. What is consciousness, and could machines have it? In: Science 358, pp. 486-492.

- Friston, K. J. 1997. Transients, metastability, and neuronal dynamics. In: Neuroimage 5, pp. 164–171.
- Graziano, M. 2017. The attention schema theory: A foundation for engineering artificial consciousness. In: Frontiers in Robotics and AI 4, art. 60, pp. 1-9.
- Graziano, M., and Webb, T. 2017. Understanding consciousness by building it. Part three: Metaphilosophy of consciousness studies. In: Bloomsbury companion to the philosophy of Consciousness, pp. 185-210.
- Haken, Hermann. 2004. Synergetics: Introduction and Advanced Topics. Springer. 758 p.
- Hendrycks, Dan, and Gimpel, Kevin 2016. Gaussian Error Linear Units (GELUs). In: arXiv:1606.08415v4 [cs.LG].
- Hinton, Geoffrey, Deng, Li, Yu, Dong, Dahl, George, Mohamed, Abdel-rahman, Jaitly, Navdeep, Senior, Andrew, Vanhoucke, Vincent, Nguyen, Patrick, Sainath, Tara, and Kingsbury, Brian 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition. In: IEEE Signal Processing Magazine 29 (6): 82–97. doi:10.1109/MSP.2012.2205597.
- Goodfellow, Ian, Bengio, Yoshua, Courville, Aaron 2016. 6.2.2.3 Softmax Units for Multinoulli Output Distributions. Deep Learning. MIT Press. 799 p.
- Seth, Anil K., Baars, Bernard J., Edelman, David B. 2004. Criteria for consciousness in humans and other mammals. In: Consciousness and cognition 14: 119-139.
- Shevchenko, A. I. 2016. Do pytannja shchodo stvorennja shtuchnogho intelektu [To the question of creating artificial intelligence]. In: Shtuchnyj intelekt 1, pp. 7–15 [in Russian].
- Shevchenko, A., Bilokobyljsjkyj, O., Vakulenko, M. et al. 2022. Regarding the draft strategy development of artificial intelligence in Ukraine (2022 – 2030). In: Shtuchnyj Intelekt, vol. 1, pp. 8–157. URL: https://jai.in.ua/archive/2022/2022-1-1.pdf.
- Vakulenko, Maksym. 2021. From Semantic Metrics to Semantic Fields. In: Proceedings of the 2021 IEEE 16th International Conference on Computer Science and Information Technologies (CSIT), 22–25 September 2021, Lviv, Ukraine. Pp. 44–47. DOI: 10.1109/CSIT52700.2021.9648675.
- Vakulenko, Maksym O. 2022a. Semantic comparison of texts by the metric approach. In: Digital Scholarship in the Humanities. Published online: 11 October 2022. DOI: 10.1093/llc/fqac059.
- Vakulenko, Maksym O. 2022b. Deep contextual disambiguation of homonyms and polysemants. In: Digital Scholarship in the Humanities. Published online: 19 December 2022. DOI: 10.1093/llc/fqac081.
- Zhang, Aston, Lipton, Zachary C., Li, Mu, and Smola, Alexander J. 2020. Dive into Deep Learning. Release 0.7.1. Apr. 14, 2020.