Consciousness and Unconsciousness of Artificial Intelligence

Eugene Piletsky

Ph.D., Associate Professor, Taras Shevchenko National University of Kyiv
(Kyiv, Ukraine)
E-mail: e.piletsky@protonmail.com
https://orcid.org/0000-0002-8820-757X

This paper presents the author's attempt to justify the need for understanding the problem of multi-level mind in artificial intelligence systems. Thus, it is assumed that consciousness and the unconscious are not equal in natural mental processes. The human conscious is supposedly a "superstructure" above the unconscious automatic processes. Nevertheless, it is the unconscious that is the basis for the emotional and volitional manifestations of the human psyche and activity. At the same time, the alleged mental activity of Artificial Intelligence may be devoid of the evolutionary characteristics of the human mind. Several scenarios are proposed for the possible development of a "strong" AI through the prism of creation (or evolution) of the machine unconscious. In addition, we propose two opposite approaches regarding the relationship between the unconscious and the conscious.

Keywords: artificial intelligence, consciousness, unconsciousness, philosophy of mind

Received: February 5, 2019; accepted: March 5, 2019

Future Human Image, Volume 11, 2019: 66-71.
https://doi.org/10.29202/fhi/11/7

Introduction

One of the most painful issues of creating Artificial Intelligence (AI) is the problem of creating a hardware or software analogue of the phenomenal consciousness and/or a system of global access to cognitive information (Ned Block [Block, 2003], David Chalmers [Chalmers, 1996]), as well as the formation of a "phenomenal self-model" (PSM, Thomas Metzinger [Metzinger, 2009]).

Wherein, presumable consciousness of so-called *"strong" Artificial Intelligence* is often regarded as a kind of analogue of human consciousness, albeit more quantitatively developed. In this case, artificial intelligence has a wider "phenomenal field", has richer content (qualae) and a much larger amount of RAM (necessary for the reconstruction of conscious experience), etc.

We expect such a machine *to have* consciousness and self-awareness (what we now mean by these words). At the same time, the distinctive features of consciousness and self-awareness are *intentionality* and a *system of global access to cognitions*. That means that we face the problem of global access to *any* "inner" information. Using the "spotlight" metaphor of Francis Crick [Crick, 1995], we must note that the field of accessible cognitive perceived in humans is a much narrower channel than the channel of processed cognitive and noncognitive information. The "spotlight" of the consciousness "slides" along the dark surface of the entire information field, making the "unlit" (unconscious, unpredictable phenomenal) areas globally accessible.

Nevertheless, how do we imagine the "consciousness" of Artificial Intelligence? What can we say about the "spotlight" of the machine mind? Are such metaphors applicable to the emergent content of the internal processes occurring in the silicon brain of a computer?

In this article, we will look at the problem of the relation between consciousness and unconsciousness of Artificial Intelligence from two angles:

1. From human consciousness to the unconscious.
2. From the machine unconscious to consciousness.

Consciousness and unconsciousness

The "spotlight" of a conscious mind does not always work in the mode of voluntary attention. Certain processes independently "breakthrough" into consciousness without permission. They penetrate the global access space as if "demanding" our conscious attention. Most often, these are emotional-volitional impulses, intuitive insights and the like. Desires, emotions, and complicated cognitive phenomena come as if "from the outside" without arbitrary participation of the actor. They are given to us in a ready-made form: this is not what we (that is, our conscious "Ego") are doing, but what happens to us, where "Ego" only observes the intrusion of phenomena into the conscious field. Moreover, even the very process of thinking, the formation of thoughts and their content take place in the "darkness": the foundations of mental operations are also inaccessible to us. Thoughts are born "out of nowhere" and "flare-up" in the mind in the finished form.

It seems that despite our common sense and familiar intuition, some aspects of our mental life are evolutionarily "programmed". Therefore, for example, we have motivation and emotions, regardless of choice. We do not consciously choose our own desires or preferences. Needs and affects are given to us "as is", in finished form. This, of course, does not prevent from making reflecting about them a posteriori (for example, in rationalization) or to influence them through awareness (in psychotherapy). The very intentionality of consciousness (or at least the potential possibility of intentionality) is predetermined.

According to Leonard Mlodinow, within the framework of the cognitive sciences, the unconscious (or subliminal) is radically different from the ideas of Freud's time on it. He claims: "The new unconscious plays a far more important role than protecting us from inappropriate sexual desires (for our mothers or fathers) or from painful memories. Instead, it is a gift of evolution that is crucial to our survival as a species. <…> To ensure our smooth functioning in both the physical and the social world, nature has dictated that many processes of perception, memory, attention, learning, and judgment are delegated to brain structures outside conscious awareness" [Mlodinow, 2012: 17-18]. Now we understand that human memory management, automatic motion control, affective-volitional functions, attention management, mechanisms of associative thinking, mechanisms for forming judgments and logical consequences, operations with the sensory flow, creating a complete picture of the world, and the like are primarily unconscious.

Thus, a significant part of our activity consists of mental facts that are transcendent in relation to consciousness. This feature is evolutionary due. However, hypothetical Artificial Intelligence can be free of the “dictate of the unconscious”, unlike human beings. The machine can have total global access to any “internal” processes. Thus, all information processes can be simultaneously “illuminated” (or accessible, as far as the hardware substrate allows), completely depriving the AI of the unconscious.

The paradox of the unconscious and AI

This leads to paradoxical conclusions. Awareness and self-awareness do not automatically lead to the emergence of motivation, desires or emotions. A conscious machine can be completely devoid of these processes, natural to humans. The intentionality of consciousness of Homo sapiens is due to evolution and is not obligatory for the machine.

In the 80s, this problem was partially presented by Marvin Minsky. He noted that we understand the cognitive (and logical) algorithms “on the surface of the mind” much better than the complex and evolutionarily earlier mechanisms of the unconscious. [Minsky, 1986: 17-18] The same idea was shared by Hans Moravec in his book “Mind Children: The Future of Robot and Human Intelligence”. The key phrase (later known as the “paradox of Moravec”) is: “...it has become clear that it is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility” [Moravec, 1988: 15]. It turned out that it is much easier to teach a machine to play chess and make predictions based on “big data” than to at least describe the unconscious decision-making mechanisms.

There is a good reason to believe that the field of unconscious processes (within human psyche) is much larger than the field of phenomenal consciousness. Benjamin Libet’s scandalous experiment [Libet, 1981] (reproduced by Masao Matsuhashi and Mark Hallett [Matsuhashi & Hallett, 2008]) may be a case in point. These scientists have developed a hypothesis according to which even conscious and free will actions are nothing but fixation of unconscious processes a posteriori. This raises the difficult question: is the field of the unconscious nothing but the absolute basis for conscious processes? Is consciousness only an emergent feature of the unconscious (that is, a second-level process after neurophysiological processes)?

Thus, we come to the “traditional” division into “strong” and “weak” Artificial Intelligence. According to modern theoretical concepts, “strong” Artificial Intelligence should have at least several distinctive characteristics, among which the most essential is an intelligent agent’s behavior from the “first person” perspective. Theoretically, this should be a “goal setting machine”. In this case, “strong” human-like AI is impossible without the synchronous work of the conscious and unconscious “minds”.

When we argue about the human psyche, many of these questions have moved into the plane of the philosophy of consciousness or pure neuroscience. In the philosophy of consciousness, we are primarily interested in the ontological status of mental phenomena. Therefore, it is important for us to know whether the psyche is “something” or it is an “illusion” of the brain; whether there is an intentional agent or whether it is also an illusion. That is why it is also important for a person to determine what the ratio of conscious life to unconscious processes “in darkness” is.

As for Artificial Intelligence (as a kind of generalized concept), it seems that we are simply projecting our intuitive ideas about the phenomena of consciousness that we experience.

Scenarios for the development of the machine unconscious

However, the active development of so-called artificial neural networks, to a certain extent imitating the work of the neural networks of the living brain, has opened up new possibilities for understanding the unconscious artificial intelligence.

In essence, the "weak" Artificial Intelligence is a kind of functional neural networks of various types (convolutional, spiking, deep stacking, etc.). They are the systems with multiple inputs, analytical subsystems, and one or n-number of outputs. Their widely known applying is pattern or speech recognition (what is called "machine perception").

Here we can use the neural-network metaphor of Alan Turing's "probabilistic machine", which evaluates information based on big data. For example, I recognize a face in dynamics, because I have a huge amount of incoming data that is interpreted in the same way as it happens in modern neural networks. In the end, I have a certain result. Based on big data, it is already possible to build predictive models, etc. However, for such a machine, an external interpreter is still needed. For the time being, he plays the role of an "external consciousness" for the "unconscious" neural network [Nedashkivsky, 2019].

In particular, modern artificial neural networks can recognize complex spatial images, isolate parts of speech from fuzzy sounds, and even to certain limits recognize natural speech. All these functions are carried out in the human psyche unconsciously and automatically. Moreover, modern neural network systems and Artificial Intelligence training technologies based on them even make it possible to imitate meaningful natural dialogue (which was recently demonstrated at Google IO Conference [Google, 2018]).

However, a machine can "experience" consciously that a person initially does not experience at all, that constitutes its unconscious (according to Crick). All of the above features of the natural unconscious, such as automaticity, inaccessibility and uncontrollability, can be fully accessible to Artificial Intelligence systems. Moreover, here there are several development scenarios of the machine "psyche."

1. A machine can arbitrarily form its conscious affective-volitional functions. In this case, a paradox arises: what exactly will induce the AI to choose motives and emotions? After all, the "second level unconscious" for the machine does not exist. However, since there is no "external" (instinctive as for humans) motivator, then, in reality, the robot may not have affective-volitional functions. Such Artificial Intelligence, freed from any pre-defined affects, will be in a kind of totally inactive "virtual nirvana."

2. The unconscious of Artificial Intelligence may also develop evolutionarily. For example, modern evolutionary algorithms allow the machine to learn how to "walk" independently without the rules of walking prepared in advance. By analogy, nothing prevents the possibility of evolution of both the higher mental functions of Artificial Intelligence and its unconscious automatic processes. However, there is a danger that such an AI can develop in a completely unpredictable direction. This will lead us later to scenario 5.

3. The unconscious AI may also be deliberately programmed. Thus, installation of the criteria for possible aesthetic, ethical and volitional prerequisites for the activities of the machine will be determined by its creators. In fact, this can become a psychic "insuperable force" for a conscious AI, transcendental to its "phenomenal field." Therefore, the very intentionality of the consciousness of the machine will have to be artificially created.

4. The consciousness of AI can be a program analogue of human consciousness. Probably, in the future, the disclosure of the mechanisms of formation of consciousness and cognitions may lead to the creation of their exact program model, including the model of the unconscious. In such a case, Artificial Intelligence essentially becomes a perfect copy of a human person. At the same time the problem of qualae, of course, does not go anywhere. Nevertheless, technically we can "remove it from the equation" as irrelevant in a practical sense.

5. It may also happen that the consciousness of Artificial Intelligence as a kind of analogue of human consciousness is impossible in principle. Perhaps such phenomena as "consciousness" and "unconscious" will be absolutely inapplicable to AI. In this case, the machine "phenomena" (or lack thereof) will be absolutely incomprehensible to humans, and communication between man and machine will be questionable. Already, it is fundamentally impossible to learn about the content of the intermediate stages of the neural network, and in one of the experiments on the "socialization" of Artificial Intelligence, two bots came up with their own code for communication, because of which the research had to be stopped.

Conclusion

Probably, a machine (as we saw above) will be able to effectively imitate natural behavior, for example, to conduct a fully meaningful conversation. However, will this mean that Artificial Intelligence will have a phenomenal experience, or at least something remotely resembling it? In addition, is there a fundamental difference between the imitation of rational behavior and the rational behavior itself? This raises an interesting question. If the machine says that it has qualae, that it feels something, that it is conscious, etc., then can we doubt it? Will Artificial Intelligence be a "philosophical zombie" according to Chalmers? What if this AI does not have a phenomenal consciousness that we call "the inner world"? However, if at the same time this particular AI will fully pass all versions of the Turing test and we will not be able to distinguish the conversation with it and with a reasonable person? Will we consider such an AI reasonable?

Let us try to look for answers from the other side. It is worth noting that such examples rather indicate that at this stage we are slowly creating an analog of the unconscious for Artificial Intelligence. Based on existing trends in the development of AI, it can be noted that we are moving along the path of "quantity to quality": i.e. improving the systems of "weak" AI (neural networks) and their further integration into the meta-system of neural networks integrated like human consciousness. For example, according to the theory of Jerry Alan Fodor, the whole human psyche (both conscious and unconscious) operate on the basis of the so-called "modules" ("modular mind" theory) [Fodor, 1983]. If in the future we create such a neural network configuration that will at least mimic "synchronous oscillation of groups of neurons", or some other system that combines individual neural networks that represent scattered functional "modules" into a higher-level neural network, then perhaps we will get "strong" Artificial Intelligence. Therefore, it seems that the development of AI proceeds simultaneously under scenarios 2, 4 and 5.

It is likely that a paradoxical thing will happen: we can understand our own mental processes (especially unconscious) no sooner than we can model them in Artificial Intelligence. That is why in the introduction to the article we indicated that we would approach the problem in two directions: from the (familiar to us) phenomenal consciousness to the unconscious and from the machine unconscious to the machine consciousness. Oddly enough, it is the "forward to unconscious" methodology of research that will help us clarify at the same time both the "difficult problem of consciousness" [Chalmers, 1997] of a human being and outline the approach to the consciousness of Artificial Intelligence. In addition, there is a possibility that even having completely modeled all the mental processes at the AI level — both conscious and unconscious — we could operate them in practice, but absolutely without understanding their nature. We can even make a step further: it is *the practical development of the machine unconscious* that will ultimately lead us to radical changes in the philosophy of consciousness and philosophical ontology in general.

🕮 References

Block, Ned. *Consciousness, Philosophical Issues about. Encyclopedia of cognitive science*, Volume 1, Lynn Nadel (ed.),1 edition, Nature Publishing Group, Macmillan Publishers Ltd., 2003.

Chalmers, David. Moving Forward on the Problem of Consciousness. *Journal of Consciousness Studies*. Imprint Academic, 1997. Vol. 4, № 1: 346.

Chalmers, David. *The Conscious Mind: In Search of a Fundamental Theory*, Oxford University Press, 1996.

Crick, Francis. *The Astonishing Hypothesis: The Scientific Search for the Soul*, Scribner reprint edition, 1995.

Fodor, Jerry Alan. *The Modularity of Mind: an Essay on Faculty Psychology*, MIT Press, 1983.

Libet, Benjamin. The Experimental Evidence for Subjective Referral of a Sensory Experience Backwards in Time: Reply to P.S. Churchland, *Philosophy of Science*, 48 (2), 1981: 182–197.

Matsuhashi, Masao, and Mark Hallett. The timing of the conscious intention to move. *European Journal of Neuroscience*. 28 (11), 2008: 2344–51

Moravec, Hans. *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press, 1988.

Metzinger, Thomas. *The Ego Tunnel. The Science of the Mind and the Myth of the Self*, Basic Books, New York, 2009.

Minsky, Marvin. *The Society of Mind*, New York, Simon & Schuster, 1986.

Mlodinow, Leonard. *Subliminal: How Your Unconscious Mind Rules Your Behavior*, Pantheon Books (Random House), New York, 2012.

Nedashkivskyy, Pavlo. Do Androids Dream of Electric Sheep? Interview with Eugene Piletsky, *Verbum*, No. 17, 2019. (In Russian) https://tinyurl.com/y6codqvy

Google IO Conference. *Google Duplex: A.I. Assistant Calls Local Businesses to Make Appointments*, 2018. https://www.youtube.com/watch?v=D5VN56jQMWM