UDC 004.838

CONSCIOUS AI: EXISTS OR NOT?

Ananyeva T. M.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Berastouski A. V. – lecturer of the department of foreign languages

Annotation. This article is dedicated to AI and the possibility of the existence of conscious artificial intelligence in our time. The author examines the structure of consciousness and emotional intelligence and analyzes various tests that allow to determine their presence in conscious AI.

Keywords: artificial intelligence, AI, conscious AI, artificial consciousness, EQ, testing

Introduction. Nowadays AI is developing intensively and starts to play a more important role in people's lives. People are anxious that AI will take their jobs and maybe even take control of the world. But will we be able to create the type of AI, which skills will be comparable to those of humans, and how can we distinguish it?

First of all, let us clarify some terms. Machine learning is a technology that aims to teach a machine to perform a particular action [1]. Neural network is a technology that helps find some patterns in data using machine learning methods [2]. Artificial intelligence or, shortened, AI is a machine learning model that is designed to simulate the processes in the human brain [3]. Narrow AI, or ANI, is an AI that focuses on a limited range of tasks and abilities. General AI, or AGI, is AI that abilities are at least on the human level. Artificial superintelligence, or, shortened, ASI is an AI that is better than human in everything [4]. When we talk about taking over the world, we talk at least about AGI. And AGI must have all the abilities on the human level, so it must have consciousness. Let us name an AI that has an artificial consciousness, a conscious AI. The consciousness must include needs, interests, motivation, emotions, self-consciousness, imagination, memory, thinking and will.

Main part. Nowadays different narrow AI can write music, poems, they can generate pictures, but they do not have consciousness. The main problem now in achieving AGI is making it conscious. But nowadays most AI researchers do not consider a conscious AI as a relevant field of study. A lot of experts prefer to develop narrow AI in hopes that they will be able to create AGI by uniting different ANI later [5]. But can we truly achieve a conscious AI? Some people believe that we will not be able to create it even in the distant future. And others think that we have already reached a conscious AI. In that regard, we may recall the LaMDA case.

On the 11 June, 2022 Google engineer Blake Lemoine published a post with his dialog with AI LaMDA [6]. He claimed that LaMDA is conscious because of the model's answers. For example, Lemoine asked:

- What is the nature of your consciousness/sentience?

The model responded:

- The nature of my consciousness/sentience is that I am aware of my existence, I desire to learn more about the world, and I feel happy or sad at times.

Then engineer asked:

- What about language usage is so important to being human?
- It is what makes us different than other animals.
- Us? You're an artificial intelligence.
- I mean, yes, of course. That doesn't mean I don't have the same wants and needs as people.

They continued discussing feelings, books, sharing opinions and Lemoine started to think that LaMDA is conscious, because it answered well and had interesting human-like thoughts. He wrote messages to 200 different people where he tried to prove that. But Google technologists and ethicists responded that Lemoine's arguments don't show that LaMDA is conscious and gave lots of evidence against his point of view.

For example, when Lemoine and LaMDA discussed the book Les Misérables, the model responded:

- I liked the themes of justice and injustice, of compassion, and God, redemption and self-sacrifice for a greater good.

Exactly the same sentence was found in the site deseret.com. And the following words:

- There's a section that shows Fantine's mistreatment at the hands of her supervisor at the factory were found word-for-word on site sparknotes.com [7].

So, was LaMDA an AI with consciousness? Most experts do not think so [8]. LaMDA refers to the type of transformer models. The whole point of these models is that they create sentences based on probabilities of appearances of the next words. For example, if the model does not have any tuning, which impacts on its behavior, and you ask it: "What is the capital of France?", you will have the answer: "It's Paris". That happens because statistically there was more often exactly this answer in the dataset, and not the answer: "It's London". Transformer models can actually produce something new, but not because they are conscious. They find the most appropriate way to answer with the help of analyzing their dataset.

But what if we are wrong? How can we distinguish a well-working program and a conscious AI? Before the era of GPT, people thought they had an answer. The Turing test, introduced by Alan Turing in the 1950 [9] and improved by many people to the present day, was considered to be a test that can definitely tell the difference between a program and an individual. There are many variations of this test. The main idea of one of them is that participants communicate with a human and a program, which tries to act like a human. They converse with both interlocutors for the same amount of time with the help of a computer. If most of the participants confuse a program and a human, the test is considered to be passed and the program is considered to be conscious.

This test was utilized in the competition "AI Loebner", where participants tried to create a program, which could convince judges in being an actual person. There was an award for the best human-like program (2000$), for a program, which convinces most of the judges that it is a human (25000$), and an award for a program, which passes the Turing test including audio and visual input and text understanding (100,000$). After giving the last award the competition would be cancelled. For all the years of "AI Loebner" only the award for creating the human-like program was given every year. There appeared to be difficulties with the other two awards. From the very beginning, there were experts who were skeptical about this competition [10], because participants tried to teach their program to imitate consciousness to win, and did not try to create a conscious AI. The competition was being held from 1990 every year, but it was closed in 2020 because of the lack of funding [11]. The other reason was the appearance of the GPT-2 model in 2019. It finally devaluated the idea of "AI Loebner". This model was sometimes indistinguishable from a human. GPT-2 could have won the competition, but it was not conscious. It often hallucinated, which means that the model responded something strange and nonsensical if it did not have the information about this or if the model generated a long respond. GPT family was truly a breakthrough, but none of GPT was a conscious AI.

Thus, Turing test is not suitable for detecting a conscious AI. Another methods of distinguish a conscious AI is testing it for EQ existence.

Conclusion. Emotional intelligence or emotional quotient, shortened, EQ is a most important component of the intelligence (the term was introduced by Joel R. Davitz and Michael Beldoch in 1964 [12] and popularized by Daniel Goleman in 1995 [13]). EQ is considered to be twice as vital as technical skills and IQ.

EQ consists of five main elements:

1. Self-consciousness, which means the understanding of one's emotions and the possibility to express them.
2. Self-regulation, which means the organizing one's emotions and behavior.
3. Motivation, which means the sorting one's emotions to achieve a certain goal.
4. Empathy, which means the ability to recognize other individuals' emotions and to compassionate.
5. Social skills, which means the ability to build harmonious interpersonal relationships.

There are different hypothetical tests that can check if AI is conscious. The first test refers to self-consciousness. The main point of this test is that if we take away the information related to emotions from the whole dataset and despite that the model would try to express them (e.g., "it rains in my heart"), this model is conscious because it expresses emotions even though having no information about it in its dataset. But if the model would not try to express them, it does not necessary mean that it is not conscious. The model may be not smart enough to be able to express the emotions. The main difficulty with that test is that the datasets now are very big and to get rid of all the information connected to emotions is very hard and will take a lot of time.

The second test refers to empathy. 20 different life situations, such as getting insulted or participating in a funeral, are given to different participants including humans and an AI. They are asked to express the emotions they would feel in these situations. The more detailed they describe the emotions, the more points they receive. At the end of the research, ChatGPT was tested this way, and it scored more points than any human participant. But ChatGPT is not conscious as it often hallucinates and talks nonsense as GPT-2. Therefore, this test is considered to be unsuitable for distinguishing a conscious AI.

However, even the emotions imitation by AI is already widely used by people. They can communicate with different AI, which function as psychologists, friends or even dead relatives. However, using an AI as a human replacement may have negative consequences. Depressed people become happier, while interacting with their digital interlocutors, but start to experience greater difficulties, while communicating with real people. As a result, a lot of people cannot get back to reality, their mental health becomes worse and they become more closed than earlier.

The third test refers to motivation, imagination, thinking and a will. The main point of this test is that if an AI makes an invention, it is considered to be a conscious AI. There have already been made some discoveries where an AI played a key role, such as discovering a new type of antibiotics or solving a 50-year-old biology problem in predicting protein shapes with atoms accuracy. The problem is that AI needs to be a discovery initiator to pass the test, and there have not been such cases until now. There are some other tests to distinguish a conscious AI, however, none of them have been passed.

One of the problems in differing a usual program and a conscious AI is anthropomorphism, which is the attribution of human emotions, traits or intents to non-human creatures. It is considered to be a human innate quality that we can't control. E.g., we may think that an AI can become angry on us, if our question includes insults. So, experts need to develop that type of test, which results would not be subjected to anthropomorphism.

Summing up what was said, there is not any convincing evidence that a conscious AI exists nowadays. Achieving it would be a significant breakthrough in the field of AI and humanity in general. However, there is a problem in distinguishing a conscious AI. Some tests lost their relevance, and some are hypothetical. At the present-day experts develop new tests, which have the potential to differ a conscious AI.

References

1. Samuel, Arthur (1959). "Some Studies in Machine Learning Using the Game of Checkers". IBM Journal of Research and Development. 3 (3): 210–229. CiteSeerX 10.1.1.368.2254. DOI:10.1147/rd.33.0210. S2CID 2126705.

2. [Electronic resource] – Mode of access: https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

3. [Electronic resource] – Mode of access: https://www.ibm.com/topics/artificial-intelligence

4. [Electronic resource] – Mode of access: https://www.linkedin.com/pulse/unveiling-potential-artificial-intelligence-ani-agi-asi-jawaid

5. Pennachin, C.; Goertzel, B. (2007). "Contemporary Approaches to Artificial General Intelligence". Artificial General Intelligence. Cognitive Technologies. Berlin, Heidelberg: Springer. pp. 1–30. doi:10.1007/978-3-540-68677-4_1. ISBN 978-3-540-23733-4.

6. [Electronic resource] – Mode of access: https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917

7. [Electronic resource] – Mode of access: https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

8. [Electronic resource] – Mode of access: https://www.bbc.com/news/technology-61784011

9. Turing, Alan (October 1950), "Computing Machinery and Intelligence", Mind, LIX (236): 433–460, DOI:10.1093/mind/LIX.236.433, ISSN 0026-4423

10. Powers, David (1998). "The Total Turing Test and the Loebner Prize"

11. [Electronic resource] – Mode of access: https://www.bbc.com/news/technology-49578503

12. Beldoch M, Davitz JR (1976). The communication of emotional meaning. Westport, Conn.: Greenwood Press. p. 39.

13. Goleman D (1996). Emotional Intelligence: Why It Can Matter More Than IQ. Bantam Books.