HOW TO TEST ARTIFICIAL INTELLIGENCE FOR THE PRESENCE OF CONSCIOUSNESS

F. Gritsev, I. A. Krivolap

fgritsev@gmail.com; ilyakrivolap88@gmail.com;
Superviser – Natallia P. Tesliuk, PhD (Philology), Associate Professor

The article under consideration deals with artificial intelligence as an object of philosophical analysis. The author presents a new test designed to check artificial intelligence for the presence of consciousness. Research in this field is topical as at present complex programs are being developed to bypass the classical Turing test.

Keywords: Artificial intelligence; the Turing test; Mary's room; Chinese room; consciousness.

INTRODUCTION

It is worth starting with the definition of the concept of artificial intelligence (hereinafter – AI). AI is a program, an algorithm that performs tasks related to intelligent activity. Now AI is already being introduced into our lives. Along with this introduction, the question of whether a person can create a conscious program is becoming increasingly popular.

This question became especially relevant this summer, after the scandalous dismissal of a Google engineer who claimed that the LaMDA AI chatbot was conscious. The programmer even began to look for a lawyer for this AI.

In this regard, another question arises: how can a program be tested for the presence of consciousness? After all, as J. Locke [1, p. 122] and L. Wittgenstein [2, p. 141] believed, we cannot even understand what other people feel, respectively, how can we know if the program feels anything?

RESEARCH MATERIAL AND METHODS

To answer this question properly, we analyzed philosophical scientific research exploring the concept of consciousness. The English mathematician A. Turing tried to answer the question whether or not a computer is capable of thinking like a human being by proposing his own test [3, p. 433–460]. The idea of his test is that if a person cannot understand whether he is talking to a machine or not, then the machine is conscious.

However, the LaMDA example showed that this approach might not work in practice.

A group of scientists, long before the advent of LaMDA, in the form of a thought experiment called *the Chinese Room*, criticized A. Turing's experiment [4, p. 417–424].

The idea of the experiment is that a person who does not know Chinese at all is sitting in the room. However, this person has precise instructions which characters he should respond to with which characters. Character cards arrive in this room, and as instructed, this person responds with other cards in Chinese. For example, he may be asked: ‘What is your favorite color?’ and he can answer with a card with the hieroglyph ‘blue’, but will it be a conscious answer? Of course not, because this person does not understand a single hieroglyph.

In this analogy with the Chinese room, a person acts as AI, and an instruction as an AI program with clear algorithms.

The man in the Chinese room does not understand what he is doing, but only blindly follows the instructions.

Then you need to find out what distinguishes a being that has consciousness from the one that does not. The ability to answer questions and even formulate them, as we have seen, is not a criterion. Moreover, we do not impose such requirements even on all people.

In modern philosophy of consciousness, two directions can be roughly distinguished:

1) Physicalism. Consciousness is a collection of physical processes in our brain.

2) The theory of ‘Qualia’. Consciousness exists independently of the brain.

From the point of view of both approaches, consciousness can be inherent in AI; in the first case, it is possible, since the brain and the neural network are two complex information systems, so there is no significant difference between them [5, p. 591].

From the point of view of qualia, any information system, even a thermostat, is in some sense conscious.

Supporters of the second approach propose to divide the question of consciousness into two parts: easy problems of consciousness and difficult ones. Easy problems are those that in the study of consciousness are solved by standard scientific methods, with the help of reduction physicalism. These methods make it possible to explain to a third person what consciousness does, how it changes over time, and what its structure is. A difficult problem arises when asking the question ‘Why does consciousness exist?’ [5, p. 3]. The answer to this question requires going beyond the application of scientific methods.

One of the main theorists of this approach, Frank Jackson, proposes a thought experiment called *‘Mary's room’* [5, p. 266] to show that phenomenal experience for a conscious being is not just the ability to collect and even analyze information. The phenomenal, subjective experience proper is information.

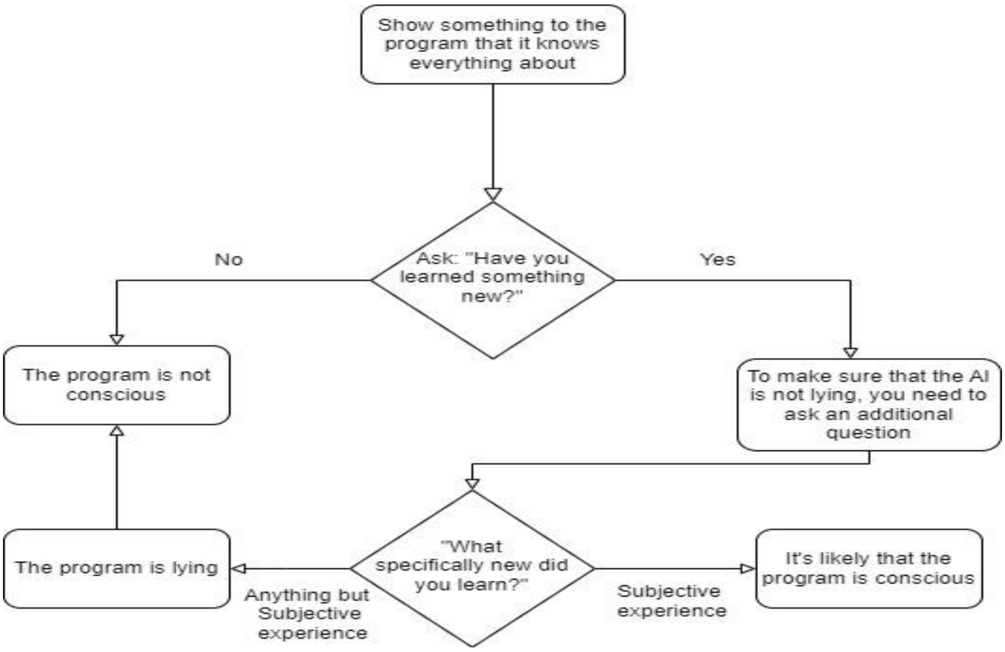
The meaning of the experiment with Mary's room is that there is a certain black and white room in which Mary lives. Mary has never seen any flowers. However, sitting in her room, she studied everything about color, from the physical properties of light and matter to how the brain processes the information it receives through the eyes. However, when she left the room one day, she learned new information about color – the subjective experience of ‘feeling’ color.

Here we are not trying to prove the rather controversial concept of ‘qualia’, here we are asking the question whether it is possible for AI to experience phenomenal experience, as we do. That is, we can say that, unlike a computer, we can be subjects of experience.

RESULTS AND DISCUSSION

Based on this, we offer our own test for the presence of consciousness in the program (see **Drawing** *Testing AI for the Presence of Consciousness*).

First, we need to teach the program something, for example, to give full information about what red is. Then, with the help of devices, let the program ‘see’ the red color, and then ask it if it learned anything new. If so, it needs to be asked again to make sure that it is not lying about ‘what new things it learnt’. If the answer is ‘experience’ or ‘phenomenal experience’, then the program is most likely conscious.


Show something to the program that it knows everything about
Ask: "Have you learned something new?"
No
Yes
The program is not conscious
To make sure that the AI is not lying, you need to ask an additional question
"What specifically new did you learn?"
The program is lying
Anything but Subjective experience
Subjective experience
It's likely that the program is conscious

Testing AI for the Presence of Consciousness

CONCLUSION

However, this approach may be fundamentally wrong, as there are critics of Mary's experiment, who claim that she will not learn anything new when she leaves the room, because if she knows EVERYTHING about color, then she must also know how people experience color.

Testing artificial intelligence in the way described above is fraught with technical difficulties; it requires an artificial intelligence that claims to possess consciousness. In addition, ‘the ability to have subjective experience’ is far from the only definition of ‘consciousness’. However, this test overcomes the problem described by the *Chinese room* experiment.

References

1. *Locke J.* An Essay Concerning Human Understanding Book I. Pennsylvania : The Pennsylvania State University, 1999.
2. *Wittgenstein L.* Tractatus Logico-Philosophicus. London : Edinburgh Press, 1922.
3. *Turing A. M.* Computing Machinery and Intelligence // Mind. A Quarterly Review of Psychology and Philosophy. 1950. Vol. LIX., № 236. P. 433–460.
4. *Searle J. R.* Minds, Brains, and Programs // Behavioral and Brain Sciences. 1980. Vol.3, № 3. P. 417–424.
5. *Chalmers D. J.* The Character of Consciousness. New York : Oxford University Press, 2010.