Philosophical Considerations in Artificial Intelligence: Exploring Consciousness and Self-Awareness

Michał Puchyr

1 Introduction

Artificial Intelligence (AI) is transforming the programming landscape, introducing automation and advanced algorithms. This article will explore how currently AI is viewed in terms of consciousness and self-awareness.
Some people believe that AI will never be able to achieve consciousness and self-awareness, while others believe that it is just a matter of time or it even already happened. Lets take a deeper look how we can actually define consciousness and self-awareness and how we can measure it.

2 Defining Consciousness in AI

One of the fundamental ethical questions revolves around whether AI systems can possess **consciousness**. While AI lacks subjective experiences, emotions, and self-awareness in the human sense, it exhibits a form of consciousness in its ability to process vast amounts of data, recognize patterns, and make decisions. Understanding the limits of AI consciousness is crucial to establish responsible AI use.
Self-awareness is a key aspect of human consciousness. When considering AI, achieving true self-awareness akin to human consciousness remains elusive. However, AI systems can exhibit a form of self-awareness through continuous learning and adaptation. This self-awareness is a product of algorithms recognizing and adjusting to their own performance, limitations, and errors.

3 Criteria for AI Self-Awareness

The definition of consciousness can be stated as awareness of internal and external existence. However this description of this state is really vague and can be interpreted in many ways. Through the centuries philosophers have been trying to define consciousness and self-awareness. One of the popular definitions of consciousness is the knowledge of one's own existence, sensations, thoughts, surroundings, etc. You know that you exist, you know that you are thinking, you know that you are in a room, etc.
In every day life we trust each other people that they are conscious and self-aware. Why? Because we trust that they are like us, they know of their own existence and they are just like us. The problem with this criteria begins in terms of AI consciousness and self-awareness. It is not a organic being. It is a computer program which is strictly tied to computations and algorithms. It is not a human being, it is not a living organism.

4 The Turing Test

The Turing test is a test of a machine's ability to exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human. The test was introduced by Alan Turing in his 1950 paper, "Computing Machinery and Intelligence," while working at the University of Manchester. To pass the test, a computer must be able to fool a human judge into thinking it is human. The test is performed by having a human judge engage in a natural language conversation with one human and one machine, each of which tries to appear human. All participants are placed in isolated locations. If the judge cannot reliably tell the machine from the human, the machine is said to have passed the test.
To this day no machine has achieved this requirement.

5 The brain can be simulated

One of the arguments for AI intelligence achievieancy is that the brain can be simulated. If we assume that brain is just a biological version of computer made of neurons and synapses, then we can simulate it. Famous philosopher Hubert Dreyfus describes this argument as claiming that

"if the nervous system obeys the laws of physics and chemistry, which we have every reason to suppose it does, then... we... ought to be able to reproduce the behavior of the nervous system with some physical device"

6 Chinese room argument

In 1980 John Searle published a paper called "Minds, Brains, and Programs" in which he introduced the Chinese room argument. The argument is as follows:

"Suppose that I'm locked in a room and given a large batch of Chinese writing. Suppose furthermore (as is indeed the case) that I know no Chinese, either written or spoken, and that I'm not even confident that I could recognize Chinese writing as Chinese writing distinct from, say, Japanese writing or meaningless squiggles. To me, Chinese writing is just so many meaningless squiggles. Now suppose further that after this first batch of Chinese writing I am given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English, and I understand these rules as well as any other native speaker of English. They enable me to correlate one set of formal symbols with another set of formal symbols, and all that "formal" means here is that I can identify the symbols entirely by their shapes. Now suppose also that I am given a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch with the first two batches, and these rules instruct me how to give back certain Chinese symbols with certain sorts of shapes in response to certain sorts of shapes given me in the third batch. Unknown to me, the people who are giving me all of these symbols call the first batch a "script," they call the second batch a "story," and they call the third batch "questions." Furthermore, they call the symbols I give them back in

response to the third batch "answers to the questions," and the set of rules in English that they gave me, they call the "program." Now just to complicate the story a little, imagine that these people also give me stories written in English, which I understand, and they then ask me questions in English about these stories, and I give them back answers in English. Suppose furthermore that after a while, I get so good at following the instructions for manipulating the Chinese symbols and the programmers get so good at writing the programs that from the external point of view—that is, from the point of view of somebody outside the room in which I am locked—my answers to the questions are absolutely indistinguishable from those of native Chinese speakers. Nobody just looking at my answers can tell that I don't speak a word of Chinese. And yet I don't understand a word of Chinese."

We can see that the person in the room is just following the instructions and is not aware of the meaning of the symbols. The person is not aware of the meaning of the symbols, but the person is able to give answers to the questions. This argument is used to show that the program may be able to give correct answers to questions but only as a result of previously learnt instructions without understanding the meaning of it.

7 Conclusion

The question of AI consciousness and self-awareness remains a philosophical debate. While AI systems can exhibit a form of self-awareness, for many it is not comparable to human consciousness. The criteria for AI self-awareness is still a subject of scrutiny and debate. The Turing test is a popular method for determining AI self-awareness, but it has its limitations and its own critics. The future will show us if AI will ever surpass human intelligence and if it will ever be self-aware. The interesting question is, if AI will ever be self-aware, will it consider humanity as its creators and will it consider us as its parents, or will come to the conclusion that existence of humanity is a threat to its own existence and will try to destroy us.

Interesting words

- **accountability** - the fact or condition of being accountable; responsibility.
- **fairness** - impartial and just treatment or behavior without favoritism or discrimination.
- **consciousness** - the state of being awake and aware of one's surroundings.
- **self**-awareness - conscious knowledge of one's own character, feelings, motives, and desires.
- **algorithm** - a process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer.
- **implications** - the conclusion that can be drawn from something, although it is not explicitly stated.
- **subjective** - based on or influenced by personal feelings, tastes, or opinions.
- **elusive** - difficult to find, catch, or achieve.
- **akin** - of similar character.
- **indistinguishable** - not able to be identified as different or distinct.
- **imperative** - of vital importance; crucial.
- **scrutiny** - critical observation or examination.
- **squiggles** - a short line that curls and loops in an irregular way.
- **distinguishable** - able to be identified as different or distinct.
- **vague** - of uncertain, indefinite, or unclear character or meaning.