# 2020 Regional Philosothon

## Stimulus 2

Can a Robot be Conscious?

What exactly is artificial intelligence (AI)? AI is an inquiry which involves computer simulation of human capacities such as visual perception, problem solving, reasoning and using mental imagery. In other words, AI involves the attempt to design machines ("robots") which can perform tasks requiring intelligence. For example the AI program "General Problem Solver", developed in the 1950's by Herbert Simon and Alan Newell, was able to solve mathematical puzzles such as the Tower of Hanoi and also to play chess. More recently the chess-playing program "Deep Blue" beat the world champion Gary Gasparov.

A distinction is usually made between "weak" and "strong" AI. According to weak AI a computer or robot is simply a useful tool for testing theories about human psychological processes such as thought, reasoning, memory, problem solving and visual perception. However according to strong AI a computer or robot is literally able to remember, solve problems, think, reason and so on. Strong AI asserts that it is possible that a robot actually has a mind - is actually intelligent. A robot which is able to simulate visual perception, for example, can actually see. It is no mere simulation. Weak AI is relatively uncontroversial. It is clear that computers are at the very least a useful tool for the study of the workings of the human mind. The interesting question, then, is whether or not strong AI is a plausible view of artificial intelligence.

The problem: Is it possible to construct a robot which is conscious? Could AI ever design a robot which has a mind?

Some problems for strong AI

1. The Chinese Room: This is a hypothetical counterexample to strong AI. Suppose that you are in a room where your task is to answer questions about a story written in Chinese. The questions are in Chinese and the answers have to be given in Chinese but you can neither read nor write any Chinese at all. The only thing you have is an instruction manual written in English. The manual contains rules for associating certain Chinese characters (the questions) with other Chinese characters (the answers). The manual allows you to answer the questions in a way which satisfies a Chinese speaker. It is suggested that this example presents a problem for strong AI in the following way. Suppose it were possible to program a robot to do what you can do in the room – answer questions in Chinese in a way which satisfies a Chinese speaker. In fact, it is suggested, this is the most the robot would be able to

do. The point is that, like you, the robot does not understand Chinese. So the robot can be programmed to follow certain instructions and behave, for all intents and purposes, like a native speaker but can't literally understand anything. It can go through the motions of understanding but that doesn't amount to actual understanding.

Is there a reply on behalf of strong AI? Maybe this: The Chinese Room example does not make clear exactly what you and the robot are able to do. The claim is that merely being able to follow the instruction manual (which you and the robot can do) is not sufficient for an understanding of Chinese. But what if the robot were able to respond to questions in a sufficiently flexible and inventive way? The more the robot is able to do this the more we are inclined to say that it does understand Chinese. Thus if it were possible to program the robot to follow instructions in a flexible and inventive way, then that may be enough for actual robot understanding. Strong AI may survive the Chinese Room.

2. No originality: A computer does just what it is programmed to do. It doesn't do anything new, original or unpredictable. People use initiative and can think and act in novel and unpredictable ways. Thus a robot can never actually simulate a human being since humans have consciousness and rationality.

Question: How do we know that a robot can't (one day) be programmed to act in novel and unpredictable ways? Is this a matter of clever engineering or is there a fundamental principle which excludes the possibility of a rational and conscious robot?

3. The frame problem: It is too hard to give computers the sort of commonsense knowledge we all have and which is necessary to get by in the world. The computer scientist Roger Schank gives the following example. Mary goes into a restaurant and orders a hamburger. It is delivered burnt to a crisp. Mary storms out of the restaurant without paying the bill. Did Mary eat the hamburger? Most people would realise that she did not, but this fact is not actually stated in the example. It is part of our background knowledge – our "commonsense". The problem for AI is that it seems impossible to program this sort of background knowledge. The store of information is too large (and is being continually updated) and it is simply too difficult to program a computer to pick out the relevant information in a particular circumstance.

4. No emotions: A computer can be programmed to perform logical tasks such as mathematics, chess and problem solving but not to feel emotions such as anger, love, fear and anxiety. These emotions involve certain basic feelings and can't be represented formally and symbolically. Also, our emotions seem tied to our biochemistry. Could this be recreated in a robot? Perhaps biochip technology will be possible in the future. This involves imprinting circuits on a protein molecule rather than on silicon.

5. No point of view: A robot doesn't have a point of view i.e. there is no such thing as a "what it is like to be robot", whereas there is such a thing as a "what it is like to be a person" or a "what it is like to be a dog".

Maybe biochip technology will produce a class of robots which have such a point of view.

R. Neurath