Assignment 01

Summary on Philosophy of Artificial Intelligence

Dubi Sao , 19111025

5th Semester , Biomedical Engineering

The Philosophy of AI is mainly concerned with understanding AI in terms of ethics, consciousness, epistemology, free will and intelligence. It's implications for knowledge is also an area of interest for researchers. Associating philosophy with AI is a vital part of its emergence as it allows us to answer various questions like can a machine act intelligently? If yes, is it's intelligence comparable to that of a human being? How similar is it to humans in terms of having a mind , consciousness and mental states? Does it have feelings?

Over the years many propositions have been put forth by researches to define terms like "intelligence" , "consciousness" and "machine". For example, The Turing's polite convention stated that if a machine behaves as intelligently as a human being, then it is as intelligent as a human being. John Searle's strong AI hypothesis proposed that The appropriately programmed computer with the right inputs and outputs would thereby have a mind in exactly the same sense human beings have minds. The Dartmouth proposal stated that every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it.

The approach of Philosophy of AI can be broadly understood by asking the following main questions:

1. Can machine display general intelligence?

To answer this question it is first necessary to clearly define what intelligence is. According to Alan Turing intelligence can be defined using a simple test called the Turing test which states that if a machine answers any question using the same words as a human would, then the machine can be considered intelligent. One of the criticisms that the test received is that it is more concerned with measuring the humanness of the machine rather than its intelligence and hence it fails. Another way in which intelligence can be defined is in

terms of Intelligent agents. An agent can be described as an entity that perceives its environment and acts on it to achieve a specific goal. According to the intelligent agent definition,"If an agent acts so as to maximize the expected value of a performance measure based on past experience and knowledge then it is intelligent."

Several arguments has been made to prove that a machine can display intelligence. One of them, given by Hubert Dreyfus, points out that we should be able reproduce the nervous system of humans and hence simulate a complete brain since the nervous system follows the laws of physics and chemistry. Allen Newell and Herbert A. Simon argued that human thinking is similar to symbol manipulation and hence is a criteria enough to define intelligence thus claiming that machines can be intelligent too. On the contrary researchers like Kurt Gödel argued that humans have a very consistent system of mathematical reasoning which are beyond what a machine could ever duplicate. Similarly Hubert Dreyfus argued that human intelligence are based on implicit skills and unconscious reasoning which can only be understood but not captured in formal rules. However much progress has been made since these arguments which has led to simulation of unconscious reasoning and learning in machines.

2. Can a machine have a mind, consciousness, and mental states?

Philosophers, neuroscientists and cognitive scientists define consciousness as thoughts in the head like perception,dream, an intention etc. There are various arguments that refute the notion that machine can have consciousness given by renowned researchers. John Searle's Chinese room experiment concludes that for a machine to have consciousness, it should have the physical-chemical properties of an actual brain. Similar arguments were made by Gottfried Leibniz and Ned Block. Several points were made to discount Searle's Chinese room experiment like the systems and virtual mind replies , the robot reply, the brain simulator reply and many more.

3.Is thinking a kind of computation?

Computationalism or computational theory of mind claims that human intelligence derives from a form of calculation like arithmetics. This theory is similar to the physical symbol system hypothesis and was supported by philosopher like Hobbes who claimed that reasoning is nothing but reckoning. These hypothesis implies that artificial intelligence is possible.

Other important points that can be questioned when discussing philosophy of AI are the connection of machines with emotions, awareness, creativity, behaviour and various other human characteristics.