Philosophy of Artificial Intelligence

Krishna Agrawal, 19111031

July 19, 2021

Introduction

The philosophy of artificial intelligence is a branch of the philosophy of technology that explores artificial intelligence and its implications for knowledge and understanding of intelligence, ethics, consciousness, epistemology, and free will.

Propositions in Philosophy of AI

• Turing's "polite convention: If a machine behaves as intelligently as a human being, then it is as intelligent as a human being.
• The Dartmouth proposal: "Every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it."
• Allen Newell and Herbert A. Simon's physical symbol system hypothesis: "A physical symbol system has the necessary and sufficient means of general intelligent action."
• John Searle's strong AI hypothesis:"The appropriately programmed computer with the right inputs and outputs would thereby have a mind in exactly the same sense human beings have minds."
• Hobbes' mechanism:"For 'reason' ... is nothing but 'reckoning,' that is adding and subtracting, of the consequences of general names agreed upon for the 'marking' and 'signifying' of our thoughts..."

The Philosophy of Artificial Intelligence Attempts to Answer:

1. **Can a machine display general intelligence?**
   The basic position of most AI researchers is "Every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it."

   Intelligence

   - Turing test: Turing's test extends: If a machine acts as intelligently as a human being, then it is as intelligent as a human being.One criticism of the Turing test is that it only measures the "humanness" of the machine's behaviour, rather than the "intelligence" of the behaviour.
   - 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". Were an "agent" is something which perceives and acts in an environment and "performance measure" defines what counts as success for the agent.

   Arguments that a machine can display general intelligence

   • The brain can be simulated

   • Human thinking is symbol processing: In 1963, Allen Newell and Herbert A. Simon proposed that "symbol manipulation" was the essence of both human and machine intelligence.

   • Arguments against symbol processing:These arguments show that human thinking does not consist (solely) of high level symbol manipulation. Also arguments by Gödelian anti-mechanist arguments, Dreyfus.

2. **Can a machine have a mind, consciousness, and mental states?**
   John Searle defined "strong AI" as "A physical symbol system can have a mind and mental states." Searle defined "weak AI" as "A physical symbol system can act intelligently".

   (a) Consciousness, minds, mental states, meaning: "Consciousness" - an invisible, energetic fluid that permeates life and especially the mind. For philosophers, neuroscientists and cognitive scientists: everyday experience of having a "thought in your head", like a perception, a dream, an intention or a plan, and to the way we know something, or mean something or understand something.

   (b) Arguments that a computer cannot have a mind and mental states: • Searle's Chinese room: Searle goes on to argue that actual mental states and consciousness require "actual physical-chemical properties of actual human brains." • Related arguments: Leibniz' mill, Davis's telephone exchange, Block's Chinese nation and Blockhead • Responses to the Chinese room emphasize several different points.

3. **Is thinking a kind of computation?**
   The computational theory of mind or "computationalism" claims that the relationship between mind and brain is similar (if not identical) to the relationship between a running program and a computer.

   • Reasoning is nothing but reckoning.

   • Mental states are just implementations of (the right) computer programs.

4. **Other related questions?**

   • Can a machine have emotions? - If "emotions" are defined only in terms of their effect on behaviour, then emotions can be viewed as a mechanism that an intelligent agent uses to maximize the utility of its actions. Hans Moravec believes that "robots in general will be quite emotional about being nice people".

   • Can a machine be self-aware? - "Self-awareness" is used as a name for the essential human property that makes a character fully human. Turing strips away all other properties of human beings and reduces the question to "can a machine be the subject of its own thought?" Can it think about itself? Viewed in this way, a program can be written that can report on its own internal states, such as a debugger.

   • Can a machine be original or creative? - Turing reduces this to the question of whether a machine can "take us by surprise". It must be possible, for a computer that can represent ideas to combine them in new ways.

   • Can a machine be benevolent or hostile? - One issue is that machines may acquire the autonomy and intelligence required to be dangerous very quickly

   • Can a machine imitate all human characteristics? - Turing noted that there are many arguments of the form "a machine will never do X", where X can be many things, such as: Be kind, resourceful, beautiful, friendly, have initiative, have a sense of humour, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream.

   • Can a machine have a soul? - Alan Turing writes "In attempting to construct such machines we should not be irreverently usurping His power of creating souls, any more than we are in the procreation of children: rather we are, in either case, instruments of His will providing mansions for the souls that He creates."

Views on the role of philosophy

Questions like these reflect the divergent interests of AI researchers, cognitive scientists and philosophers respectively. Some scholars argue that the AI community's dismissal of philosophy is detrimental, while some philosophers argue that the role of philosophy in AI is underappreciated. Physicist David Deutsch argues that without an understanding of philosophy or its concepts, AI development would suffer from a lack of progress