AI and Human Cognition: Theories and Implications

Author: Daniel Moon

AS.140.396.01.FA21 Encoding Bias: Algorithms, Artificial Intelligence, and the History of Computing

Table of Contents:

Abstract

Whether or not machines could posess a human mind has been a debate ever since the advent of Artificial Intelligence. While scholars like Herbert Simon argued that there was no reason not to believe that machines couldn't possess a human mind, other scholars such as John Searle and Ulrich Neisser opposed this view by stating that there were mental contents inherently missing in machines that was required in order to replicate the human mind, of these mental contents being emotion. However, the implications of these two views were far-reaching as popular media from the 1960s onwards would start incorporating human-like programs into films and shows, as seen from the character HAL-9000 in the 2001: A Space Odyssey. The two tests that influenced these views are the Turing Test and the Chinese room experiment. The implications of these tests are still up for debate as the results of these experiments can be interpreted differently based on one's views and inherently have philosophical limitations. These tests aim to answer the fundamental question, "Is imitating a human mind enough to conclude that the object has a human mind?"

Intro

Can a machine think?

For many computer scientists, the answer is yes. If one defines "machines" as a physical system capable of performing certain functions, humans are, by definition, machines. Consequently, if humans are machines of a special biological kind and humans can think, machines can theoretically think as well given the right programs. Attributing the verb "think" to machines is commonplace in today's era; when AlphaGo contemplates its next move on a Go board or when a self-driving car calculates the fastest route to its destination, we can argue that they are "thinking," whether or not the said program is conscious or not.[1] [2]

The natural follow-up question to that, is "Can machines replicate the human mind?" In other words, is the human mind simply a program that can be deciphered and implemented into machines? Can a machine think like a human does?

This particular question, unlike the first, is a debate that has been ongoing ever since the advent of artificial intelligence. There are mainly two points of view that argue for both sides of this problem. The first is that yes, the human mind is simply a collection of problem-solving and pathfinding mechanisms with serial processors that carry out given goals, so a human mind can be replicated by virtue of a program.[3] The other side claims that no, machines as they exist today cannot replicate the human mind because its program is solely based on syntax, or programs, while human minds have mental contents, or semantics,[4] and are strongly interwoven with emotional experiences.[5]

[1] Puccetti, Roland. "On Thinking Machines and Feeling Machines." *The British Journal for the Philosophy of Science* 18, no. 1 (1967): pp. 41.

[2] Minsky, Marvin. "Why People Think Computers Can't." AI Magazine Volume 3 Number 4 (1982): pp. 4.

[3] Simon, Herbert. "A Theory of Emotional Behavior," Carnegie Mellon University Complex Information Processing (CIP) Working Paper #55, June 1, 1963: pp. 8 - 15

[4] Searle, John R. "Is the Brain's Mind a Computer Program?" *Scientific American* 262, no. 1 (1990): pp. 27.

[5] Neisser, Ulric. "The Imitation of Man by Machine." *Science* 139, no. 3551 (1963): pp. 1

The dichotomy between these two arguments presents a philosophical dilemma in identifying what it truly means to "think" and act like a human. Is it enough to simply imitate what a human does in order to be considered human? Or does the machine require an underlying consciousness?

Chapter 1

Attempts to decipher and replicate the human mind dates back to a scientist named Alan Turing. Widely regarded as the father of computer science, Turing theorized in 1950 that for any algorithm, there existed some Turing machine that could implement the said algorithm: the Universal Turing Machine. But now, what made this result so exciting? Well, what made it send shivers up and down the spines of the workers in artificial intelligence was the following thought: suppose that the brain was a Universal Turing Machine. In other words, what if brains were just like computers running programs? Turing's ground-breaking theory marked the birth of the field of cognitive computing which focused on mimicking human behavior and reasoning to solve complex problems, treating the human mind as a computer program.[6]

As cognitivists, Herbert Simon and Allen Newell began working on the Logic Theory Machine in 1955, a primitive form of AI that was designed to embody human problem-solving behavior in the domain of elementary logic.[7] The Logic Theory Machine was a resounding success; with modifications implementing heuristics (bounded rationality) to govern their calculation processes,[8] the program was able to prove 38 of the 52 proofs in

[6] Searle, John R. "Is the Brain a Digital Computer?" *Proceedings and Addresses of the American Philosophical Association* 64, no. 3 (1990): pp. 23.

[7] Dick, Stephanie. "Of Models and Machines: Implementing Bounded Rationality." *Isis* 106, no. 3 (September 2015): 626–34

in chapter 2 of the Principia Mathematica, with few of the proofs being more elegant than the ones done by humans.[9]

This success would prompt Simon and Newell to go a step further and hypothesize the problem-solving mechanisms of the human mind with their proposal of the General Problem Solver in 1961. In this defining work, the two postulated that human problem solving was governed by a set of elementary information from a set of elementary information processes, and that the human mind was simply a symbol-manipulating device and nothing more.[10] With these findings, the two claimed that "the process of [human] thinking [could] no longer be regarded as completely mysterious," coming one step closer to the possibility of a machine possessing a human mind.

However, dissenters of this cognitivist view refuted these claims by stating that Simon and Newell completely disregarded one extremely influential factor in human cognition: emotion.

It is common logic that emotions have a direct, or at least an indirect connection to our everyday activities, whether it be taking a test, walking down a sidewalk, or learning a new skill. The interrelation of emotion and intelligence is especially apparent in infants whose first accommodation to basic features of the world such as time, distance, and causality are strongly interwoven with emotional experiences and in juveniles who learn to increasingly cope with

[8] Simon, Herbert A. "A Behavioral Model of Rational Choice." *The Quarterly Journal of Economics* 69, no. 1 (1955): pp. 104.

[9] McCorduck, Pamela (2004), *Machines Who Think* (2nd ed.), Natick, MA: A. K. Peters, Ltd.

[10] Newell, Allen, and Herbert A. Simon. "Computer Simulation of Human Thinking." *Science* 134, no. 3495 (1961): 2011–17.

their emotions through personal experiences.[11] Ulrich Neisser, being one of the dissenters of the Cognitivist view, argued in 1961 that computer programs, in order to truly think like a man, would need to be similarly endowed with powerful internal states comparable to that of emotion since human thinking was intimately associated with emotions and feelings. He added that these internal states must have a significant influence on processing information at the earliest stages of learning, just like that of an infant or a teen.[12] For Neisser, the cognitivists failed to account for this relationship and foolishly assumed that man's intelligence was independent of the rest of human life.

However, Herbert Simon would later respond to Ulrich Neisser in 1963 with his paper "A Theory of Emotional Behavior," where he proposed that emotion was simply an interruption mechanism to reassess one's hierarchy of goals based on certain sensory stimuli. In his theory, emotion was not a derivative of a higher-order intellectual feature exclusive to humans, but rather one that could be easily implemented in computers to simulate human thinking. He stated that information-processing machines could readily be endowed with precisely the properties that Neisser listed as characterizing human thinking. Thus, when it came to the cognitivists, emotion was not an insurmountable barrier forever separating them from their ultimate goal, but just another feature that had to be mechanically implemented in order to program a perfect computer model of the human mind. Put simply, there were no critical differences between man and a machine that could distinguish one from the other.[13]

[11] Mackaye, David L. "The Interrelation of Emotion and Intelligence." *American Journal of Sociology* 34, no. 3 (1928): 451–64.

[12] Neisser, Ulric. "The Imitation of Man by Machine." *Science* 139, no. 3551 (1963): 193–97.

Chapter 2

The possibility of there existing a machine that could think and act like a human was one that was quickly endorsed by the popular media from the 1960s onwards. The first media portrayal of a program that possessed human intelligence, albeit imperfectly, was the HAL-9000 featured in 2001: A Space Odyssey. In the movie, HAL-9000 decides to kill all the astronauts after realizing that the astronauts were going to disconnect his cognitive circuits. He ultimately succeeds in doing so, going on to kill everyone except one man, David Bowman. In his final moments before David completely shuts him down, HAL-9000 states:

> "I'm afraid, David."[14]

The presence of human emotions in this cold, soulless machine evokes an uneasy dissonance that poses a question to the audience in attendance: is HAL truly afraid? Does HAL know what it means to be afraid of its impending doom?

Again, we are entering the realm of philosophy with this question. Is the portrayal of fear by HAL-9000 enough to guarantee his emotional cognition and therefore his human cognition?

This particular question leads us back to the question I posed at the start of this video. Is imitation enough for a machine to be considered having a human mind?

[13] Boden, Margaret A. "How Artificial Is Artificial Intelligence?" Edited by F. J. Crosson, B. Meltzer, and D. Michie. *The British Journal for the Philosophy of Science* 24, no. 1 (1973): 61–72.

[14] Kubrick, Stanely, director. 1968. *2001:A Space Odyssey*. Metro-Goldwyn-Mayer

In order to formulate our answer, we need to discuss two very important tests in computer science: the Turing test and the Chinese room experiment.

The Turing test, created by Alan Turing in 1950, is a method of inquiry in AI for determining whether or not a computer is capable of exhibiting intelligent behavior equivalent to or indistinguishable from that of a human.[15]

In the original imitation game, 3 players are involved. Player A is a man, player B is a woman, and player C, who plays the role of the interrogator, is of either sex. In this game, player C is unable to see either player A or player B, and can only communicate with them through written notes. By asking questions to player A and player B, player C tries to determine which of the two is the man and which is the woman. Player A's role is to trick the interrogator into making the wrong decision, and player B’s role is to assist the interrogator into making the right decision.

The Turing test makes one modification into this game and replaces player A with a computer. The goal of the interrogator, now, is to determine which is a computer and which is a man. The computer is said to have passed the test if the interrogator cannot reliably distinguish between the two.

[15] A. M. Turing (1950), “Computing Machinery and Intelligence.” Mind 49: pp. 433 - 450

Just like the original name of the game suggests - if the computer can imitate the thoughts of a human well enough to pass the test, it is said to exhibit equivalent intelligent *human* behavior. The test certainly resonates with behaviorist ideals of assessing human intelligence as it confines its observations exclusively to the external behaviors of computers (i.e. their responses to the questions). In other words, the test tempts people to think that if machines can flawlessly imitate a human mind, it must possess a human mind.

Chapter 3

A direct antithesis to the Turing test comes from the Chinese Room experiment, created by John Searle in 1980 to dispute the implications of the Turing test.

In this thought experiment, Searle begins with a hypothetical premise: suppose AI research had successfully created an AI that behaves as if it understood Chinese. In fact, it performs its job so convincingly that it passes the Turing test, convincing the human Chinese speaker that the program itself is a live Chinese speaker, not a computer.

The question Searle poses is this: does the AI truly understand Chinese in a literal sense? Or is it going through the motions of simulating the ability to understand Chinese? In the context of HAL, is HAL afraid in the most literal sense or is HAL simply simulating the ability to fear?

Searle then supposes he is placed in a room with an English version of the said computer program. He receives Chinese letters through a window, processes them with the given instructions, then produces Chinese characters as output. To the person outside of the room, Searle is demonstrating an intelligent conversation in Chinese even though Searle himself does not understand a single letter of Chinese. Searle asserts that this is the exact scenario that unfolds in computer programs as well; the computer does not understand nor truly comprehend Chinese, it is simply manipulating symbols to give an illusion that it does.[16]

Similarly, no one believes that Searle, in his room, understands Chinese even if he produces convincing results. So why should anyone believe that a computer program will be any different? This thought experiment implies that even if the Turing test is passed, that fact alone cannot guarantee mental contents of the machine behind the curtain. For Searle, HAL is not truly afraid - he is imitating as if it does.

However, this does not mean that Searle's experiment is foolproof. In the field of epistemology, or the study of knowledge, there is a question called "The Problem of Other Minds."

It states: given that I can only observe the behaviors of others, how can I **know** that others have minds? Rewording this problem, we can ask: given that I can only observe the behavior of a computer, how can I know that it **doesn't** have a mind?

In everyday life, we never consider the problem of other minds when interacting with people; we just assume that everyone does as sort of a polite convention since we can never determine this truth ourselves.

[16] Searle, John R. "Is the Brain's Mind a Computer Program?" *Scientific American* 262, no. 1 (1990): 2

The bottom line is that this problem also applies to machines as well. Realistically, there is no way of determining whether a machine has a human mind unless we become the machines themselves. And obviously, we can't. While the Chinese room experiment denies the possibilities of there being a human mind inside a program and asserts that imitating is fundamentally different from understanding, how are we to distinguish the difference?[17] Even if HAL did not understand what it meant to be afraid, we can never dismiss the possibility that HAL, in fact, did have a human mind. The inherent uncertainty in determining whether or not a man or a machine possesses a mind suggests that we can never know the true answer to this question.

The only thing that we can observe from a machine is its behavior. And, if the machine perfectly thinks and acts like a human does, who are we to dismiss the possible existence of their minds?

On the surface, this whole conclusion might seem disingenuous. After all, this philosophical dilemma goes in a loop, leaving behind questions that are fundamentally open-ended for all.

[17] Nilsson, Nils. "A Short Rebuttal to Searle," 1984. .

Conclusion

Before we end, however, let's consider Mary Shelly's magnum opus, Frankenstein. In her Sci-Fi novel, Dr. Frankenstein brings to life a creature named "the Monster," but immediately runs away in fear, frightened of his own creation. The Monster, stripped of all his humanity, matures to become just like his name: a Monster.[18] Do you think the results would have been different had Dr. Frankenstein acknowledged that the Monster was a conscious entity and treated him as such?

As we enter an era where science fiction becomes a reality, we are constantly reminded of the responsibilities that follow. The ethics involved in creating a cognitive AI is one that certainly needs more research and consideration as we prepare for the inevitable. The discussions presented in this essay deals not with whether such AI is possible, but instead what its implications are.

Can a machine truly feel or love like us all? Depending on your answers, the path we undertake will dramatically differ.

And when the future becomes the present, will we be frightened of our own creation just like Dr. Frankenstein, carelessly abandoning all responsibility? Or will we set aside our differences in opinion and be able to cater to the unknown?

[18] John Turvey and Mary Shelley, *Frankenstein* (Harlow: Longman, 1998

Bibliography

1. Kubrick, Stanely, director. 1968. *2001:A Space Odyssey*. Metro-Goldwyn-Mayer
2. Neisser, Ulric. “The Imitation of Man by Machine.” *Science,* vol. 139, (1963): 193-197.
3. Boden, Margaret A. “How Artificial Is Artificial Intelligence?” Edited by F. J. Crosson, B. Meltzer, and D. Michie. *The British Journal for the Philosophy of Science* 24, no. 1 (1973): 61–72.
4. Mackaye, David L. “The Interrelation of Emotion and Intelligence.” *American Journal of Sociology* 34, no. 3 (1928): 451–64.
5. Minsky, Marvin. “Why People Think Computers Can’t.” *AI Magazine*, vol. 3 no. 4 (1982).
6. Newell, Allen, and Herbert A. Simon. “Computer Simulation of Human Thinking.” *Science* 134, no. 3495 (1961): 2011–17.
7. Puccetti, Roland. “On Thinking Machines and Feeling Machines.” *The British Journal for the Philosophy of Science* 18, no. 1 (1967): 39–51.
8. Searle, John R. “Is the Brain’s Mind a Computer Program?” *Scientific American* 262, no. 1 (1990): 25–31.
9. Simon, Herbert A. “A Behavioral Model of Rational Choice.” *The Quarterly Journal of Economics* 69, no. 1 (1955): 99–118. https://doi.org/10.2307/1884852.
10. Dick, Stephanie. “Of Models and Machines: Implementing Bounded Rationality.” *Isis* 106, no. 3 (2015): 623–34. https://doi.org/10.1086/683527.
11. Searle, John R. “Is the Brain a Digital Computer?” *Proceedings and Addresses of the American Philosophical Association* 64, no. 3 (1990): 21–37. https://doi.org/10.2307/3130074.

12. Turvey, John, and Mary Shelley. *Frankenstein*. Harlow: Longman, 1998.

13. A. M. Turing (1950), "Computing Machinery and Intelligence." Mind 49: pp. 433 – 450

14. McCorduck, Pamela (2004), *Machines Who Think* (2nd ed.), Natick, MA: A. K. Peters, Ltd.

15. Simon, Herbert. "A Theory of Emotional Behavior," Carnegie Mellon University Complex Information Processing (CIP) Working Paper #55, June 1, 1963: pp. 8 - 15