Playing the Imitation Game: Artificial Intelligence and Human Mind

Elise Jing

05/05/2017

Introduction

The attempt to understand the human mind is often intertwined with the attempt to build machines that have “minds”. Although standing as distinct disciplines, theories of mind and theories of artificial intelligence are aware of, and have inspired each other: the classical computational theory of mind was heavily influenced by the theory of computation; artificial neural networks, being the current most powerful tool for artificial intelligence, started as simple models imitating real neurons (Rosenblatt, 1957).

However, underneath this surface, there lies much deeper, complicated discrepancies and (perhaps) resolutions. Is human intelligence fundamentally the same thing as artificial intelligence? Shall we approach artificial intelligence by simulating human intelligence? Finally, if we create a perfect artificial intelligence, will it have consciousness, mental states, feelings, love and desire — in a word, a mind — as we (questionably) do? These questions have long been the topics of debate for philosophers of artificial intelligence and philosophers of mind, and with the most recent wave of AI sweeping industries, investors and media, it seems not likely that they will settle anytime soon.

In this article, I would like to provide a brief review on the relation between AI and mind. I will start by going over the history of relevant theories starting from mid-20th century, then discuss some major positions on the questions presented above.

History

The theory of computation started in 1930s by pioneers such as Alan Turing and Alonzo Church. The most iconic event was the invention of the Turing machine, an abstract

machine that can carry out general computation by manipulating symbols. This model and the Von Neumann architecture together became the foundations of computer science, and soon after, researchers found the ambition to create intelligence with a computer. The Dartmouth workshop in 1956 was considered by many to be the start of artificial intelligence as a field (Dartmouth workshop, 2017). Meanwhile, going along another direction, psychologists started to think about whether the Turing machine can work as a model of the human mind. This idea, soon developing into the computational theory of mind, played a central role when cognitive science rose as an interdisciplinary area, making itself the “default” paradigm for the following fifty years.

However, while the Turing machine model was widely accepted, what it actually meant varied between schools of thoughts. In some senses, a computer was only used as a metaphor for how the brain functioned, while in others, the brain was thought to be literally carrying out computation. The latter approach was accepted by Hilary Putnam (Putnam 1961) and introduced into the philosophy of mind, constructing the widely accepted theory: machine functionalism. In this view, mental states are functional states; the brain moves between functional states similar to how a Turing machine moves between machine states.

About one decade later, Jerry Fodor introduced another main thread of the computational theory, the representational theory of mind (Fodor 1975, 1981). According to him, the brain as a Turing-style machine operates on mental symbols, which represents both basic and complex concepts about the world.

Notably also during the same time period, the artificial neural network model was developed, although not widely known until many years later. The influence goes in the

other direction in this case: artificial neural networks were influenced by psychology and neuroscience. When creating the Perceptron, Frank Rosenbratt, a psychologist, was said to be inspired by real neurons (Rosenblatt, 1957). The closely associated Back Propagation algorithm, without which artificial neural networks would not work well, was also loosely inspired by Freud's psychological theories (Werbos 1974).

It may not be a coincidence that as the computational theory of mind becomes "the only game in town", the symbolic approach to AI was also at the best of its days, before going into the first AI winter in mid-1970s. Leading researchers at CMU, MIT, and other institutes believed that symbolic is the ultimate approach to AI, claiming that it will completely solve the problem of creating AI in 20 years (Minsky 1967). They even go further to argue that symbol manipulation is at the root of all intelligence, and when combined with proper physical components, "any physical symbol system of sufficient size can be organized further to exhibit general intelligence" (Newell and Simon, 1976). As we all know, the task was soon proved to be much more difficult than they optimistically thought, and by 1980s symbolic AI was no longer the prevailing paradigm.

It may also be not a coincidence that connectionism, challenger of the classical computational theory in influence, also rose at this point. In this case, however, the psychologists didn't directly borrow ideas from computational theorists. Rather, the paradigm found its way into both disciplines, fueled by neuroscience and computer science together (Medler, 1998). Connectionism is a major diverging from the symbolic approach: connected units (more specifically, neurons) operate under certain rules, and what were traditionally thought to be stored as symbols — memory, concepts, propositions — are claimed to be contained in these connections.

More alternatives to the computational theory of mind have been proposed since the late 20th century, although none was strong enough to be a new dominating paradigm. Meanwhile in artificial intelligence, deep learning has shown capabilities that people once thought machines couldn't have; but the driving force is mainly an increase of computing power. Compared to the perceptrons in 1960s, the advanced neural network models are still based on the same core ideas: layers and back propagation. Many improvements have been made, such as increasing the number of layers, finding good cost functions, and adding input and output gates. However, the core ideas have not fundamentally changed.

This brief overview has attempted to outline the relation and interaction between artificial intelligence and philosophy of mind. It can be found that this relation is complex: rather than having unidirectional influence and sharing superficial concepts, the two fields shared important ideas and experienced paradigm shifts at approximately the same time. Additionally, many great scientists in one area also had substantial contributions to the other, for example, Alan Turing and Herbert Simon. These all imply that the problem of understanding human mind and that of understanding artificial intelligence are fundamentally intertwined. In the following sections, I will examine some major topics on the connections between these two fields.

The ontological question

It is noteworthy that while pioneers of AI considered it AI's goal to show intelligence like that of humans' (for example, Simon 1965), current researchers often just focus on

“intelligence” and not “human”. This may imply a shift of attitude on an ontological question: are human intelligence and artificial intelligence fundamentally the same?

To consider this question, we have to start with a definition for intelligence. A common definition is that an intelligent being should be able to perceive and react to the environment, which is what human and many animals are capable of; but according to this, plants such as sun flowers will also have intelligence. Stricter definitions involve the ability to think and reason, but when it comes to machines, these abilities are hard to identify. Can we say that a machine learning program is thinking and reasoning when it makes classifications?

This is exactly the question Turing asked more than 60 years ago (Turing 1950), and he got around it by proposing a “polite convention”: if a machine acts as intelligently as a human, then it is as intelligent as a human. The well-known Turing test is based on this premise. However, although it has been popular as both a thought experiment and a real experiment, we should be aware that the original question had been modified in a behaviorist-style way. A machine acting as if it is thinking does not necessarily imply that it is thinking, as indicated by the problem of other minds.

As Turing pointed out, to face the question directly requires us to go into the definition of thinking. If as the classical computational theory of mind suggests, thinking is the manipulation of symbols, then it seems reasonable to consider a computer to be thinking. The latter may be what the symbolic AI supporters have in mind: when Newell and Simon described their model of physical symbol systems, they did not limit it to either computers or human, because “the symbolic behavior of man arises because he has the characteristics of a physical symbol system” (Newell and Simon, 1975), and

computer and human thinking are, in their roots, the same thing. If instead as connectionists say, thinking is the result of activities of connected neurons, then it's likely not going to be found in the computers we currently have.

This suggests us to turn once more to the long-lasting debate between classical computational theory of mind and connectionism. However, many have also suggested that the two paradigms are not mutually exclusive. A neural network can be implemented in a computer that does symbolic manipulation; oppositely, Turing-style computation can also be achieved using a neural network (Rescorla, 2015). Regarding physical implementations, although currently we don't have computers made of silicon neurons, in the future they may come into being. Reversely, it has been attempted to simulate the whole human brain in a supercomputer, mapping over a hundred billion neurons and their connections. Although the project has not been successful (Dvorsky 2014), it implies the possibility of completely "transplanting" a brain to a computer. In that case, is the simulated brain or the computer thinking? Perhaps the ontological question will become unnecessary at all, if the boundary between brains and machines gets obscure.

Approaching AI

Naturally, the question discussed above leads to a relevant one: is it necessary to simulate the human minds to create artificial intelligence? A famous analogy is often mentioned — do we have to know how birds fly in order to make airplanes? On one hand, people studied how birds fly to understand aerodynamics; on the other, most planes do not actually fly in a wing-flapping way.

This question partly depends on a premise — do we know how the human mind works in the first place? Depending on the answer to these two questions, various positions may be taken. Some symbolic AI supporters would say “yes” to both: for example, Newell and Simon intentionally mimicked human’s problem solving behavior, found through psychological experiments (Daniel 1993), to create their algorithms. Some others did not think it was necessary to simulate human behavior, and created algorithms for AI that might or might not behave as some humans would have. Most statistical approaches to AI are thought to be different from human thinking; however, artificial neural network models are often considered an analogy to human’s learning process.

It may be worthwhile to go deeper into this example. A type of neural networks excelling at computer vision tasks is the convolutional neural network model. Intuitively, a convolution of two matrices is a way of “blending” them. In convolutional neural networks, it is applied to draw a representation from every local region of the image, therefore reducing the number of parameters to train. When the features are passed down hidden layers, this process is repeated. The primary purpose of applying this algorithm is to optimize the model. However, when used in computer vision tasks, it appears to work similarly to how the human visual cortex processes images: after sufficient training, a layer will often be found to specialize in identifying the outline of objects, a second layer will specialize in processing colors, etc. In a 2014 paper, Cox and Dean suggested a rough mapping between layers in a convolutional neural network and regions in a visual cortex (Cox and Dean, 2014).

In this example, artificial intelligence works like the human brain to a certain degree (although major differences still exist, for example, the brain doesn't have a back propagation mechanism). However, there was no intentional simulation when the model was created. Inspiration from the brain, although may have existed, is superficial. Similarly, most neural network models are not directly simulating human's cognitive processes.

But these models perform well, being able to recognize cats in pictures (Le, 2013), find semantic similarities, and play Go as well as humans do. With such success, it seems natural that most AI researchers nowadays do not care much about how human mind works. However, there are still people who insist that the only way to create "full" AI is by replicating human mind in a machine. This leads us to more consideration about minds of machines.

The problem of consciousness

Despite the recent success of deep learning with neural network models, some people, for example Douglas Hofstadter, still insists that this is fundamentally wrong, because a machine like IBM's Watson "is finding text without having a clue as to what the text means. In that sense, there's no intelligence there." (Herkewitz, 2014).

This easily reminds us of Searle's Chinese room case. The thought experiment was used to refute the strong AI hypothesis, that a properly programmed computer handling inputs and outputs as a human does has the same mind as the human does. In Searle's argument, a computer that matches Chinese output with English input according to a set of rules cannot be said to understand Chinese (Searle 1980). He further concluded

that no physical symbol system could ever be said to have a mind. There has not been a full consensus on this argument, but many consider it to be sound.

Searle's attack being directed to the symbolic AI approach, we may wonder if it's the case with neural network models, which are famous for being "black boxes" that don't let people understand what is happening between their layers. Will a complicated neural network have a mind? An interesting reply to Searle's argument is the emergence hypothesis. In a complex system, the properties of a system as a whole are not the same as adding up the properties of the parts (Simon, 1962). Some suggest that if an artificial intelligence is complex enough, a mind may emerge spontaneously by interaction between its parts. This is a popular imagination in contemporary culture: AIs in many cyberpunk fictions gain consciousness in this way. But in this way, the hard problem of consciousness is still not solved. Just as with ourselves, we observe the existence of consciousness, but don't know how it rises from a lot of neurons (or a lot of circuits).

Furthermore, the relation between mind and general intelligence is also unsolved. Considered to be the "ultimate goal" of AI, general intelligence is usually defined as the ability of performing a wide range of intelligent actions: learning, reasoning, working towards a goal, and more — in a word, things that a normal human adult can do. Naturally, the question is often asked: does a being must have a mind to have general intelligence, or is mind not relevant to having general intelligence at all?

If we take the first option, the current mainstream approach to AI will be fundamentally wrong. However, in the brief historical review above, it can be observed that unless development of technology becomes stilled, there will be little motivation for

researchers to push for paradigm shifts. When symbolic AI and expert systems are making significant advances, few people considered them to be wrong approaches (although we should acknowledge that at that time, the computational power was also not sufficient to support other approaches such as neural networks). With the current success of deep learning models, it may be reasonable to infer a similar situation: unless at a certain point the advance of such models hit a dead end, it will be unlikely for people to switch to a different approach.

The second option, naturally, is favored by most current researchers. Why is it necessary to have a mind in order to display general intelligence? It seems not hard to put a Chinese room, a German room, a Russian room,… together to create a general machine translator that doesn't understand any of these languages, which is essentially Google Translate. Similarly, it may be possible that by putting the specific artificial intelligences together, the AI built will display general intelligence without anything like mind or intentionality.

This invites us to go back to Turing's hypothesis from a slightly different view. If an AI displays general intelligence as a human does, how do we know if it has a mind or not? In the famous thought experiment of philosophical zombies, it is claimed impossible to tell if anyone other than oneself has a mind, or is a complicated automata that behaves just like it has a mind: take the example of pain, it can be programmed to wince when approaching fire as if it feels pain, while it actually has no feelings of pain. If we are unable to tell whether another person has a mind, are we able to judge whether an AI has a mind?

If we take this claim, the boundary between humans and computers is again obscured: a human can now be considered as a computer that successfully passed the Turing test. Moreover, being able to behave like a human will have nothing to do with having a mind. It seems yet another support for the claim that having a mind is not relevant to displaying general intelligence. However, most people are likely not comfortable to be skeptical enough to accept a world walked by zombies. If in the future a computer exhibiting perfect general intelligence is created, we may be not able to tell if it has a mind at all. But this will surely jeopardize the belief held by many, as expressed in the movie *Ghost in the Shell:* that what distinguishes a human from a machine is the awareness of a self: a “ghost”, a mind, a soul.

Conclusion

The history of interactions between theories of mind and theories of AI is long and subtle. This paper provides a brief overview of how the two areas has influenced each other since their establishment in mid 20th century: how they evolved, borrowed ideas from each other, and experienced paradigm shifts at similar times.

I have also introduced and discussed several major topics on the relation between AI and human mind. The first one is the ontological problem: are human intelligence and machine intelligence essentially the same thing? The second question is about approach: is it necessary to simulate human mind in order to create AI? Lastly, I focused on the minds of machines — is having a mind necessary to displaying general intelligence? In our endeavor to creating artificial general intelligence, how will it change the way we understand the world and ourselves?

It is beyond the scope of this paper to propose answers for these questions, and most of them are not likely to reach any resolution soon. Some neuroscientists estimate it would take another 100 years or more to understand how the human brain works (Epstein, 2016), and I hope that researches on AI can help us reaching this goal.

Reference

Cox, D. D., & Dean, T. (2014). Neural networks and neuroscience-inspired computer vision. Current Biology, 24(18), R921-R929.

“Dartmouth workshop”. (2017, April 23). Retrieved from https://en.wikipedia.org/wiki/Dartmouth_workshop

Dvorsky, G. (2014, July 10). “Europe’s $1.6 Billion Human Brain Project Is On The Verge Of Collapse”. Retrieved from http://io9.gizmodo.com/europes-1-6-billion-human-brain-project-is-on-the-verg-1602991993

Epstein, R. (2016, May). “The empty brain”. Retrieved from https://aeon.co/essays/your-brain-does-not-process-information-and-it-is-not-a-computer

Fodor, J., 1975, The Language of Thought, New York: Thomas Y. Crowell.

Fodor, J., 1981, Representations, Cambridge: MIT Press.

Herkewitz, W. (2014, February). “Why Watson and Siri Are Not Real AI”. Retrieved from http://www.popularmechanics.com/science/a3278/why-watson-and-siri-are-not-real-ai-16477207/

Le, Q. V. (2013, May). Building high-level features using large scale unsupervised learning. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on (pp. 8595-8598). IEEE.

McCulloch, W. and W. Pitts, 1943, “A Logical Calculus of the Ideas Immanent in Nervous Activity”, Bulletin of Mathematical Biophysics, 7: 115–133.

Medler, D. A. (1998). A brief history of connectionism. Neural Computing Surveys, 1, 18-72.

Minsky, Marvin (1967). Computation: Finite and Infinite Machines. Englewood Cliffs, N.J.: Prentice-Hall. ISBN 0-13-165449-7. Quoted in Crevier, Daniel (1993), AI: The Tumultuous Search for Artificial Intelligence, New York, NY: BasicBooks, ISBN 0-465-02997-3.

Newell, A., & Simon, H. A. (1976). Computer Science as Empirical Enquiry: Symbols and Search, 19 COMM. ASS'N FOR COMPUTING MACHINERY, 113.

Putnam, Hilary. 1961. "Brains and Behavior", originally read as part of the program of the American Association for the Advancement of Science, Section L (History and Philosophy of Science), December 27, 1961. Reprinted in Block (1980).

Rescorla, M. (2015, Oct 16). "The Computational Theory of Mind". Retrieved from https://plato.stanford.edu/entries/computational-mind/

Rosenblatt, F. The perceptron, a perceiving and recognizing automaton Project Para. Cornell Aeronautical Laboratory, 1957.

Searle, J. R. (1980). Minds, brains, and programs. Behavioral and brain sciences, 3(03), 417-424.

Simon, H. A. (1962). The architecture of complexity. Proceedings of the American philosophical society, 106(6), 467-482.

Simon, H. A. (1965). The Shape of Automation for Men and Management. New York: Harper & Row.

Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59(236), 433-460.

Cervier, D. (1993). AI: The Tumultuous Search for Artificial Intelligence.

Werbos, P. (1974). Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. PhD thesis, Harvard University, Cambridge, MA.