An Analysis of Artificial Intelligence in Machines & Chinese Room Problem

[1] Priyanka Yedluri, [2]A.Nagarjuna

[1, 2] Department of Computer Science, DVR College of Engineering & Technology
Hyderabad, Andhra Pradesh 502285, India.

Abstract - The success of machines over the last few decades in performing tasks that were seeming impossible for humans to perform led to the discussion that can machines be made intelligent. The argument was based on the fact that there was no understanding and the computer merely followed human etc. is not a new one. The question evokes deep programmed rules without any consciousness. The other side countered that an argument like that was arguable, since the results were as if produced by an intelligent being and had meaning, the computer has produced proof of intelligence. In this paper, we would analyze the arguments of both the sides and present a clearer picture of the capabilities of machine. We'll begin by explaining the Turing test, a criteria to test the intelligence of a machine and then move to discuss Chinese room problem and its implications. I will be highlighting the objections raised against these problems and my own answers to these arguments.

Keywords – *AI, Chinese Room Problem*

1. Introduction

Over the past few years, computer technology has developed with a supersonic pace. The future is bright and we hope to break the barriers that seemed to be untouchable not so long ago. The computers undoubtedly outsmart humans in so many domains but one. The capability to think. This is one area where the computers are still lagging far behind us.

The machines may have been able to beat us in brainy games like chess etc., but this success is attributed not to their intelligence, but to their capability to process monstrous amount of data very quickly. After all, chess strategies are nothing but a set of algorithms to anticipate the best possible move from the current move. The machines do this by considering every possible move and then selecting the best out of it based on some priority (called heuristic) function. But what if one day machines excelled at thinking too. Would it prove to be a blessing or the doom day (as predicted by terminator movies) when we would have to be the slaves of our creations. The debate on whether a mechanical device could be made to express feelings like love, hunger, jealousy philosophical issues like what does it means to think. What is mind and how is it different from brain. What does it really means to feel. Can I improve thinking ability if I implant a chip in my brain.

2. Intelligence and Turing Test

The question of whether or not machines can emulate human thinking is not new. However, it was first addressed from an Artificial Intelligence point of view by Alan Turing, acclaimed as the father of Theoretical Computer science. Turing was interested in the question of what it means for a task to be computable, which is one of the fundamental questions in the philosophy of computer science. He defined a computational model which was based on the sequential thinking of human brain. This computational model, popularly known as "Turing machine", was the foundation step for the computational models used in present day computers.

In his seminal paper entitled "Computing machinery and Intelligence", which appeared in 1950 in the philosophical journal Mind (Turing 1950), he proposed a test to compare the thinking capability of humans to computational capability of machines. This test, popularly known as "Turing test" tells whether a machine a machine can actually think. Let us propose a hypothetical experiment to explain the concept of Turing Test. Suppose the Stanford AI lab claims that it has developed a machine which can think. The machine has very large memory to store all the facts known to humans and can perform googol that their machine is a conscious being.

Compare a computer with human on the basis of looks, i.e. if computer looks like a human or not. Such matters would be irrelevant to the purpose of such a computer. We say that a computer thinks if it acts indistinguishably from the way a human acts when he/she thinks, minus the physical motion. Also note that we don't expect it to produce

answers to personal questions that we may shoot at it which only that person can know.

The Turing test goes as follows: There is an interrogator who will ask a series of questions to both human and computer and will differentiate between the two based upon the responses that they give. The interrogator cannot see who is answering. The responses are typed at a constant speed on a computer screen. No other information is available about the two sides apart from what could be inferred from the question and answers. The human subject answers honestly to justify him being a human. On the other hand, the computer is programmed to lie in order to convince the interrogator that its human. At the end of questions, if the interrogator is unable to decide which side is human, the computer is said to have passed the Turing test.

The unfairness that I described in the previously is not however, something that should worry us. In his paper, Turing suggested a 30 percent success rate for a computer with an "average" interrogator and five minutes of questioning by the year 2000. Keeping in mind the rapid strides that the computer industry is making, we should be able to achieve a hundred percent success rate by 2020. Thus, if such a computer is made, its passing of the Turing test can only be delayed but can't be avoided. The more important question, however, is that does the passing of Turing test is a real measure of a computer's ability to think. One may argue against the Turing test by several possible arguments1. It may so happen that after a long sequence of failures at test, the computer puts together all the human responses and next time when a similar question is asked, it appends some random element to it and answers the question. The interrogator may also run out of original questions soon. So, the computer then searches its vast database for repetition and if founds it, it could give the answers that a human would have given. Imitation, no matter how well executed, is not the real thing. Thus, learning to give the answers in a way indistinguishable from humans is not thinking.

However, the contrary viewpoint is, should the running out of original questions be considered cheating on the part of Computer. If such a intelligent machine exists which can imitate human responses, why not construct a similar machine which could ask questions like a human would have asked and have infinite questions in its database. We could give the task of asking questions to another machine and detect imitation. Thus, if the computer lacks intelligence, it is not difficult to detect it using second machine.

Both of the arguments seem to be reasonable. I myself believe that imitation, no matter how strong can always be detected by probing for long enough time and with suitable questions, if we are unable to do so, it's a shortcoming on human part and we shouldn't belittle the achievement of computer by questioning the authenticity of test. There is one more factor which comes into play here: the Interrogator's perspective. It may happens that he/she senses intelligence, but is unwilling to - acknowledge it. It appears that asking a machine to imitate human so closely is asking to much of it.

So, I can give the benefit of doubt in the favor of machine and I am ready to accept the result of Turing test as a measure of intelligence.

3. Strong Artificial Intelligence

One of the challenges before Artificial Intelligence is to make machines feel the emotions like hunger, pleasure, anger, pain, love etc. there is a point of view in AI, referred to as Strong AI which believes that devices could be as intelligent as humans.

Let us consider an example2 to illustrate this point of view better. We have a robot that detects light source in a room. Once it detects this source, it goes as close to it as possible, stays there for a certain time to recharge its photo-voltaic batteries, and then goes away to do other task like cleaning floor etc. We define the pleasure -pain score of our machine on a scale of 0 to 100. We says that that this score increases steeply when it gets light, increases slightly when it's in the presence of another robot (preferably with female attributes) in the room and decreases when its asked to perform some work like cleaning. Our robot tries to maintain its score above a certain level. When the level falls below a certain threshold, it abandons all tasks and starts searching for a light source.

Now can we say that our robot have emotions Does it feel hunger when its score decreases. Does it feels pain when its score falls below threshold. Does it feels love when in the company of female robot. And does it feels happiness when it's near a light bulb. In summary, how do we know that the score of the robot really indicates its pleasure or pain.

From an operational point of view, the answer would be to simply judge it from the way the robot acts. Thus, as it acts to increase its positive and score and tries to avoid negative score, we could define its pleasure as the degree of positivity of its score while we could define pain as the decrease in its score. But is this a justified explanation. If you ask a believer of strong AI, he would definitely say yes and dismiss the debate. But to me the situation seems to demand much more debate.

In our day to day life, we often say in a humorous way that "my computer hates me" or "my car likes to give me troubles". We certainly don't mean that computer hates us in a literary way. Thus, when we say that our robot "likes" the light source, we say so as it aids in our understanding, but not because it bears any literary meaning. What I mean to say is that from my point of view, there is lot more to view, but he has provided many examples where a machine could pass. Understanding the feeling by machines than only acting as if to feel, just because they have been programmed to do so. The strong AI has to present a stronger case to convince of their viewpoint.

The strong AI viewpoint is that not only the device that we talked about earlier has intelligence, but any sort of mental activity can be broken down to a set of set of logical operations and can be implemented on a hardware. Of course, the algorithm underlying the thinking process is highly complicated but its the same in principle. The believers of Strong AI say that whenever such an algorithm is found, it would not only pass the Turing test; but also, whenever it would be run, it would express feelings, have a consciousness and in short, have a mind.

4. The Chinese Room Problem

One of the main opponents to the Strong AI point of view is American philosopher John Searle. He has not only strongly disputed the simplified versions of Turing test but the quality of "understanding" was absent. To convince that the quality of "understanding" was absent, Searle proposed a thought experiment, popularly known as Chinese room Problem. The Chinese room first appeared in his paper "Minds, Brains, and Programs", published in Behavioral and Brain Sciences in 1980.

In the experiment setup, as mentioned on Wikipedia, Searle imagines himself in a room, acting as a computer by manually executing a program that convincingly simulates the behavior of a native Chinese speaker. People outside the room slide Chinese characters under the door and Searle, to whom "Chinese writing is just so many meaningless squiggles", is able to create sensible replies, in Chinese, by following the instructions of the program; that is, by moving papers around. The question arises whether Searle can be said to understand Chinese in the same way. Searle himself strongly argues that mere carrying out of an algorithm successfully doesn't imply an understanding of the Chinese. Searle would not have understood even a single word of Chinese by working like this.

Several objections could be raised over the Searle's stand2. A trite objection that comes to mind after the first encounter with the problem is that if an algorithm to simulate mind exists, it is going to be horribly complicated and no human being could compute it even in ten lifetimes. This question doesn't appeal much to me as we are not concerned with complexity of experiment here but with the implications of it.

Nevertheless, it could be encountered by replacing one man by a whole city of people (say Hyderabad), none of whom knows Chinese and dividing the complicated task into smaller parts and each doing it in parallel. This could be thought of like the assembly of neurons in brain which work together to produce intelligent decisions. However, another point could be raised on the following lines.

The other objection goes as follows – When a bunch of people are carrying out the algorithm, does this means that each person's brain understands the problem or do they collectively understand the problem . If we draw an analogy with neurons, of course each individual neuron doesn't understand what the person is thinking as a whole. In a similar manner, we could say that it is not required for each and every person to understand the whole thing. But if no one understands it fully, could it be understood completely as a whole. I would say in affirmative but I would also like to add that Searle's argument works best for a single person computation instead of a group. Of course we could justify the argument with bunch of people, it sounds more convincing when we talk about only one person Another objection that has been raised to Searle's Chinese room is that the symbols that the Searle processes are not meaningless codes, but they are meaningful Chinese alphabets. Hence, the processing that he is doing is also meaningful. We can counter this by saying that whatever meaning the computation derives is not an attribute of the process or the processor itself, but it is observer relative. It may have meaning to the Chinese speakers but in general, such a computation is meaningless.

Searle also tries to distinguish between the human brains and computer logic on biological grounds. He says that the difference lies in the machinery over which the algorithm would execute in the two cases. He insists that biological brain has "intentionality" and "semantics". Although he could not justify his belief scientifically, he says that these are the defining characteristics of intelligent thinking. However to me, this seems to be as dogmatic as the Strong AI perspective itself as there is no solid ground to it, just intuition only.

5. Continuing Debate

To call the Chinese room controversial would be an understatement. The number of objections published along with Searle's original (1980) presentation have exploded

since then, not only about whether the Chinese room argument is relevant; but, among those who think it is, as to why it is; and, among those who think it is not, as to why not. This discussion includes several noteworthy threads. Searle's Chinese Room experiment is closely linked to the Turing test and echoes Mathematician Rene Descartes' suggested means for distinguishing thinking souls from unthinking automata. Since "it is not conceivable," Descartes says, that a machine "should produce different arrangements of words so as to give an appropriately meaningful answer to whatever is said in its presence, as even the dullest of men can do" (1637, Part V), whatever has such ability evidently thinks.

The philosophy of Strong AI could be said to be broadly based on two hypotheses. First, known as the Behavioristic hypotheses denies that anything besides acting intelligent is required. The second one, known as the Functionalistic hypotheses holds that the intelligent-seeming behavior must be produced by the computations.

On the other hand, non-believers sympathize with two other hypotheses: Dualistic hypotheses holds that, besides (or instead of) intelligent-seeming behavior, thought requires having the right subjective conscious experiences. It asserts that there are two separate kinds of substance: 'mind- stuff and ordinary matter. The point is that the mind-stuff is not formed of matter, and is able to exist independently of it. Identity theoretic hypotheses holds it to be essential that the intelligent-seeming performances proceed from the right underlying mental (neurological) states. Searle, through his Chinese room, takes his shot at Behavioristic and Functionalist hypotheses. Searle in the experiment behaves as if he understands Chinese; although this is not true: so, contrary to Behaviorism, acting (as-if) intelligent is not necessary argument for being so; something else is required. But contrary to Functionalism this something else is not – or at least, not just – a matter of by what underlying procedures (or programming) the intelligent-seeming behavior is brought about:

Searle, according to the thought-experiment, may be implementing whatever program you please, yet still be lacking the mental state (e.g., understanding Chinese) that his behavior would seem to evidence.

Thus, Searle claims that Behaviorism and Functionalism are strongly discarded by this experiment; Arguments against strong AI somehow seem to be in unison with dualistic and identity theoretic hypotheses. Searle's own hypothesis of Biological Naturalism may be characterized as an attempt to merge – or at least as an attempt to find some certain degree of correlation between – the remaining dualistic and identity-theoretic alternatives.

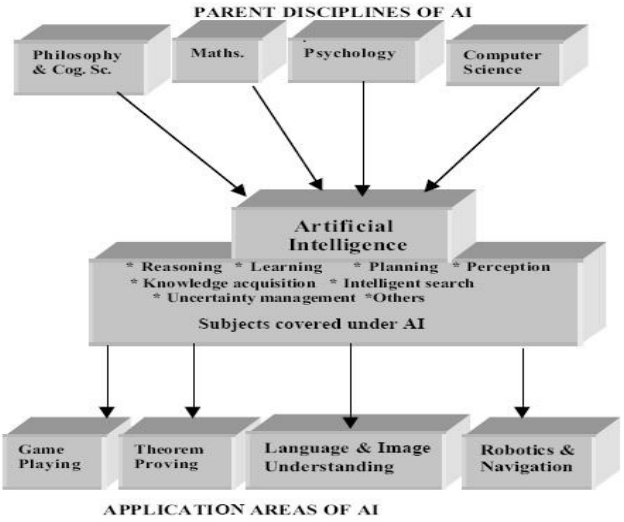

Figure 1 – Applications of AI

6. Conclusion

Chinese room fails to end the debate between the supporters and non-supporters of Strong AI. In fact, no hypotheses presented so far is close to being the deciding argument of this argument. Non supporters propose newer experiments every now and then to counter strong AI. On the other hand, with the rapid advancement in AI and Robotics, advanced machines are being develop every day. Every day, one could hear the news of yet another human attribute being successfully modeled by robots. Thus, machines are steadily progressing towards the ultimate goal Would a virtual mind, in a body of wires, circuits and chips would ever be able to experience the warmth of virtual love, in the same way as we do. Based on the current state of our knowledge and philosophical understanding of the topic, I find it hard to imagine that a machine would ever be able to have the same mental state as humans. Thus, even if it did understand love, the feeling won't be the same as what humans feel.

However, this doesn't rule out the possibility of a machine having brain. A machine may be made to understand, but this understanding won't be same as that of humans. Searle doesn't deny the belief that a machine would be able to pass the Turing test in near future. He is in fact was ready to accept the assertion that brain could be modeled by electronic circuits. Since every device having the capability to reason is some form of finite state machine, brains too, lies in the same category. But I am of the opinion that simulations, no matter how accurate or by how much powerful machine are not the same thing as original installation. I could clarify my position as follows: While a Linux machine might be able to emulate a Windows machine, and seem equivalent to those programs

run on it, to a person familiar with both, the differences will be apparent. The two machines will differ radically in speed as well as in other ways. It is clear that they are different on at least one level. The best way to run Windows is on a PC compatible machine. In the similar manner, the most efficient way to run the human 'program' has got to be on human hardware. The assertion is that this is a unique configuration, running human software on any other hardware is not going to have the same effects. Before concluding, I would like to address one last question that pops up in mind. If we are ready to let go the assumption that humans and machines are going to be equivalent, could we change the direction from pursuit of Strong AI to something which is more than weaker AI. I think yes, properly configured programs, capable of learning, sensing and acting in a meaningful manner, will be capable of understanding. The idea that they should understand in the equivalent way to humans is misconceived.

References

1. Turing, Alan M. "Computing Machinery and Intelligence." Philosophy of Mind: A Guide and Anthology. Ed. John Heil. New York: 2004.
2. Based on an example provided in “Emperor's new mind” by Roger Penrose.
3. Mariano de Dompablo Cordio “Searle's Chinese Room Argument and its Replies: A Constructive Re- Warming and the Future of Artificial Intelligence”
Indiana Undergraduate Journal of Cognitive Science 3 (2008).