Artificial Intelligence Lecture Notes by Dr. Belal Al-Khateeb

Week No.: 2

Lecture Title: The History of AI.

*** AI History**

Turing Test:

- **The test is conducted with two people and a machine.**
- **One person plays the role of an interrogator and is in a separate room from the machine and the other person.**
- **The interrogator only knows the person and machine as A and B. The interrogator does not know which is the person and which is the machine.**
- **Using a teletype, the interrogator, can ask A and B any question he/she wishes. The aim of the interrogator is to determine which is the person and which is the machine.**
- **The aim of the machine is to fool the interrogator into thinking that it is a person.**
- **If the machine succeeds then we can conclude that machines can think.**

- **Often "forget" the second person.**
- **Informally, the test is whether the "machine" behaves like it is intelligent.**
- **This is a test of behaviour.**
- **It is does not ask "does the machine really think?".**
- **It is too culturally specific?**
- **If B had never heard of "The X-Factor" then does it preclude intelligence?**
- **What if B only speaks Italian?**
- **It tests only behaviour not real intelligence?**

I. Background: Turing Tests, Behavior and Reduction

Alan Turing argued, in effect, that we should grant that a computer is intelligent if it can pass for human in online chat. To deny that such a talented computer is intelligent, a computer that can pass what we now call "*The Turing Test*", is just prejudice, *bio-chauvinism*. Turing thought that machines might pass for human by the year 2000.

On the face of it we might have worries about Turing's argument. Turing introduces The Turing Test by way of the "*Imitation Game*" in which unseen humans answering questions pretend to be of the opposite sex. But when say a woman wins the Imitation Game by fooling the judge into thinking she is a man, we do not conclude that she really is a man. Critics of the Turing Test argue that there should be no difference when a machine passes for human – we should conclude the judges were fooled, and the cleverly-programmed machine does not possess the traits – in this case intelligence and understanding – that are attributed to it by the judge. But perhaps intelligence is a very different kind of trait than sex, one defined by behavioral abilities such as problem solving, and not by biology and physical traits. Indeed, that seems to be Turing's view.

In the Turing Test, a successful computer would exhibit apparent verbal intelligence, responding appropriately to questions and so apparently understanding language. If the computer does not actually understand the language of its responses then it merely fools the judges. But how could one show that the machine does not actually understand the questions it is asked and the responses it makes? In 1980 U.C. Berkeley philosopher *John Searle* argued he could *prove* computers do not actually understand the questions they may answer so well, and hence Turing and other *AI* enthusiasts are wrong. Searle's argument is now known as *The Chinese Room Argument*, or CRA. It has become one of the best known arguments and thought-experiments in contemporary philosophy.

One way of looking at AI is that it is the claim that intelligence reduces to stupidity. The claim is that intelligent behavior can be produced by very simple operations. This is just an extension to the biological realm of Turing's *1937 thesis* that anything computable is computable by a Turing Machine, a very simple machine that can only perform a handful of very simple operations. The magic is in the sequence of steps, not the steps themselves. The view Searle calls "**Strong AI**" holds that intelligence is computable, and hence reducible to very simple operations. Many cognitive scientists (e.g. leading psychologist *Steven Pinker*) hold that human cognition is computation by neurons. Thus the Chinese Room Argument occurs within a long history of reductionist debates. Searle is no *dualist*; he thinks mental states are *products of our brains*, not spirits or souls. But he believes he can show computer-like manipulations of strings of symbols can't be the operations that give rise to understanding and consciousness.

II. The Chinese Room Argument

By 1980 AI researchers had claimed that by running their programs a computer could come to understand a sub-set of English. For example, the computer might be given a story about a visit to a restaurant, and then correctly answer questions about what happened there, where the answers to the questions require inference and are not explicit in the story. Searle argued that the best way to test a claim like this would be to do what the computer does – run a program – and see if running a program enables a human to understand a language the person did not know before running the program. The idea of humans running computer programs was not new: Turing himself had mentioned in his writings that he ran through the steps of a program to play chess – but Turing was not interested in what this was like, nor the philosophical question of whether he or something else (the program?) was playing the resulting chess game.

The core of the CRA is a very simple thought-experiment. Thus Searle imagines himself inside a sealed room. In the room are a large number of manuals in English – an English version of the program and its databases. People outside the room slip pieces of paper under the door with Chinese characters on them. Searle runs the program and follows its instructions which eventually result in his putting Chinese characters on another piece of paper and pushing it back out under the door. Unbeknownst to Searle, the incoming slips of paper are questions, and the slips he pushes out contain appropriate answers.

To those outside the room, it appears that someone in the room understands Chinese. But no one does. For example, a question might be about hamburgers, but Searle would never know this. Since Searle just does what a computer does, and he merely appears to understand the language of the input, then running a program cannot make a computer understand language either, no matter how apt its responses. Searle argues that the larger lesson from the thought experiment is that you can't get **semantics** (meaning) from **syntax** (sequences of marks) alone. Computers are very good at mechanically manipulating strings of digits – but they have no understanding of what any of it means.

Searle does not specify the nature or level of the program he follows in the Chinese Room. That matters. Suppose the program says "if you see this Chinese character, then go to the entry "hamburger" in the internal copy of Wikipedia". Then Searle might soon get at least a vague idea of what was going on. So I think it is best to think of the programs as the English equivalent of machine language, specifying manipulation of strings of 1s and 0s, with databases that are arrays of 1s and 0s as well. That is what the hardware of a computer "sees", and so if the events in the room are to be most like what a computer does, the level of processing should be a low one.

III. Critics

There is no consensus about the merits of the CRA. Some believe it does prove that digital computers cannot be programmed to understand language, and hence the Turing Test is excessively behavioral. Others believe that the CRA does not prove anything. Leading philosopher *Dan Dennett* calls it an "*Intuition Pump*", guiding us to wrong conclusions about complex systems. I think that the many criticisms of the CRA are best understood in terms of what they concede. Some critics hold that Searle is right about the original thought-experiment, but that some variation on the Chinese Room would result in understanding and intelligence. Others do not concede that the original CRA shows Chinese is not being understood. Let us look at the major criticisms.

Sympathetic critics concede that Searle is right about the inadequacies of the Turing Test, but that he is mistaken in thinking that he has shown computers are intrinsically incapable of producing understanding. What if we add sensors – cameras, microphones, thermistors, maybe even scent detectors – and let the computer roam the world in a robot body? Might it not then be able to learn the meaning of words just as humans do, by associating them with the world? A Total Turing Test would grant understanding to a robot that could interact with its environment just as intelligent humans do.

Searle argues that this "**Robot Reply**" to the CRA fails. For we could simply modify the Chinese Room scenario so that the operator in the room receives additional digital output from the robot sensors – just a very large real-time stream of 1's and 0's. Of course there must also be a much larger and more complex program for processing all this binary data. Let the busy operator output a stream of 1's and 0's that appropriately controls a robot body, a body that navigates through the world and speaks in Chinese. But through all this the room operator understands no Chinese and also understands nothing about the non-linguistic robot interaction

with the world. Thus Searle says such a Total Turing Test is also inadequate; the Robot Reply fails.

Other objectors asked, "But what if the program is a neuron-by-neuron model of the brain of a Chinese speaker?" In a brain, neurons act on neurons to cause them to fire, and this eventually results in behavior. A program could model the interactions of all these neurons, and produce whatever thought and behavior the modeled brain would produce, including understanding and speaking Chinese. Searle calls this "**the Brain Simulator Reply**"; *Paul* and *Patricia Churchland* are leading proponents of understanding minds at the level of neuron function. Searle argues this still fails to get around the core of the argument – you could run this simulation and still not know what any Chinese word meant. Furthermore, a simulation by a computer, whether of digestion or storm system or a brain, is *just* a *simulation* and not the real thing. No one gets rained on by a computer simulation of a weather system, and no one's questions are understood by a computer simulating understanding.

The Robot Reply and Brain Simulator Reply argue that Searle may be right about the original room story, but wrong about modified scenarios of robot or brain models. But the **System Reply**, which Searle says is the most common reply to the CRA, holds that there is understanding going on in the unaltered CRA scenario. The key claim is that the thing that understands Chinese is not the room operator, it is the entire system. It is not Searle running around the room, but Searle plus the massive program and all the information that it embodies that is doing the understanding. Searle counters this System Reply with his own modification of the room scenario: in principle he could memorize the entire program and thereby *become* the system. But this would change nothing; he might be faster but would still just be manipulating strings of 1s and 0s in his head and still wouldn't understand Chinese – there would be no real change at all (except maybe a headache).

A final "**Other Minds Reply**" asks: isn't it only by behavior that we know that other people understand what we are saying"? Searle agrees – but we know that they are people, and hence like ourselves. The situation is very different with a computer, where, as Searle says he has just proven, we know they just manipulate strings of 1's and 0's and don't understand. Searle's counter-argument here raises many questions – we don't really know how much variability there is in brain structure and human cognition. It appears biologically possible that some people's brains work by manipulating Mentalese syntax – just like an Imitation Game judge, evolution can select only on the basis of behavior. Hence humans might be a mix of those with brains that are computational, and those that achieve the same result some other way. Thus, it is not clear that it is so easy to dismiss the Other Minds criticism.

IV. Minds and Machines

My own view is to concur with Searle that the CRA proves that no computer comes to understand Chinese just by running a program. I think the logic of the argument is impeccable. At the same time, I don't think the argument shows anything important about the limits of AI or refutes computational views of human cognition. That may seem a paradox, but it is as simple as thinking of the CRA as a very natural strawman argument. The initial claim – that computers understand language – is refuted, but it is not the claim we should be interested in. The interesting claim is that computers running a program gives rise to a mind which understands – a mind which is not identical with the computer. On this view the computer is just a box that is not identical with the entity that understands. But minds and bodies are not identical. The physical box can exist when the understander does not, e.g. if its hard drive is erased. And vice versa, if the program runs on different hardware – so if computational processes give rise to a mind, the computer and the mind that understands are not

identical things. Minds in general are not identical with the systems that realize them (hence the System Reply isn't quite right). I think this delicate but important metaphysical point shows that Turing and strong AI has not been refuted by the CRA.

I have tried to illustrate this point in various ways; I'll try a new one here. Re-visit the Robot Reply. Suppose that Searle in the room – unbeknownst to him, of course – is controlling two remote robots over the internet. **Johnny 4** is a robot in Beijing, and converses only in Chinese. Johnny 4 says in Chinese that it is 3 years old, and that it has never been outside of China, but someday hopes to learn English. When a Big Mac is held up in front of Johnny 4's camera eyes, and it is asked what is being offered, it responds (in Chinese, as always): "That looks like a hamburger." Meanwhile a second robot up in Oregon called **Johnny 5** (the robots are named in honor of Searle's first name) converses only in English. It remarks "I love Astoria. Of course, even though I am former military, I have never been anywhere else." Now one robot is an English monoglot, understanding no Chinese, and the other is a Chinese monoglot, understanding no English. They differ in what they say on a host of topics, and they display different personalities. Since they have different properties, they are not identical to each other. Meanwhile Searle, back in the room in Berkeley, is manipulating strings of 1s and 0s, making it all happen. Although the claims made by the robots are true, we might suppose, they are not true of Searle and yet Searle is not lying. The robot responses to questions are not Searle's responses, although he is a part of the causal process that makes them happen. It is entirely irrelevant what languages Searle understands or doesn't understand to the understanding, if any, of the robots.

Searle is not the author of what the robots say. So who is talking when the robots talk? How about: Johnny 4 and Johnny 5. Each robot has its own personality, apparent beliefs, and linguistic capacities. Johnny 4's answers are not Johnny 5's, and their answers are not Searle's. Nor does Searle see

what each robot sees with its camera eyes. The robots move and pick up and manipulate objects; Searle is pushing paper in Berkeley. Hence the agents that speak and see and roam around Beijing and Astoria respectively are distinct agents from Searle. The thought-experiment I have imagined, with two robots with non-identical cognitive processes and abilities, underscores the irrelevance of the room-operator's abilities and mental states to the states of the robots. Yet the cognition of the robots is causally dependent on the low-level activity of another mind, another conscious agent, Searle, with his own distinct thoughts ("Man, is this boring!") and abilities. If Searle is doing shift-work in the Chinese Room and is replaced at noon by someone else, e.g. someone who does understand Chinese, it makes no difference whatever. Running the program creates a metaphysical wall between the agent in the room and the agent – or agents – that the room activity creates. The robots, it seems to me, are best described as virtual minds that see and speak and make decisions, created by the activity of a distinct mind. Nothing in this story is a *proof* that the robots understand, but I think it does show that the CRA fails to prove that they do not.

What about syntax and semantics? That is beyond the scope of my discussion here, except to say that it is not clear that if running a program created understanding it would be doing the impossible, producing semantics from syntax, meaning from mere marks. A computer is an electronic device – it runs on voltages and currents. *We interpret* some of these electronic states as syntactic, as binary numerals – the computer does not. So the real question is whether you can get understanding and meaning from a massive electronic causal system. That is a hard question, perhaps as hard as how nature produces understanding and meaning from neuron firings. It is unlikely these hard problems can be settled by simple arguments. Nevertheless, at the least, the CRA and its replies are paradigms of the joys and perils of thought experiments.