CSCI-UA 9472. Assignment 1 – **Solutions**

Augustin Cosse

December 23, 2021

1 Question 1 (5pts)

Read Turing's original paper on AI (Computing machinery and Intelligence, 1950). In the paper, Turing discusses several potential objections to his proposed enterprise and his test for intelligence. Answer the following questions:

1. Describe the imitation game (rephrase in your own words) The original imitation game is played by three people: a man (A), a woman (B) and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by label and at the end of the game, he should be able to say either 'X is A and Y is B' or 'X is B and Y is A'. The interrogator is allowed to put questions to A and B. The main question is then what will happen when a machine takes the part of A in the game? Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?

2. What are some of the advantages of the imitation game that are listed by Turing ? The problem draws a sharp line between the physical and intellectual capacities of a man. There is indeed little point in trying to make a 'thinking machine' more human by dressing it up in a material that is indistinguishable from the human skin. The form of the problem reflects this idea. Moreover the question and answer method seems suitable for introducing almost any of the fields of human endeavour. We don't want to penalize a man for its inability to shine in beauty competitions or for losing in a race against an aeroplane. The conditions of the game makes these inabilities irrelevant.

3. What are the restrictions on the machines that are allowed to play the game? (be as complete in your characterization as possible) In particular, what is the interesting variant that Turing mentions?

   At first, the idea of Turing is to permit every kind of engineering to be used in the test (including if the functioning cannot be described because it is experimental). Turing however excludes men born in the usual manner. Finally, to prevent the use of biological techniques (such as cloning) Turing narrows his description of the machine to a digital computer.

4. What are the three parts of the digital computer according to Turing (describe each part in detail)?

   According to Turing a digital computer can be described as consisting of three main components:

- The storage unit (which stores the information and corresponds to the human computer's paper or his memory, since he does part of the calculations in his head)
- The execution unit (which is the part that carries out the various individual operations involved in the calculation)
- The control unit (whose objective is to make sure that the instructions contained in the storage unit are obeyed correctly and in the right order.)

5. What was the problem with Babbage analytical engine ?

   The machine was never completed. Although Babbage had the ideas, the machine at the time, was not a very attractive prospect. Part of the reason for this being that the speed which would have been available at the time (although faster than a human computer, would have been 10 times slower than the Manchester machine (itself one of the slower of the modern machines)). Moreover the storage was entirely mechanical, using wheel and cards.

6. What is, according to Turing, a difference between the discrete state machine and Laplace's description of the universe?

   Unlike in the Universe, where a small error in the initial conditions can have an overwhelming effect at a later time, Turing considers as a property of the discrete state machines that such a phenomenon does not occur. I.e reasonably accurate knowledge of the state at one moment yields reasonably accurate knowledge any number of steps later.

7. What does Turing mean by a *universal machine*?

   By universal machines, Turing means the property that digital computers can mimic any discrete state machine

8. Turing suggests replacing the original question "Can machines think" by a more precise formulation ? What is this refined formulation?

   Are there imaginable digital computers which would do well in the imitation game?

9. List and provide a short (no more than a couple of lines) summary of the objections that Turing addresses and what are his refutations (2/3 lines summary). In particular, what does Gödel's theorem states ? and what is the answer of Turing to the objection saying that machines cannot make mistakes?

   Among the objections that Turing considers, we can mention

   - The Theological Objection. This first argument considers thinking to be related to the soul. To counter that argument, Turing indicates that there is no reason that a soul should not be conferred to an animal and hence, by extension to a machine. Turing also cites the example of Galileo against whom where used texts such as "And the sun stood still.. and hasted not to go down about a whole day", ... As additional argument against this objection, Turing highlight the lack of a common definition for the concept of soul across religions. In summary Turing takes the view that if there is any soul, it is conferred by God for reasons that He is the only one to know, so that no one could pretend whether or not a machine could be conferred such a soul at any point in the future.
   - The Heads in the Sand Objection. This second argument is a safety argument which considers that designing thinking machines would be dreadful.
   - The Mathematical Objection. This objection relies on mathematical logic to show that there are limitations to the power of discrete state machines. As an example, Turing cites Gödel's theorem which shows that in any sufficiently powerful logical system, one can formulate statements that can neither be proved nor disproved within the system, unless the system itself is inconsistent.

- The Argument from Consciousness (which can be found in Professor Jefferson' Lister Oration) relies on the idea that sentiments are out of reach to digital computers. Turing discards it by saying that the only way to satisfy this objection and quantify the sentiments of the machine would be to be the machine itself.

10. How does Turing connect artificial intelligence to sub-critical and super-critical piles?

    Turing takes the example of subcritical and supercritical piles to counter Lady Lovelace's objection which states that a machine can only do what we tell it to do. In an atomic pile of less than critical size, an neutron entering the pile will cause a disturbance that will eventually fade away. If the size of the pile is sufficiently increased however, the disturbance caused by the incomming neutron will go on and on increasing until the whole pile is destroyed. Turing uses this example to show that just as while most of the ideas that occur in the human mind are subcritical and only some of them can be labeled as super-critical, there does not seem to be any reason why a machine could not at some point be made supercritical.

11. What is the opinion of Turing regarding rewards and punishments in the design of intelligent programs?

    An important feature of a learning machine is that its teacher will often be very largely ignorant of quite what is going on inside. This is in clear contrast with normal procedure when using a machine to do computations: one's object is then to have a clear mental picture of the state of the machine at each moment in the computation

12. Can you think of new objections (aside from those mentioned by Turing) arising from developments since he wrote the paper? This last question is clearly an open question. However, we could mention the ecological cost or computational cost of current model and the limited amount of natural resources on earth. Another one is the amount of time needed to train current models. We can also mention the multifaceted nature of intelligence (logical reasoning vs learning vs exploring) and the lack of a proper formalism describing all the facets.

2 Question 2 (5pts)

Read Searle's refutation of strong AI (The Behavioral and Brain Sciences, 1980). In the paper, Searle introduces his famous Chinese room Gedankenexperiment. Answer the following questions:

1. How does Searle define strong and weak AI?

   Searle calls strong AI, the school of thought according to which the computer is not merely a tool in the study of the mind but is instead a mind in itself in the sense that given the right program, it can literally be said to understand and have other cognitive states. Searle calls weak AI, the school of though according to which the "only" value of the computer in the study of the mind is thjat it gives us a very powerful tool.

2. What does Searle mean by intentionality?

   By intentionality Searle means the ability to understand the processes or manipulations through which the machine goes.

3. What is the main thesis of Searle? The main idea of Searle is that only very special kinds of machines, namely the brains and machines with internal causal powers equivalent to the brains can think

4. How would you describe Searle gedankenexperiment ? (rephrase in your own words) In particular, how does Searle use his gedankenexperiment to refute the strong AI view on Shank's program?

   Searle considers the example of someone locked in a room and who would be given a large batch of Chinese writing. Furthermore, the person is assumed to know no Chinese and is unable to recognise Chinese writing. Together with the first batch of writing, the person locked in the room is given a second batch of Chinese script together with a set of rules for connecting the second batch to the first batch. The rules are, as for themselves, written in English so that the person can understand identify the symbols by their shapes. The person is finally given a third batch of Chinese symbols together with an additional set of instructions, again written in English, which makes it possible to correlate the elements from the third batch with those from the first and second batches.

   The first batch could be compared to a script, the second to a story and third corresponds to questions.

   In this framework the symbols that the person locked in the room outputs (without understanding them) are considered as the answers.

   Searle further considers that the candidate locked in the room receives the stories in plain English.

   If we assume that the employee in the room becomes better and better at manipulating the (english) instructions on the (chinese) symbols, nobody just looking at the answers of this employee could tell that he does not speak a word of chinese, while the employee is in fact manipulating uninterpreted formal symbols.

5. What 1963 paper does Searle cite as an example of an erroneous attribution of intentionality to a program?

   The Newell and Simon paper which, according to Searle, considers that computer cognition is exactly the same as human being cognition.

6. What is the Berkeley reply to Searle's argument? and what is Searle refutation of this reply? What is your opinion on this first reply?

   The Berkeley replies consists in considering the system as a whole (including not only the individual in the room but also the three batch). In that sense one can consider that the system does understand the task.

   As his refutation Searle suggests to incorporate the whole system within the individual (assuming the individual can memorize the batches of symbols along with the set of rules written in English). According to Searle, the result is the same

7. At some point in the paper, Searle mentions that the two systems (Chinese and English) can be considered to pass the Turing test while the first system exhibits a clear understanding of the language and the second exhibits no understanding at all. What, in your opinion is a danger with such a statement? (hint: you can check the work of Hector J Levesque and the paper "Is it enough to get the behavior right?")

   One difficulty with Searle considering that his experiment is a refutation of Turing's imitation game lies in the memorization of the books that enable the employee to understand chinese. As indicated by Levesque, memorizing a full book would be equivalent to knowing perfect chinese.

8. To provide an additional refutation of strong AI, Searle mentions other organs as information processing subsystems. What point is he trying to make?

   Searle's mention of biological subsystems arises in his refutation of the Berkeley system's reply. According to Searle we cannot assume, simply on the basis that a subsystem processes inputs by means of a program to produce given outputs, that there is cognition in the program. If this idea was true, at a certain level of description, a stomach could be considered to do information processing.

9. What is, according to Searle, a distinction that strong AI should be able to make? (Searle in particular provides a citation from McCarthy. What is the citation?)

   McCarthy is quoted by Searle to have said that ”Machines as simple as thermostats can be said to have beliefs, and having belief seems to be a chracteristic of most machines capable of problem solving performances”. The quotation is given by Searle to insist on the fact that according to him, for strong AI to become a branch of psychology, it should be able to make a proper distinction between subsystems that are genuinely mental from those that are not.

10. What is the Yale reply? and the corresponding refutation from Searle?

    The Yale reply considers the design of a program that would focus on replicating a human being instead of succeeding at the imitation game. I.e, the reply considers a program that would be combined with a robot, as well as a camera attached to its head so that it can replicate human vision, arms and legs so that it could act and walk. The idea of the Yale reply is that in this framework, the robot could be considered to exhibit some form of understanding.

    Searle's reply consists in saying that the addition of perception and/or motor capacities adds nothing regarding the presence of any understanding. Searle even suggests replacing the central program (i.e. located in the robot's head) to be replaced by a Chinese room, hence showing that the given exhibits no more understanding than the employee locked in the room.

11. What is the Brain simulator reply? and what is Searle's corresponding refutation (in particular what is the water pipe example)?

    The brain simulator reply is one of the most interesting replies. Although not rely a refutation of Searle's chinese room experiment (it only provides an example of a program that could be said to have consciousness instead of showing that every program passing the Turing test can be considered to have consciousness), it is one of the few replies that can be considered to really highlight the transition between human cognition and artificial intelligence.

    The reply focuses on the simulation of the neural processes occuring in the brain. If we can write a program that simulates the sequence of neurons firing at the synapses in the brain of a native Chinese speaker, in this case, we would have to say that

12. According to Searle, what is the problem with the combined reply? (as part of his refutation, in particular, Searle mentions animal consciousness. What is he trying to demonstrate?)

    According to Searle, the only proof that the combined reply gives is that if a sustem gets sufficiently close in its design to a human being, then in all likelihood we can assume that there is a point at which it will be correct to assume that it exhibits consciousness. However Searle insists on the fact that this is not what strong AI posits. Searle makes the distinction between mimicking every aspect of human consciousness and assuming that a formal program can achieve consciousness.

    Searle compares the case of a robot, whose main and only component would be a formal program that could be replaced by the employee locked, then insisting on the fact that in this case, the employee, although receiving all the inputs to the robot and producing all the outputs, has no idea what the robot is doing. Searle compares this framework to the intentionality that is assumed to be present in apes or other animals, indicating that the reason one can consider the latter to exhibit intentionality is because they are ”made up of similar stuff to ourselves”

13. The other mind reply from Yale provides an answer to the question ”How does one know that other people understand a particular language?” What does Searle consider to be essential processes coming on top of the computational processes (over formally defined elements) that make the difference between intentionality and the lack of it?

    Cognitive states

14. Towards the end, Searle compares the brain and the mind in term of information processing. What is his conclusion?

    For Searle, the distinction between the program and the realization is the same as the distinction between the level of mental operations and the level of brain operations. the analogy seems to suggest that if we could describe the level of mental operations as a program, then we could describe what is essential about the mind without doing introspective psychology of the brain. There are however, according to Searle, 3 points at which the equation "mind is to brain as program is to hardware".

    - We need to keep in mind the distinction between a program and its realization. A same program can have all sorts of crazy realization that do not exhibit any intentionality.
    - While the program is formal, the intentional states are not.
    - Mental states and events are literally a product of the operations in the brain, but the program is not a product of the computer.

15. What is behaviorism? What is functionalism?

    In this case, what is meant by *Behaviorism* is that AI practitioners tend to think that if you can build a machine that behaves intelligently, then it really is intelligent.

    Functionalism holds that a mental state is what a mental state does – the causal (or "functional") role that the state plays determines what state it is. In Functionalism, what matters is again what the mind does, hence the program.

16. Quite surprisingly Searle connects strong AI and dualism. How does he justify this connection? and why is it at odds with the traditional view on strong AI?

    According to Searle, Strong AI implies that when considering the mind, since the mind is viewed as a pure program, the support, hence the brain does not matter. This is at odd with the traditional definition of strong AI as strong AI considers that the mind is a direct extension of the processes that are generated at the level of the brain.

3 Question 3 (5pts)

Read Christine Kenneally's paper Mind Machines from The New Yorker and provide a complete answer to the following questions:

1. What is the name and university of the philosopher who gets interested in the case of Legget? Frederic Gilbert was a philosopher at the University of Tasmania
2. What are the research interests of this philosopher? (Try to be as exhaustive as possible)

   Gilbert did his PhD on free will. While hanging out with scientists working in genetics, he got interested in studying determinism in a scientific way (in particular following a biological approach)
3. To what extent was the device NeuroVista able to help Legget with her medical condition?

   The device from NeuroVista sent an alert each time she was about to have a seizure.
4. What were the main steps of the operation needed to implant the device and what problem initially took place during the first few days after the implant? Leggett had a small hole drilled in her skull inside which the surgeon slid a cross shaped silicone strip which was laid across the surface of her brain. Initialy the recordings recovered by the NeuroVista device because the brain was simply reacting

5. How does Legget describe her relation to the device? Thanks to the device, Leggett felt like an entirely new person. Leggett speaks of the device as if it were a partner.

6. What are some of the positive and negative side effects of medical implants?

   implants enable the restoration or enhancement of human abilities. On the other hand, it's becoming apparent that many people develop an intense relationship with their device, often with profound effects on their sense of identity. Those effects are still little studied.

7. What was the problem with Hannah Galvin?

   Unlike Legett, Galvin ended up hating her device. Her antipathy to her device was almost instant. It felt as if there were someone inside her head, but it wasn't her. She hated the telemetry unit embedded in her chest

8. What conclusions would you draw from the opposite experiences of Legget and Galvin?

   patients' perspectives are vital. Ethical issues are in constant danger of being overshadowed because of how rapidly technologies are developing

9. What is one of the issues with scientific papers from the medical devices industry and what is the solutions proposed by Gilbert?

   Most published papers don't mention ethics or risk, and, he said, because companies have no obligation to publish the outcomes of failed trials, the results over all appear to be ninety-nine per cent positive. Gilbert has been working on protocols to prevent harm: Neurosurgeons must declare financial interests. The risks described on consent forms need to be better articulated. Participants in early trials must understand that irreversible consequences of the trial might prevent them from receiving the better therapy they are helping to develop. All trials should express interest in the autonomy of a patient after implantation and after explantation. International research projects must also contend with national differences in ethical standards

10. What problem can happen when considering the removal of an implant? Patients can undergo some form of identity change

11. At the end of the day, and despite the removal of the implant, what did Legget get out of her experience with this implant? Although the implant was repmoved, Legget kept a better sense of when a seizure might occur