CS 486 - Assignment 1

Alex Klen
20372654

January 15, 2015

1 Computing Machinery and Intelligence

i) **What is the imitation game?**

The imitation game is a test to see if an interrogator can tell the difference between a human being and an AI agent through a textual conversation. It is setup with the interrogator and two subjects all in separate rooms with a text-based communication method between the interrogator and each subject. Subject A is human, and subject B is an AI program or machine. The interrogator communicates with each subject one at a time for an allotted period of time, and then has to decide which subject is A and which is B. If the interrogator cannot choose correctly more often than not then the AI system is of human intelligence.

ii) **Why does Turing think that the imitation game is worth studying as a test of intelligence?**

Turing believes the imitation game is a fair test for a machine to demonstrate intelligence without having to display other qualities of humans, such as appearance. The game, however, can test knowledge, understanding, and learning of an AI system in any area, as long as questions can be expressed through text. Since humans mainly use language to convey thought, the game is an adequate test of intelligence.

iii) **Summarizing Objections**

The mathematical objection is that Gödel's Incompleteness Theorem proves that sufficiently complex formal logic systems, and therefore digital computers, cannot both be consistent and complete. This means there are valid theorems for a particular system which it cannot prove to be either true or false. The claim is that the human mind doesn't suffer from this deficiency, and therefore a digital computer cannot imitate a human. Turing states that there is no supporting evidence that humans don't suffer from this flaw. He also states that while a particular question can be designed so that a particular machine can't answer it, other machines could answer it, and so there is no way to triumph over all machines in this manner.

There are numerous objections from various disabilities, notably that machines cannot make mistakes like humans can since machines follow instructions exactly. Turing's argument is that there are two kinds of errors - errors of functioning and errors of conclusion. Errors of functioning mean the machine had a fault in its code or a hardware problem and is unintentional, and ideal, abstract machines never suffer from this error. However, a machine's output can contain errors of conclusion - logical errors in its output. When given an arithmetic expression, an intelligent computer wouldn't execute it as machine instructions, but it might pause for some time and randomly introduce errors to mimic a typical human's performance.

iv) **Objections still relevant today**

The argument from consciousness is still an important question today. We have come a long way for mimicking specific abilities that humans have, such as speech and visual object recognition. It's still an important question whether adding more complexity to AI systems, for example deeper artificial neural networks, can give rise to general intelligence - machines that can learn anything like the human mind can. It's a big question whether consciousness and self-awareness can be created with a good enough learning machine - with a level or density of intelligent machinery beyond a critical level. It is a difficult question to consider because we would be hard-pressed to say a machine has conciousness since it's so different from ourselves. But as Turing states, how can we say for sure that a particular human has understanding and conciousness as opposed to just following rules if we are not that human? In a way we are just learning and following different rules. If a machine succeeds in tricking the interrogator in the imitation game every time - it can reason, learn, and have desires just like as any human - then we could only admit that it understands and possesses conciousness.

Lady Lovelace's Objection is that a machine can't originate anything. This is still an important question because it asks whether a machine where all of its inputs are deterministic and can be observed and recorded can come up with something original or creative. If it cannot, then it shouldn't be able to succeed at the imitation game since we can pose a creative challenge. However this brings up the question of whether humans can come up with anything original. Humans process so much external stimulus constantly that it's not hard to imagine that apparent randomness from the environment combined with learned patterns and concepts is responsible for all human creativity. If then a machine is supplied with a random number generator it may be able to also mimic human creativity.

v) **Find an objection not considered by Turing**
An objection discussed in “Does the Turing Test Demonstrate Intelligence or Not?” by Stuart Shieber (http://www.eecs.harvard.edu/shieber/Biblio/Papers/turing-aaai-senior.pdf) is that a machine which simply memorizes suitable answers to every possible sequence of inputs up to some limit can beat the imitation game if it is run with up to that amount of input. So a machine could theoretically be built with huge storage banks that stores output for each possible sequence of inputs for a 10 minute conversation. This machine would pass the Turing test and then be deemed intelligent, but it won't be able to respond intelligently to any conversation longer than 10 minutes long, and can't be called intelligent because all of its outputs are canned. Shieber's response is that a Turing test of even 1 minute rules out this type of machine because it would require more matter than the universe contains in order to store outputs for all possible sequences of inputs this long.

2 Minds, Brains and Programs

i) **Differences between “strong” and “weak” AI**
Weak AI is a system that inhibits some qualities of the mind and is useful for studying how the mind works. This is what the current AI systems we have today fall under - they can do things like object recognition from images, or looking through a database of information to answer questions, like IBM's Watson, but they do not possess any understanding of what they are doing - they are just following a human-created program.
Strong AI is not just a model or tool, but in fact is a mind that exhibits intelligence and understanding. This would be a system that really does understand what it's doing, like a human, instead of just following preprogrammed rules.

ii) **What is the Chinese Room?**
The Chinese room is a scenario in which a man who does not understand any Chinese is placed in a room with a written set of rules (in English) of how to manipulate Chinese symbols from an input to create a response. The man simply follows the rules exactly and doesn't understand anything about the input and output. However, once the man gets fast enough at operating these rules, a Chinese speaker sending in input and getting back output cannot distinguish the man in the Chinese room from another Chinese speaker.

iii) **Why does Searle think that the Chinese Room invalidates strong AI?**
Searle says that a machine is exactly like the man in the Chinese room. He takes input, follows precise rules, and forms output. Searle believes that even if a machine was constructed that could pass the Turing test, it couldn't be classified as strong AI because at its core it is just manipulating symbols based off of rules, and like the man in the Chinese Room, it doesn't really understand what it's doing.

iv) **Replies and Searle's critiques**
The systems reply from Berkeley states that even though the man in the Chinese room does not understand what he is doing, the combined system of the man, the rules and any data storage material he is using to carry out the rules understands the input and response. Searle's critique is that the man could memorize all of the rules and perform them all in his head. The system, then, would be just the man, and yet the man is just following rules without understand a single character of the Chinese he is looking at.

The brain simulator reply from Berkeley and M.I.T. supposes that we could encode rules that simulate the entire brain of a Chinese speaker. The rules would simply be changing electrical signals for various neurons depending on the inputs and the brain's state. Then surely this machine can be said to understand Chinese. Searle's reply, similar to the above, is that the underlying system has no understanding since it is only simulating neuron's firing and has no idea about what the Chinese it is taking as input actually means.

v) **What does Searle think is the main difference between human cognition and Strong AI?**
Searle states that strong AI, which is a general AI system that can pass the Turing test, is different from human cognition because the machine executing the AI program does not understand anything. It is simply shunting symbols according to a series of machine instructions. Human cognition, on the other hand, is fundamentally different because people understand what is being spoken to them, their own mental states, and their reasons for their behavior. Humans have understanding, whereas AI - even AI that can behave indistinguishable from humans, does not.

vi) **Do you agree with Searle?**
I agree with the very fine technical distinction he makes between the underlying machinery running a strong AI program not understanding anything but simply following rules, and the instantiation of the program having understanding. However, I disagree with any conclusions he makes about how this invalidates strong AI. The same argument can be made for human beings, saying their collection of neurons doesn't have any understanding of the person's concious mental state or the inputs and reactions he or she makes. A human's understanding and conciousness presumably comes from the patterns of activations of neurons - the parallel to an instantiation of a computer program. For the Chinese room example where the man has memorized all of the rules, I claim that the man doesn't need to understand any Chinese for there to be understanding because in this case the man's mind is actually the machinery manipulating symbols for another "mind", which he must be simulating within his own (assuming he has enormous mental capacity relative to the mind he is simulating) .

3 AI in the Media

i) **What is the piece's main concern?**
The piece states that the AI community needs to exercise a lot more caution then they currently do. It likens the arrival of human-level intelligent AI like an alien race that has warned us they will invade and we're patiently waiting for them instead of preparing to defend ourselves. It speculates that if an AI intelligent enough to improve on its own design were created, then the singularity will occur and the future after that is unpredictable, for better or for worse. AI might bring about the greatest technological change every, and with this comes huge risks that have to be remedied in the near future before it's too late.

ii) **How plausible do you find the scenarios Hawking poses?**
I think the scenarios are plausible, although the amount of time before an AI can pose a real threat is likely quite distant. Very little progress has been made in general AI, which is required for a real thinking machine. AI research currently is more focussed on performing and automating practical tasks that only humans have been able to do in the past, such as driving cars. While these systems have intelligence, this category of programs have a very limited scope of intelligence and no hope of developing their own intentions and posing risks. The only risks they can pose beyond their creator's wishes is working sub-optimally and making mistakes. Of course more intelligent weapons could be utilized, but we've had weapons that don't require intelligence to be enormously destructive for decades. I think we'll have plenty of warning if a real general purpose AI is on the horizon. There will need to be much incremental improvement before we really have to plan on how to avoid risks from AI becoming more intelligent than us.

iii) **What do you think AI researchers should do, if anything, to avoid these scenarios?**
In the future when we might begin to have generally intelligent AI systems, I think a lot more care would need to be taken. There might be a point at which an AI system will have reasoning capabilities that approach those of humans. At this point further development should be regulated. These systems should be carefully monitored in a sand-boxed environment (eg. a virtual world) . If the AI isn't given access to the internet or to additional computational resources then a safe level of experimentation

can be maintained. However I think if research gets to this level somebody is inevitably going to try to unleash these systems into the world.