CS 486 - A1

Alexander Maguire
amaguire@uwaterloo.ca
20396195

January 11, 2015

Problem 1 i:

The imitation game is a test proposed by Turing in which two participants (A and B) both try to convince a third party that they, the participants, are in fact person A. The participants communicate with the third party indirectly so that only their words can be used as evidence to determine who is who.

Problem 1 ii:

Turing thinks the imitation game is worth studying because it abstracts away the notion of intelligence from our pre-convictions of what it means to be intelligent; if a computer can successfully convince a human third party that it is a human, then for all intents and purposes relative to the third party, *it is*.

Problem 1 iii:

Turing anticipates many objections to his propounding that machines can think; two of which will be considered for this assignment. The first of which can be described as follows: ”if machines could think it would be very bad; so let us stick our fingers in our ears”. Turing's rebuttal of this point is that the objection in question is not in fact an objection whatsoever; he says that he doesn't think ”[the] argument is sufficiently substantial to require refutation”.

Additionally, Turing anticipates the objection that ”computers will never be able to do X” where X is some poorly defined aspect of humanity, for example ”love” or ”do something really new”. Turing refutes this by alluding to the fact that these things are black boxes, and that we humans only won't assign the ability of machines to do so because we haven't yet seen them do so.

Problem 1 iv:

I am unable to find two objections from the paper which I believe to be relevant. The mathematical objection is one which might be relevant; Gödel's incompleteness theorem does indeed exclude explicitly self-referential computers, though this limitation is likely capable of being skirted around with more modern theorems in mathematical logic. For example, Löb's theorem allows proving things about the underlying universe by transforming proofs of proofs into proofs themselves.

Problem 1 v:

TODO: Find or invent an objection not considered by Turing. Describe it briefly (with citation if applicable), and offer a response to it.

Problem 2 i:

The Chinese Room is a thought experiment proposed by Searle in which a human operator, whom by assumption understands no Chinese whatsoever, is given a book in English which describes how to manipulate Chinese writing in a deterministic fashion. By definition, Searle says, this book produces Chinese output – based on Chinese input – which is indistinguishable from a native Chinese speaker.

Problem 2 ii:

Searle thinks the Chinese Room thought experiment invalidates strong AI, in that the operator with the book is acting isomorphically to a computer, but by assumption doesn't understand any Chinese. Therefore, no computer can ever understand Chinese, and thus no computer can ever do everything a human can do.

Problem 2 iii:

The ”many mansions” reply, is an objection posited to Searle that just because current digital computers can't create strong AI (assuming Searle's argument holds), that is no argument that such strong AI is unconstructable *in general*. Searle's rebuttal is that this is missing the trees for the forest, that such an argument is ”you might be wrong” without any actual reasons behind it.

ANOTHER ONE

Problem 2 iv:

What does Searle think is the main difference between human cognition and Strong AI?

Problem 2 v:

I completely disagree with Searle; his choice of imagery for framing the discussion should be considered deceptive at best. As is so often the case in bad philosophy, the issue can be cleared up with a brief interlude into complexity theory. Searle's script and story construction can be viewed as isomorphic to an embedding of a deterministic finite automaton (DFA).

If we assume, as Searle does, that a human operator with this book is performing equivalently to a computer, we must find a means of transforming an arbitrary program into a DFA. This is possible in general by creating a DFA with 2^n states, where n is the size of the input space, putting such a DFA firmly in **EXPSPACE**. The resulting DFA is therefore non-instantiable in any universe.

If we then assume that the number of possible Chinese sentences is infinite (which is trivially constructable by mathematical statements into Chinese), the number of states to perfectly respond to any Chinese input must be in **EXPSPACE**. For any input space ¿ 240 (likely a safe assumption for Chinese), such a book would use more atoms than exist in the universe. Consulting any such book would be its own insurmountable task.

Searle's argument thus breaks down in its assumptions; such a book can not possibly be equivalent to a computer which would understand Chinese, in that a Chinese-understanding computer would *by necessity* need to do something smarter than table lookups in order to avoid complexity limitations.

Problem 3 i:

Hawking's main concern is that we are not taking AI research seriously enough; that any significant AI is going to be much smarter than we are, and at that point, it will be too late to stop it if the AI in question is not friendly to humans.

Problem 3 ii:

I find Hawking's scenario extremely plausible. Nick Bostrom's ”Superintelligence” (2014) outlines what he calls ”the orthogonality thesis”, that intelligence levels and goals are completely independent to one another; that just because an AI is smart doesn't mean it will care about any of the same things we care about. I find such a scenario very frightening.

Problem 3 iii:

I think AI researchers should spend their time working on finding means of proving friendliness of AI. I suspect that the AI industry will not do this research for them, since there is economic value in having the first AI, but none in having a safe AI, and thus the AI researchers must figure this friendliness out beforehand so it is usable when the field has a need for it.