Strong Artificial Intelligence, the Chinese Room, and Intuition

Alexander Meinhof

Ger 179C: Computing Cultures

Prof. Offert

17 March 2021

Introduction

In this paper I will evaluate John Searle's *Chinese Room Argument* against Strong AI. I will object to the conclusion of the *Chinese Room Argument* on the grounds that it depends on unreliable intuitions.

The *Chinese Room Argument* is an argument against what Searle calls 'Strong AI'. Strong AI is the claim that, given the right program, a computer could literally be a mind. That is, that a computer could literally have mental states like you or me. Put more succinctly, we can state the claim of Strong AI as the following:

> *Strong AI*: An appropriately programmed computer really is a mind.

The *Chinese Room Argument* is an argument that Strong AI is false, an appropriately programmed computer can never be a mind. The *Chinese Room Argument* asks us to consider the following scenario:

> *The Chinese Room Scenario*[1]
>
> Imagine I am locked in a room with a large set of cards with Chinese characters written on them. I do not know any Chinese. To me, different symbols in Chinese have no meaning. Imagine now that I am given a set of rules, in English (which I understand) which relate certain Chinese characters to others. Outside the room, someone slides a small set of Chinese symbols under the door. I consult the rules and return the appropriate set of Chinese characters from my large set. Unbeknownst to me, the characters I am given are questions, and the characters I return are answers. Eventually, I become so good at consulting the rulebook and returning the right sets of characters that the answers I slip under the door become indistinguishable from those of a native Chinese speaker.

Searle thinks that the various elements of this are analogous to the actions of a computer executing a program. The set of symbols initially in the room is 'the database' and the English ruleset is the 'the program'. Searle's argument is that despite my answers being indistinguishable from those of a native Chinese speaker, I do not *understand* Chinese. I am merely manipulating formal symbols, and the mere manipulation of symbols does not produce understanding. Since

[1] Searle (1980 p.417-418)

every element of the *Chinese Room Scenario* models a computer running a program, Searle thinks that the conclusion we should draw from this scenario is that **no** computer program can produce understanding. We can express this argument as follows:

> *The Chinese Room Argument*
>
> P1: Programs manipulate purely formal, syntactical objects.
>
> P2: The manipulation of purely formal, syntactical objects does not produce mental content.
>
> P3: A mind must have mental content.
>
> C: Therefore, programs cannot be minds.

Premises 1 and 3 of this argument are uncontroversial. The bulk of the work in this argument is done by premise 2. The *Chinese Room Scenario* is a defense of premise 2. In the scenario I am manipulating purely formal, syntactical objects, yet I do not understand the content of those objects. Searle responds to a number of objections to this argument. I will not consider those objections here. Rather I wish to object to Searle's argument on the grounds that it is an intuition pump. The Chinese Room Scenario seriously manipulates our intuitions by conflating two systems with vastly different levels of complexity.

Distorted Intuition in the Chinese Room Scenario

The inconceivability of Searle's scenario distorts our intuitions to the point where they are unreliable. The *Chinese Room Scenario* seriously manipulates our intuitions by conflating the execution of a set of rules of a rulebook, by hand, on a set of cards with symbols, with to the execution of a computer program and natural language processing. There are two significant factors which serve this manipulation, time and scale. To see how time and scale can color our judgement in the ability of formal systems, consider the following scenario:

> *Water Pipe System*
>
> Consider a system of water pipes, much like the standard PVC pipes available at a hardware store. These pipes, along with pumps, switches, gates etc. are arranged in such a way that they can produce logic gates. For example, one arrangement can produce a

> not-and gate, in which an output pipe is filled with water when 1 or 2 of the input pipes are empty. The pipe system is Turing-complete.

It is unintuitive to suggest that this water pipe system would be capable of rendering complex 3d geometry and lighting, or real time physics simulation, or the synthesis of realistic human faces though a neural network. When presented with a system of water pipes, the thought that the system could be expanded to do such operations is unfathomable. However, there is no logical reason to think that the water pipe system could not do those things. Insofar as the pipe system is Turing-complete, the pipe system could *theoretically* perform any of these operations. So, what can explain why our intuitions about the pipe system depart so far from its actual capabilities? The answers are time and scale.

Water flows much slower than electrons, so our water computer would run in orders of magnitude far slower than an average computer. Simple addition, for example, might take a few minutes. A simple image render might take years. A 3d model render might take centuries. The water pipe system would be capable of doing these things, despite it being outside the timescales of human lives.

Furthermore, our intuitions are colored by the fact that a computer constructed under the water pipe system would take up an imperceptible amount of space. Let us imagine that the analog to an electronic ‘transistor’ in the water pipe system would take up roughly an area of one square meter. Let us assume that a modern computer processor (like the one in the device I am writing this on) has around 1 billion transistors[2]. With these numbers, the water-pipe analog of the transistors of a common household CPU would cover the entire city of Los Angeles. If we factor in a GPU, a screen, as well as all of the additional machinery required such as power distribution, potential cooling etc., we can easily imagine that our water pipe computer would span the area of three or four major metropolitan areas. Both the issues of time and scale heavily color our intuitions about the capabilities of the water pipe system. However, to adopt Searle’s phrase, I take no “philosophical ice to be cut”[3] by this fact. Time and scale can heavily influence our

[2] This is a conservative estimate.
[3] Searle (1980 p.419)

intuitions to the point where the depart from the logical truth. With the water pipe example in mind, we can see how the time and scale color our intuitions about Searle's argument.

Let us first consider the time. Searle asks us to imagine ourselves acting out a computer program by hand. However, actually doing this would take immense amounts time. To manipulate a set of Chinese symbols, by hand, in a way emulating even primitive question-answer programs is an inconceivable task for a human to do in a reasonable amount of time. This is a point Searle glosses over, one which contributes to his misrepresentation of the relationship between symbol manipulation and the mind. Consider how long it takes a human to sort a deck of shuffled cards versus a computer to sort a similarly organized digital dataset. To sort a shuffled deck of cards would require, in the best-case scenario, a couple minutes. A computer could perform this operation in microseconds. Now consider how fast you can recall the meaning of the word 'phantom'. Insofar as you know the meaning of this word, you can recall that meaning very quickly. The symbolic manipulation in the *Chinese Room Scenario* is so far divorced from both the speed of the operations of a computer and the speed of the operations of natural language processing. This point is succinctly stated by Dennett and Hofstadter:

> "the reader is invited to identify with Searle as he hand-simulates an existing AI program that can, in a limited way answer questions of limited sort, in a few limited domains. Now, for a person to hand-simulate this, or any currently existing AI program that is, to step through it at the level of detail that the computer does would involve days, if not weeks or months, of arduous, horrendous boredom" (374).

Dennett also states:

> "speed, however, is 'of the essence' for intelligence'" (326).

Our intuition that the subject of the *Chinese Room Scenario* lacks understanding can be attributed to the vast differences in speed between operations of natural language versus the manipulation of symbols from a set of rules. The argument of whether computers possess the necessary processing speed to execute the operations of natural language is a question of speculative science, far removed from the a-priori argument Searle hopes to propose[4].

[4] Dennett advances an argument that computers are **not** capable of this level of processing. But he to admits this is a speculative argument of 'scientific likelihood'. (p.327).

Let us now consider space. In the *Water Pipe Example*, space colored our intuitions because the water pipe computer would have to occupy and inconceivable amount of physical space to match the ability of a common household computer. In the *Chinese Room Scenario,* the issue of space is not an issue of physical space, but rather the size of the ruleset which the subject would have to consult. Dennett and Hofstadter state:

> "Searle envisions [the subject] as having absorbed what in all likelihood would amount to millions, if not billions, of pages densely covered with abstract symbols, and moreover having all this information available, whenever needed, with no retrieval problems" (375)

Let us seriously consider the number of rules necessary in the *Chinese Room Scenario*. Firstly, there would have to be rules categorizing symbols as nouns, verbs, adjective, adverbs etc. These rules would take the form of conditionals of the form:

> If [(input) = Symbol X] then [(input) = noun or adverb or verb or adjective]

One such conditional would have to exist for every available symbol. Already we have a set of rules as large as the set of symbols, meaning that properly categorizing the symbols would require a number of operations the size of the set of symbols squared. These sort of equivalences between large sets of information are trivial for a computer to perform, but inconceivable for a human to do by hand by consulting a set of rules. Furthermore, in cases of natural language processing, we are capable of implicitly internalizing large sets of rules, such as the ability to distinguish words into different classes or the immense number of rules which govern the creation of sentences. Again, we see a huge gap between the operations of a computer, natural language processing and the actions of the subject in the *Chinese Room Scenario*. Space, in this case the size of the ruleset, colors our intuitions in service of Searle's conclusion.

In conclusion, time and scale are both axis by which the *Chinese Room Scenario* alters our intuitions about the relationship between language understanding, computer programs, and the execution of symbolic manipulations by hand from a set of rules.

Implications for Computing

Supposing the points I raise in this paper are sound, the broader implication of this paper is that it is still up in the air whether correctly instantiated computer programs could really be *minds.* That is, the project of strong AI is not a-priori doomed. Interestingly, the *Chinese Room Argument* echoes an argument which far preceded it: Lady Lovelace's Objection to the idea that machines could think, an objection responded to by Turing in 1950[5]. Lady Lovelace's objection is that machines cannot 'take us by surprise' since we always know how they will respond to certain inputs. Turing's response is that machines can 'take us by surprise' since the functions which they perform can be complex enough to produce unpredicted results. Searle's claim is an adjacent one. Searle believes that the complexities of the mind cannot be represented by syntactical input-output functions. But Searle's view of what input-output functions are is severely limited. Just like Lady Lovelace, Searle fails to grasp the possible complexity of input-output functions.

Citations

Dennett, Daniel C, and Douglas R Hofstadter. "22: Minds, Brains, and Programs." *The Mind's I*, Bantam Books, 1981, pp. 353–382.

"Fast Thinking." *The Intentional Stance*, by Daniel Clement Dennett, MIT Press, 1987, pp. 323–337.

Searle, John. "Minds, Brains, and Programs." *The Behavioral and Brain Sciences*, 1980, pp. 417–457.

Turing, Alan. "Computing Machinery and Intelligence." *Mind*, Oct. 1950, pp. 433–460.

[5] Turing (1950 pp.450-451)