JOHN SEARLE'S CHINESE ROOM ARGUMENT

John Searle begins his (1990) ``Consciousness, Explanatory Inversion and Cognitive Science" with

> ``Ten years ago in this journal I published an article (Searle, 1980a and 1980b) criticising what I call Strong AI, the view that for a system to have mental states it is sufficient for the system to implement the right sort of program with right inputs and outputs. Strong AI is rather easy to refute and the basic argument can be summarized in one sentence: {\it a system, me for example, could implement a program for understanding Chinese, for example, without understanding any Chinese at all.} This idea, when developed, became known as the Chinese Room Argument."

The Chinese Room Argument can be refuted in one sentence:

Searle confuses the mental qualities of one computational process, himself for example, with those of another process that the first process might be interpreting, a process that understands Chinese, for example.

Here's the argument in more detail.

A man is in a room with a book of rules. Chinese sentences are passed under the door to him. The man looks up in his book of rules how to process the sentences. Eventually the rules tell him to copy some Chinese characters onto paper and pass the resulting Chinese sentences as a reply to the message he has received. The dialog continues.

To follow these rules the man need not understand Chinese.

Searle concludes from this that a computer program carrying out the rules doesn't understand Chinese either, and therefore no computer program can understand anything. He goes on to argue about biology being necessary for understanding.

Here's the refutation in still more detail.

Assume the process is a good participant in an intelligent Chinese conversation, i.e. behaves as though it understands Chinese. What is required for that we'll discuss shortly. The so-called Berkeley answer is that the system, consisting of the man and the book of rules, understands Chinese.

Our answer is an elaboration of the Berkeley answer. A computer interprets computer programs, i.e. carries them out instruction by instruction. Indeed a program can interpret other programs, e.g. a Lisp or Java interpreter interprets, i.e. carries out, Lisp or Java programs. We speak of the interpreter as carrying out the Lisp program, although this could be elaborated to saying that the computer carries out the Lisp interpreter which is carrying out the Lisp program step by step.

Indeed a time-shared operating system can carry out many different programs at once, some may be in machine language, others may be in Lisp, C, Fortran or Java. Suppose one of these programs is a Lisp program carrying out an intelligent Chinese conversation with someone at a terminal. Suppose another program is carrying out an intelligent French conversation or a different Chinese conversation with someone at a different terminal. Assume that these conversations are normally considered to require an understanding of Chinese or French. *What understands Chinese?*

We don't want to say that the computer understands Chinese and French but rather that the respective programs understand Chinese and French respectively. Indeed if we have two Chinese conversation programs, one may understand Chinese well and the other hardly at all.

Returning to the man in the room. He can be carrying out a conversation in English or playing chess while he is interpreting the book of rules for a Chinese conversation. Indeed he may have memorized the book of rules and be carrying them out in his head. As with the computer programs, it's the process that understands Chinese well or badly.

Let's consider some practicalities that may help us understand the question better. There are two extreme levels on which the man may be carrying out the Chinese conversation. One level is that of Joseph Weizenbaum's 1965 program ELIZA. It makes sentences by re-arranging and transforming the words in the input sentence. Thus one version, called DOCTOR, and included in the Xemacs editor, replies to "My mother hates me?" with "Why do you say mother hates you". According to Weizenbaum (personal communication), ELIZA requires so little computation that it can be carried out by hand. Thus an ELIZA level Chinese room is entirely feasible.

Does an ELIZA level Chinese room understand Chinese? It depends on what you mean by "understand", but I would prefer to say that a Chinese ELIZA does not understand Chinese. We'll see why?

Now consider a Chinese room that passes the Turing test, i.e. the Chinese interlocutor cannot be sure whether he is conversing with an intelligent fellow Chinese speaker. This is not feasible with a man and a book of rules. In fact it is beyond the present state of the art in artificial intelligence. While the book of rules probably needn't be bigger than an ordinary encyclopedia, I doubt that a human could carry out the rules at better than 10^{-9} of the speed required for conversation.

What is required for a Chinese room that passes the Turing test?

1. A knowledge base of facts about the world, e.g. about 3-dimensional objects and the fact that they fall when unsupported and end up on the floor or ground.
2. A knowledge base of facts about Chinese life and the Chinese language.
3. A representation of the conversational purpose of the program.
4. A program that translates the sentences into some internal form and responds appropriately, given the motivations we have

given the program.
5. A program that translates the output sentences into Chinese, prints the result, and pushes it back under the door.

These requirements can, at least in principle, be implemented in a variety of ways, e.g. by a sequentially operating neural net or by a logic based reasoner. I think the latter approach can do more now and will approach the goal of a human level conversation sooner.

So what is it to understand Chinese?

Understanding Chinese involves being able to translate Chinese sentences into some internal representation and to reason with the internal representation and some knowledge base. Thus understanding "Tom is an airplane pilot." requires being able to correctly answer, "Does Tom know how rotating the control column left affects the ailerons?"

More about understanding is discussed in my **Making Robots Conscious of their Mental States**.

More Searle arguments

> ``Once we get out of that confusion, once we escape the clutches of two thousand years of dualism, we can see that consciousness is a biological phenomenon like any other and ultimately our understanding out it is most likely to come through biological investigation'' John Searle - New York Review of Books, letter pp 58-59, 1990 June 14.

My view is that consciousness is an abstract phenomenon, currently best realized in biology, but causal systems of the right structure can also realize it. See **Making Robots Conscious of their Mental States**.

The discussion of the Chinese Room has remained at an excessively high level on both sides. I propose to discuss what would actually be involved in a set of rules for conducting a conversation in Chinese, independently of whether these rules are to be carried out by a human or a machine.

First we must exclude various forms of cheating that aren't excluded by Searle's formulation of the problem.

1. We need to exclude a system like Weizenbaum's Eliza that merely looks for certain words in the input and makes certain syntactic transformations on each sentence to generate an output sentence. I wouldn't count such a program as understanding Chinese, and *a fortiori* Searle wouldn't either. The program must respond as though it knew the facts that would be familiar to an educated Chinese.

2. If the rules are to be executed by a human, they must not involve translating what was said into English, e.g. by giving the dictionary entries for the characters. If this were done, the English speaker could use his own understanding of the facts of the world to generate English responses that he then translates into Chinese. The database of facts must not be in English. We also suppose that the human is not allowed to do cryptanalysis to translate the inputs or the database into English.

This eliminates the forms of cheating that I can think of, but I don't guarantee that there aren't others.

How shall we construct our program? Artificial intelligence is a difficult scientific problem, and conceptual advances are required before programs with human level intelligence can be devised. Here are some considerations.

1. In discussing concrete questions of intelligence, it is useful to distinguish between a system's algorithms and its store of facts. While it is possible in principle to consider the facts as built into the algorithm, making the distinction is practically essential for studying both human and machine intelligence. We communicate mainly in facts even when we are trying to tell each other algorithms.

2. The central problem of AI is, in my opinion, achieving goals in the *commonsense informatic situation* See my **What is artificial intelligence?** for more on this.

Searle offers four axioms.

1. Brains cause minds.

"Cause" makes me a little nervous. If he only means that the human mind is an abstraction of part of the operation of the brain, I'll agree.

2. Syntax is not sufficient for semantics.

This purported axiom is slippery. Does he just mean that defining a language, whether a natural language, first order logical language, or a programming language, requires defining what the expressions of the language mean? If that's what he means, I agree.

3. Computer programs are entirely defined by their formal, or syntactic structures.

This is ok provided we remember that the programming language has a semantics, and the data structures used by the program must have semantics if the program is to be intelligent.

4. Minds have mental contents; specifically they have semantic contents.

That's ok with the above provisos.

Conclusion 1. No computer program by itself is sufficient to give a system a mind. Programs, in short, are not minds, and they are not by themselves sufficient for having minds.

The conclusion doesn't follow from the axioms, not even informally.

I should remark that Searle's Chinese room argument hasn't convinced very many of his fellow philosophers.

In his *Scientific American* article on the Chinese room Searle makes an interesting mistake, though not a new mistake. He writes that a transcript of the Chinese conversation could *equally well* represent the score of a chess game or stock market predictions. This will only be true if the Chinese conversation is very short; perhaps it would have to be less that 20 characters - or maybe it's 100 characters.

We have to haggle about what *equally well* means. We can get a 1-1 correspondence between Chinese dialogs and chess scores by enumerating Chinese dialogs and enumerating chess scores and putting the *n*th dialog correspond to the *n*th score. This isn't good enough. Both Chinese dialogs and chess scores have meaningful substructures, and the previously described correspondence does not make the substructures correspond. One structure is that of initial segments. The initial segment of a Chinese dialog is meaningful to a Chinese, and an initial segment of a chess score is meaningful to a chess player, and these meanings related to the meanings of the whole dialog and the whole score respectively.

All this relates to the notion of *unicity distance* in cryptography. A simple substitution cryptogram that has less than 21 letters is likely to have several interpretations. With more than 21 letters the interpretation is extremely likely to be unique. That's why people can solve cryptograms.

I think there is a mathematical theorem stating that meaningful strings in a structured language have unique interpretations if their lengths exceed some rather small bound. I don't know how to formulate such a theorem.

I don't know whether this mistake of Searle's is related to his Chinese room mistake. It seems to me that Quine's assertions about "the indeterminacy of radical translation" are based on too small examples. However, I may be misunderstanding what Quine was claiming.

Send comments to mccarthy@stanford.edu. I sometimes make changes suggested in them. - John McCarthy

The number of hits on this page since 2001 September 28.