*In memory of Ray Reiter, a wise man*

# Searle and the Art of Motorcycle Maintenance

Yoav Shoham
Stanford University

**Abstract** The Chinese Room thought experiment was put forward by John Searle in the early days of AI as an argument against the prospects of endowing computers with artificial intelligence. The Chinese Room argument does have some merit as a basis, albeit a confusing one, for discussing what it means to 'understand' something. But it certainly sheds no light on the prospects for AI. This note explain why this is the case, in my view.

## Preface

I recently had the occasion to hear John Searle speak about the Chinese Room Argument (CRA).[1] I was surprised to hear such prominent discussion of an argument that never had much force in my view, and that I was sure had long been put to rest. Since its inventor seemed as enthusiastic about it as ever, and a fraction of the audience seemed not obviously dismissive of it, I decided to write down my thoughts on it. During the same time Ray Reiter was dying of cancer. Although to my knowledge Ray would never have spent the time writing about such matters, I do think his dry humor and intellectual honesty would have made him like this discussion. I was too late getting this to him during his life, so this is in his memory.

## The argument

Imagine a person sitting in an sealed room. Inside the room is a machine, which provides the sole contact between the person and the outside world. The machine accepts questions written in Chinese from persons outside the room. The person answers the questions in writing, via the same machine, also in Chinese. His answers are perfect, even though he doesn't know a word of Chinese. The way he does it is by consulting a rulebook, written in English. Each rule in the book is of the form "If you get a message with the following symbols send back a message with the following symbols". Thus, the argument goes, the person displays a perfect understanding of Chinese, even though he does not understand Chinese – he has no idea what the questions mean, what the answers mean, indeed perhaps even that these are questions and answers. In just the same way, the argument continues, even if you programmed enough rules into a computer so it could function as if it understands a certain topic, in fact it would just be an electro-mechanical device going through the motions; it would no more understand anything than a toaster does.

[1] The occasion was Searle's lecture at Xerox PARC on August 15, 2002. The subject of the lecture was brain research as a new foundation for cognitive science, with ramifications to the study of consciousness. I won't have anything to say about these issues. However, in service of these issues, Searl devoted a substantial portion of his lecture – at least a quarter, in my estimate – to the CRA. It is only that part that I am addressing.

**From understanding Chinese to motorcycle maintenance**

At first there is something intuitively appealing about the CRA, which is why it struck a chord with people when first introduced and why apparently it still has some life left in it. We do feel that there is a difference between engaging in a meaningful dialog with someone on the one hand, and mechanically exchanging with them uninterpreted symbols on the other. This sentiment is often mixed up with other notions – that we relate to the semantics of language, not only its syntax; that we experience "qualia", and more.

But there is a combination of confusion and slight of hand taking place in the CRA. The confusion regards the notion of 'understanding'; the CRA doesn't discuss it explicitly, but implicitly assumes that it is a self-evident, absolute notion. In fact it is a relative notion; we understand things within certain scopes, and to different degrees. Indeed, if the CRA has merit, it is as a basis for discussion of what it means to understand something. The slight of hand consists of the assumption about the existence of the rulebook; we AI researchers should be so lucky. Here is a bit more about both.

One problem with the CRA is its extreme implausbility, and so it will be useful to consider alongside it a different example, which I'll call the 'GS argument', or GSA for short. Imagine a person who has the maintenance manual for the 2002 BMW 1100GS, a neat if somewhat odd dual-purpose motorcycle. The manual is in English, which the person understands very well even though he is from America. The person doesn't own such a bike, and has never even seen one. While we're at it, let's assume he's never heard the term 'motorcycle' until he picked up the manual, and does not know even that it is used for transportation. He certainly doesn't have a clue what a cylinder or a drive shaft is, nor the meaning of any of the hundreds of other terms in the manual. However the manual is very detailed, and its Troubleshooting section has a comprehensive list of rules; for every possible question about the bike, it contains an answer. So the person opens a website called GSguru.com, and flawlessly fields all maintenance questions about the 2002 BMW 1100GS, based solely on the manual. We won't discuss whether this person has a viable business on his hands, but we will ask this: Does this person understand motorcycle maintenance or not?

To answer this we need to understand the meaning of 'understand'; one has to define the requirements before claiming that someone (or something) doesn't meet them. The CRA narrative doesn't do it, so let's take a stab at it here. In the rhetoric heard around the CRA I could tease apart three types of requirement:

1. The person must be able to correctly answer questions; this seems to be the main requirement.
2. The person's actions must reflect his understanding.
3. The person must experience appropriate emotional responses to the information.

Let's consider these in turn. On the first count the situation is unambiguous; by assumption the person can (in the CRA) answer questions in Chinese, or (in the GSA) answer questions about motorcycle maintenance. But there is a certain slight of hand

here; we haven't been precise about the class of questions the person is required to answer. Implicit in both arguments is the assumption that the class is large. If there were only (say) two questions the person was required to field, no one would ascribe understanding to him and the argument would fall apart. In the CRA the class seems to be ludicrously large – something like all sentences in Chinese. But researchers in natural language processing would kill for such a universal rulebook. Indeed, there are strong arguments by such researchers that intelligent dialog systems require deep semantic knowledge. If these arguments are correct, then the rulebook would not be able to exist without it also capturing the meaning of sentences. Thus the anscillary argument that one hears within the CRA, namely that the person has access only to the syntax of the sentences but not their semantics, is highly suspect. At the very least it is clear that the CRA makes such an extreme assumption that one cannot apply any commonsense intuition to it.

Indeed, a similar slight of hand is present also in the GSA, if more subtly. The assumption that the online mechanic can truly answer *any* question relating to motorcycle maintenance – questions of arbitrary legal syntax and vocabulary, questions that make reference to other notions such as the danger of motorcycles or the free spirit of their riders – is again highly implausible, or at least is well ahead of the state of the art. So again, unless one carefully circumscribes the set of questions to which the mechanic is held accountable, the GSA embodies an assumption that defies intuition.

The upshot of all of this that one has to describe the *scope* of understanding being evaluated. Both the GSA and the CRA assume scopes so large so as to render the examples uninformative.

The second kind of requirement we contemplate is that the person act based on his understanding. If someone asks you in Chinese whether you know that a car is about to hit you and you reply 'yes' pleasantly but don't jump out of the way, you don't really understand what was said. I actually am not sure that this is a reasonable requirement; it could be argued that this requirement confuses understanding with rational decision making. But if one wants to make this requirement in the case of the GSA or the CRA, one has to be fair, and equip the actor with the sensors and effectors to act. In the case of the GSA it might be a visual-auditory system as well as mechanical hands to manipulate the motorcycle, and maybe a body to ride it. I'm not sure what the set of actions that might be relevant in the CRA, and thus also not what sensors and effectors would be needed. But in either case there is absolutely no reason to assume that such sensors and effectors cannot be created, and indeed no such claims have been made to my knowledge.

But one does hear arguments against meeting the third requirement, that the person experience the sentences emotionally. In the words of the late Miles Davis, "If you're not nervous, you're not paying attention." In particular, there is discussion of "qualia", that mysterious immediate experience that is unmediated by language or thought, and of "consciousness".

It seems to me that this third requirement is unclear, almost mystical. First, here again one could argue that the emotional responses are only accidentally correlated with understanding. But even ignoring this, the problem is that emotions, qualia, and consciousness are not well understood, and to the extent that they are there is no compelling argument that I am aware of that machines cannot be endowed with them. Indeed, there have been specific arguments in AI that emotions should and can be built into intelligent machines. In my opinion this is a fascinating discussion, but a very nascent one and neither side can use it to argue for or against the possibility of understanding by machines.

**Lessons from Chelm**

In East European Jewish folklore, the mythical city of Chelm is inhabited by uniquely foolish people. In a series of short stories, the wise men of Chelm display every folly imaginable. In one of the stories, one of the wise men – in some versions it's Reb Zelig the tailor, in others it's a fellow named Getsel – decides to make the long journey from Chelm to Warsaw, about which he'd heard so much.[2] Early in the morning he takes his leave from his loving wife and children, and sets out on foot. Halfway into the journey he becomes tired, and stops to sleep under a tree by the side of the road. However, in order to remember the right direction, he take off his shoes and points them in the direction he has been walking. While he is asleep a carriage drives by and one of its wheels hits the shoes, turning them around so they point in the opposite direction. When the wise man wakes up he puts on his shoes and starts walking in the direction the shoes were pointing. Eventually he gets to his destination. He is struck by its resemblance to his hometown of Chelm. Out of curiosity he follows the streets until he comes to a street that, he could swear, is the spitting image of his own street back home. In fact, there's a house there that's just like his. And what do you know, out of the house come a woman and her children who hug and kiss him, as if they were his family. The wise man realizes that the man of the house must look a lot like him, and is struck by the similarity between this family and his own. The family has obviously missed the man of the house, since they won't let the wise man go. The wise man sees he has no option but to stay until the real man of the house returns, at which time the family will realize its mistake. He spends the rest his life in that town, but, as pleasant as this new family and the town folk are to him, he remains forever homesick.

And so it is with Searle's computer. One day the computer will function perfectly, whether this means conversing on all possible matters in Chinese, or, more plausibly, answering questions on motorcycle maintenance. But the computer will always feel inferior, because it doesn't *really* understand.

*Palo Alto*
*September 2002*

[2] Different versions of the story exist; I've taken some liberties myself in the following.