A view inside the Chinese room

Mark Bishop

Introduction

Perhaps the most famous critic of computational theories of mind is John Searle. His best-known work on machine understanding, first presented in the 1980 paper 'Minds, Brains & Programs' (MBP), has become known as the Chinese Room Argument (CRA). The central claim of the CRA, are that computations alone cannot, in principle, give rise to cognitive states, and that they therefore computational theories of mind cannot fully explain human cognition. More formally Searle (1994) stated that the Chinese Room Argument (CRA) was an attempt to prove the truth of the premise:

1. Syntax is not sufficient for semantics

… which, together with the following two axioms:

2. Programs are formal, (syntactical)
3. [Human] Minds have semantics, (mental content)

… led him to conclude that 'programs are not minds' and hence that computationalism - the idea that the essence of thinking lies in computational processes and that such processes thereby underlie and explain conscious thinking - is false.

Although the last thirty years have seen tremendous controversy over the success of the CRA, there has emerged a great deal of consensus over its impact: Larry Hauser called it, "*perhaps the most influential and widely-cited argument against the claims of Artificial Intelligence*"; Stevan Harnad, editor of Behavioural and Brain Sciences, asserted that, "*it [the CRA] has already reached the status of a minor classic*"; Anatol Rapaport claims it, "*rivals the Turing test as a touchstone of philosophical inquiries into the foundations of AI*" and most recently this author has co-edited a volume of new essays reflecting on the argument, (Preston & Bishop, 2002).

In the CRA Searle emphasizes the distinction between syntax and semantics to argue that while computers can follow purely formal rules[1], they cannot be said to know the 'meaning' of the symbols they are manipulating, and hence cannot be credited with 'understanding' the results of the execution of programs those symbols compose. In short, Searle claims that Artificial Intelligence (AI) programs may simulate human intelligent behaviour, but not fully duplicate it.

And yet it is also clear that Searle believes that there is no in principle barrier to the notion that a machine can think and understand; indeed in MBP Searle explicitly states in answer to the question, "Can a machine think?" that "the answer is obviously yes". So the CRA is not a critique of machine intelligence per se; simply of any form

[1] A claim Searle has later questioned.

of computationalism, according to which a machine could carry genuine mental states (e.g. genuinely understand Chinese) purely in virtue of carrying out an appropriate series of computations.

Historical Background

Following work on the automatic analysis of simple stories, a cultural context emerged within the Artificial Intelligence (AI) community that appeared comfortable with the notion that appropriately programmed computers were able to 'understand' such stories, a concept which can be traced back to the publication of Alan Turing's seminal paper 'Computing Machinery and Intelligence' (Turing 1950).

For Turing, emerging from the fading backdrop of Logical Positivism and the Vienna Circle, conventional questions concerning 'machine thinking' were too imprecise to be answered scientifically and needed to be replaced by a question that could be unambiguously expressed in scientific language. In considering the metaphysical question 'Can a machine think?' Turing arrived at another, distinctly empirical, question of, whether, in remote interaction (e.g. via email) with both a computer and a human, a human interrogator could reliably identify which was which. If the interrogators performance was no better than chance, then the computer is said to have passed the 'Turing test'.

Yet one of the key points at issue when discussing the CRA is the adequacy of the Turing test, which many proponents of the Artificial Intelligence project continue to use as a criterion for the mental. Searle (1982) expresses this view as:

> "*The conclusive proof of the presence of mental states and capacities is the ability of a system to pass the Turing test … If a system can convince a competent expert that it has mental states then it really has those mental states. If, for example, a machine could 'converse' with a native Chinese speaker in such a way as to convince the speaker that it understood Chinese then it would literally understand Chinese*".

Then in 1977 Schank and Abelson published information on a program they created, which could accept a simple story and then answer questions about it, using a large set of rules, heuristics and scripts[2]. In the wake of this and similar work in computing labs around the world, some of the more excitable proponents of Artificial Intelligence began to claim that such programs actually understood the stories they were given, and hence offered insight into human comprehension.

It was precisely an attempt to expose the statements emerging from a vociferous proselytising AI-niks, and more generally to demonstrate the inadequacy of the Turing test, which led Searle to formulate the CRA.

In the CRA Searle argues that understanding of a Chinese story can never arise purely as a result of following the procedures proscribed by any computer program and in the MBP paper Searle offers a first-person tale outlining how he could instantiate such a

[2] A script is a detailed description of a stereotypical event unfolding through time. For example, a system dealing with restaurant stories would have a set of scripts about typical events that happen in a restaurant: entering the restaurant; choosing a table; ordering food; paying the bill etc.

program, produce correct internal and external state transitions, pass a Turing Test for understanding Chinese, and yet still not understand a word of Chinese.

The Chinese Room Argument

In the CRA Searle describes a situation where he is locked in a room and presented with a large batch of papers covered with Chinese writing that he does not understand. Indeed, Searle doesn't even recognize the symbols as being Chinese, as distinct from say Japanese or simply meaningless patterns.

A little later Searle is given a second batch of Chinese symbols together with a set of rules (in English) that describe an effective method (algorithm) for correlating the second batch with the, first purely by their form or shape.

Finally Searle is given a third batch of Chinese symbols together with another set of rules (in English) to enable him to correlate the third batch with the first two, and these rules instruct him how to return certain sets of shapes (Chinese symbols) in response to certain symbols given in the third batch.

Unknown to Searle, the people outside the room call the first batch of Chinese symbols, 'the script', the second set 'the story', the third 'questions about the story' and the symbols he returns they call 'answers to the questions about the story'. The set of rules he is obeying they call 'the program'.

To complicate matters further the people outside also give him stories in English and ask him questions about them in English to which he can reply in English.

After a while Searle gets so good at following the instructions and 'outsiders' get so good at supplying the rules which he has to follow, that the answers he gives to the questions in Chinese symbols become indistinguishable from those a true Chinese man might give.

From the external point of view, the answers to the two sets of questions, one in English the other in Chinese, are equally good. Searle, in the Chinese room, has passed the Turing test. Yet in the Chinese case, Searle behaves 'like a computer' and does not understand either the questions he is given or the answers he returns whereas in the English case he does. Searle contrasts the claim from members of the AI community - that any machine capable of following such instructions can understand the questions and answers – with his own continuing inability to understand a word of Chinese…

The 'logical reply'

In a chapter published as part of our collection, (Preston & Bishop, 2002), Jack Copeland challenged the validity of the basic CRA. Alongside many commentators he pointed out that the person in the room is not analogous to a computer executing an AI program, but rather just its CPU – its central processing unit – and that the claim of the AI researchers are not that the CPU understands Chinese, rather the computer as a whole – the CPU + memory + hard disc etc. – does. Copeland then simply observes that if the CRA is supposed to target the entire system of Searle, the room, the rulebook, the bits of paper etc., then the argument is not watertight for it becomes something of the form:

> "*No amount of symbol manipulation on the persons part will enable him to understand the Chinese input, therefore no amount of such manipulation will enable the wider system of which he is a part to understand that input*", (Copeland, 2002).

Copeland's 'logical reply' is very closely related to an earlier move which Searle anticipated in MBP – the 'systems reply'; that although the person in the room doesn't understand Chinese, the entire system of the room and its contents do.

Searle finds this response entirely unsatisfactory, (except by behaviourism and the Turing test) and responds to it by allowing the person in the room to internalise everything, (the rules, the batches of paper etc.) so that there is nothing in the system not internalised[3]. Now in response to the questions in Chinese and English there are two subsystems, the native English speaking Searle and the internalised Chinese room. Now, "all the same, he [Searle] understands nothing of Chinese, and a fortiori neither does the system, because there isn't anything in the system that is not just a part of him".

In a recent discussion of the CRA Haugeland (2002) calls the native English speaking subsystem Searle and the Chinese subsystem Huo and asks why should we should accept Searle's conclusion that that Huo doesn't understand Chinese given his response to Chinese questions are all correct? Haugeland points out that for Searle's conclusion to follow, it is must necessary be the case that if the Hao subsystem understands Chinese then so would Searle and yet concludes that Searle is not entitled to conclude this because when we say that the understanding a person has is in that person what we mean is that it is a capacity of that person whereas when we agree that what Searle means by in is that it is implemented by internal process of

Searle finds
Laughing Qualia:
Compare this to a situation where he participates in a joke in English. He may find it funny and laugh because he understands, whereas although he may make the right responses in Chinese if he executes the procedures correctly, he will never get the joke and feel laughter within...

The Simulation Fallacy:
If a computer simulation of a fire doesn't burn the neighbourhood down, why should the computer simulation of understanding, actually understand?

The Information Processing Fallacy:

[3] Searle also objects that it is somehow implausible to think that whilst a person doesn't understand Chinese, somehow a person plus bits of paper might and further that the systems reply is over liberal allowing all sorts of non-cognitive subsystems of the person in the room – stomach, heart, kidneys etc - to be cognitive, "since there is no principled way to distinguish the motivation for saying the Chinese subsystem understands from saying that the stomach understands".

The computer does not do information processing, as it does not know what it is processing. A bit might represent the day of the week, then weight of a body or the speed of a bullet.

However, in the twenty years since its publication, perhaps because of its ubiquity and the widespread background perception that, if it succeeds at all, its primarily target is Good Old-Fashioned AI (GOFAI), the focus of AI research has drifted into other areas: connectionism, evolutionary computing, embodied robotics, etc. Because such typically cybernetic approaches to AI are perceived to be the antithesis of formal, rule-based, script techniques, many working in these fields believe the CRA is not directed at them. Unfortunately it is, for Searle's rule-book of instructions could be precisely those defining learning in a neural network, search in a genetic algorithm or even controlling the behaviour of a humanoid-style robot of the type beloved by Hollywood.

The CRA argument, if sound, would undermine the very foundation any cognitive science grounded upon a computation theory of mind. The importance of Searle's argument, philosophically and practically - in its impact on the feasibility of current and proposed AI research programmes - has of course ensured that it has been widely attacked (and defended) with almost religious fervour.

That the CRA addresses both phenomenal and intentional aspects of understanding and intelligence is clear from the introduction to Searle's original paper, where we find Searle's definition of Strong AI:

But according to Strong AI the computer is not merely a tool in the study of the mind; rather the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states. (Searle 1980, p.417 (p.67 in Boden)).

An axial statement here is that, 'the appropriately programmed computer really is a mind'. This, taken in conjunction with, '[the appropriately-programmed computer] can be literally said to understand' and hence have associated 'other cognitive states', implies that the CRA also, at the very least, targets some aspects of machine consciousness – the phenomenal infrastructure that goes along with 'really having a mind'.

However it is also clear from literature on the CRA that many philosophers do not believe that prestigious practitioners of AI take the idea of machine phenomenology and artificial consciousness seriously and hence that, in this aspect at least, the CRA is supposed to target a straw man. Yet many eminent cognitive scientists - Minsky, Moravec and Kurzweil et al - have speculated widely on the subject. Further, as Searle makes clear, it was precisely such statements, emerging from a vociferous bunch of proselytising AI-niks discussing Schank & Abelson's work, that originally led him to formulate the CRA.

But, following Turing, we must rid ourselves of a popular intuition:

Importance is often attached to the fact that modern digital computers are electrical, and that the nervous system also is electrical. Since Babbage's machine was not electrical, and since all digital computers are in a sense equivalent, we see that the use of electricity cannot be of theoretical importance. (Turing 1950, p.439 (p.46 in Boden)).

Indeed, in 1976 Joseph Weizenbaum described a game-playing 'computer' that could be constructed from toilet rolls and coloured stones (Weizenbaum 1976, pp.51ff.). Certainly functionalism, as a philosophy of mind, remains silent on the underlying hardware that causes computational state transitions – whether a program is executed on a PC or a MAC the results of its execution, the computational states it enters, are functionally the same.

References

Bringsjord, S. (1992) What Robots Can and Can't Be, (Dordrecht: Kluwer).

Chalmers, D.J. (1994) 'On Implementing a Computation', Minds and Machines, vol.4, pp.391-402.

(1996a) 'Does a Rock Implement Every Finite-State Automaton?', Synthese, vol.108, pp.309-333.

(1996b) The Conscious Mind: In Search of a Fundamental Theory, (Oxford: Oxford University Press).

Dreyfus, H. (1972) What Computers Cannot Do, (New York: Harper & Row).

Hofstadter, D. (1981) 'Reflections', in The Mind's I: Fantasies and Reflections on Self and Soul, (eds.) D.Hofstadter & D.C.Dennett (London: Penguin), pp.373-382.

Kelly, J. (1993) Artificial Intelligence: A Modern Myth, (Chichester: Ellis Horwood).

Kurzweil, R. (1998) The Age of Spiritual Machines: When Computers Exceed Human Intelligence, (New York: Viking).

Minsky, M. (1985) The Society of Mind, (New York: Simon & Schuster).

Moravec, H.P. (1988) Mind Children: The Future of Robot and Human Intelligence, (Cambridge, MA: Harvard University Press).

Newell, A. & Simon, H.A. (1976) 'Computer Science as Empirical Enquiry: Symbols and Search', Communications of the ACM, vol.19, pp.113-26.

Penrose, R. (1989) The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics, (Oxford: Oxford University Press).

Putnam, H. (1988) Representation and Reality, (Cambridge, MA: MIT Press/Bradford Books).

Schank, R.C. & Abelson, R.P. (1977) Scripts, Plans, Goals & Understanding, (Hillsdale, NJ: Lawrence Erlbaum).

Searle, J.R. (1980) 'Minds, Brains, and Programs', Behavioural and Brain Sciences, vol.3, pp.417-424.

(1982) The Myth of the Computer, New York Review of Books, 29/7: 3-6.

(1984) Minds, Brains and Science, (London: BBC Publications).

(1990) 'Is the Brain a Digital Computer?', Proceedings of the American Philosophical Association, vol.64, pp.21-37.

(1992) The Rediscovery of Mind, (Cambridge, MA: MIT Press).

(1994) The Mystery of Consciousness, (London: Granta Books).
Turing, A.M. (1950) ‘Computing Machinery and Intelligence’, Mind, vol.49, pp.433-460.
Weizenbaum, J. (1976) Computer Power and Human Reason: From Judgement to Calculation, (San Francisco: W.H.Freeman).
Wittgenstein, L. (1953) Philosophical Investigations, (Oxford: Blackwell).