The Chinese Room Argument – AI's Crisis of Science

R. Smith

In 1950, an inspiring academic paper outlined theory and practice for the coming project, soon to be dubbed Artificial Intelligence,[1] to program human-like general intelligence (AGI) into the stored-program electronic digital computer, ending:

> We can only see a short distance ahead, but we can see plenty there that needs to be done.[2]

In 1980, a strong philosophical argument concluded the machine's essential nature prevents it understanding what it receives, for instance from sensors.[3] Hence it could never have or develop human-like intelligence.

Now four decades later the argument still stands: a vast literature but no rebuttal – though many have tried. And despite seven decades of prodigious funding and effort, nothing approaching AGI has been demonstrated. Instead, serious practical and theoretical difficulties have arisen including those known as the problem of design,[4] the problem of machine translation,[5] the frame problem,[6] the problem of common-sense knowledge,[7] the problem of combinatorial explosion,[8] the Chinese room argument,[9] the infinity of facts,[10] the symbol grounding problem,[11] and the problem of encodingism.[12] And as for "deep learning": edge cases,[13] noisy data-sets,[14] adversarial attack[15] and catastrophic forgetting.[16]

---

1 John McCarthy, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," (Dartmouth College, 31 August 1955); "We propose a 2 month 10 man study of artificial intelligence...".

2 A. M. Turing, "Computing Machinery and Intelligence," *Mind* 59, no. 236 (October 1950): 455.

3 John R. Searle, "Minds, Brains, and Programs," *The Behavioral and Brain Sciences* 3, no. 3 (1980): 417-457.

4 Ada Lovelace, 1843, "Note G", quoted in "The Turing Test," *Stanford Encyclopedia of Philosophy*, section 2.6, https://plato.stanford.edu/entries/turing-test. Also see John von Neumann, "First Draft of a Report on the EDVAC," (Moore School of Electrical Engineering, University of Pennsylvania, 30 June 1945), 1.

5 Yehoshua Bar-Hillel, "The Present Status of Automatic Translation of Languages," in *Advances in Computers*, ed. Franz L. Alt (Academic Press, 1960), 1: 91-163.

6 J. McCarthy and P. J. Hayes, "Some Philosophical Problems from the Standpoint to Artificial Intelligence," in Bernard Meltzer and Donald Michie (eds.) *Machine Intelligence 4* (American Elsevier, 1969), 463-502. Also see Dreyfus, "Alchemy and Artificial Intelligence," 29, 39, 68.

7 Hubert L. Dreyfus, "Alchemy and Artificial Intelligence," (The RAND Corporation, P-3244, December 1965), 39.

8 James Lighthill, "Artificial Intelligence: A General Survey," section 3 Conclusion, in *Artificial Intelligence: A Paper Symposium* (Science Research Council of Great Britain, July 1972). Also Dreyfus, "Alchemy and Artificial Intelligence," 39.

9 John R. Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3, no. 3 (1980): 417-457.

10 Dreyfus, "Alchemy and Artificial Intelligence," 39. Also Daniel C. Dennett, "Cognitive Wheels: The Frame Problem of AI," in *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, ed. Zenon W. Pylyshyn (1984; Ablex, 1987), 49.

11 Stevan Harnad, "The Symbol Grounding Problem," *Physica D* 42 (June 1990): 335-346.

12 Mark Bickhard and Lauren Terveen, *Foundational Issues in Artificial Intelligence: Impasse and Solution* (Elsevier, 1995).

13 Overview of issues: Philip Koopman, "Edge Cases and Autonomous Vehicle Safety," (paper presented at the Safety-Critical Systems Symposium, Bristol UK, 7 February 2019).

14 Overview of literature: Gupta, Shivani and Gupta, Atul, "Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review," *Procedia Computer Science* 161 (Elsevier, 2019): 466-474.

15 Survey: Anirban Chakraborty *et al*., "Adversarial Attacks and Defences: A Survey," (arXiv:1810.00069v1, 28 September 2018).

16 James Kirkpatrick *et al*., "Overcoming Catastrophic Forgetting in Neural Networks," *Proceedings of the National Academy of sciences* 114, no. 13, (2017): 3521-3526.

The argument contains a thought experiment that analogically presents AI's fundamental weakness – computation – the manipulation of extrinsically meaningful but intrinsically meaningless tokens.

So is the argument's conclusion true? Or is AI in a crisis of science? Thomas S. Kuhn:

> ...'normal science' presupposes a conceptual and instrumental framework or paradigm accepted by an entire scientific community ... [T]he resulting mode of scientific practice inevitably invokes 'crises' which cannot be resolved within this framework...[17]
>
> ...the analytical thought experimentation that bulks so large in the writings of Galileo, Einstein, Bohr and others is perfectly calculated to expose the old paradigm to existing knowledge in ways that isolate the root of crisis with a clarity unobtainable in the laboratory.[18]

AI's "old paradigm" is that of computation, the manipulation of symbols conditional on the extrinsic meanings of their shapes. The existing knowledge is the purely formal, or syntactic, nature of the symbol, plus the semantic nature of the mind. And the root of crisis is the inability of the symbol to deliver semantic content.

If in a crisis of science, AI's lack of progress towards AGI doesn't mean the computer can't understand what it manipulates. It means AI is using the wrong paradigm. Under the right paradigm, it might understand.

AI has always had its critics.

Yehoshua Bar-Hillel, 1960:

> A human translator ... is often obliged to make intelligent use of extra-linguistic knowledge which sometimes has to be of considerable breadth and depth. Without this knowledge he would often be in no position to resolve semantical ambiguities. At present, no way of constructing machines with such a knowledge is known, nor of writing programs which will ensure intelligent use of this knowledge.[19]

Mortimer Taube, 1961:

> In the final analysis, our electrical engineers and computer enthusiasts should stop talking this way [about programming a computer with the computation of intelligence] and face the serious charge that they are writing science fiction to titillate the public and to make an easy dollar...[20]

Hubert Dreyfus, 1965 (talking about the AI project):

> If the alchemist had stopped pouring over his retorts and pentagrams and had spent his time looking for the true structure of the problem ... things would have been set moving in a more promising direction. After all, three hundred years later we did

17 Thomas S. Kuhn, *The Structure of Scientific Revolutions,* (The University of Chicago Press, 1962), cover.

18 Ibid., 88.

19 Yehoshua Bar-Hillel, "The Present Status of Automatic Translation of Languages," in *Advances in Computers,* ed. F. L. Alt (Academic Press, 1960), 1, no. 1: 91-163.

20 Mortimer Taube, *Computers and Common Sense: The Myth of Thinking Machines* (Columbia University Press, 1961), 52.

> get gold from lead ... but only after we abandoned work on the alchemical level, and reached the chemical level or the even deeper level of the nucleus.[21]

James Lighthill, 1972:

> [Narrow AI has progressed] to a disappointingly smaller extent that had been hoped and expected, while progress in [building human-like robots] has been even slower and more discouraging ... tending to sap confidence in whether the field of research called AI has any true coherence.[22]

Daniel Dennett, 1984:

> ...a more realistic solution to the frame problem (and indeed, in all likelihood, to any solution) [may] require a complete rethinking of the semantic-level setting.[23]

Hubert Dreyfus, 1992:

> ....now that the twentieth century is drawing to a close, it is becoming clear that one of the great dreams of the century is ending too. Almost half a century ago computer pioneer Alan Turing suggested that a high-speed digital computer programmed with rules and facts, might exhibit intelligent behavior. ... however, it is now clear to all but a few diehards that this attempt to produce general intelligence has failed.[24]

Kenneth Sayre, 1993:

> Artificial intelligence pursued within the cult of computationalism stands not even a ghost of a chance of producing durable results ... it is time to divert the efforts of AI researchers – and the considerable monies made available for their support – into avenues other than the computational approach.[25]

Jerry Fodor, 1995 then 2000:

> ...the Artificial Intelligence project has failed.[26] ...there are some very deep problems with viewing cognition as computational.[27]

John McCarthy, 2007:

> ...most AI researchers believe that new fundamental ideas are required, and therefore it cannot be predicted when human-level intelligence will be achieved.[28]

David Gelernter, 2007:

> Unfortunately, AI, cognitive science, and the philosophy of mind are nowhere near knowing how to build [an intelligent machine].[29]

---

21 Dreyfus, "Alchemy and Artificial Intelligence," 86.

22 James Lighthill, "Artificial Intelligence: A General Survey," section 3 Conclusion, in *Artificial Intelligence: A Paper Symposium,* (Science Research Council of Great Britain, July 1972).

23 Daniel C. Dennett, "Cognitive Wheels: The Frame Problem of AI," in *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, ed. Zenon W. Pylyshyn (1984; Ablex, 1988), 61.

24 Hubert L. Dreyfus, *What Computers Still Can't Do* (1992; MIT Press, 1994), ix.

25 Kenneth Sayre, "Three More Flaws in the computational Model," (conference paper, *American Philosophical Association (Central Division) Annual Conference*, Chicago, Illinois, 1993, quoted in Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed. (Prentice Hall, 2008), 947.

26 Jerry Fodor, "The Folly of Simulation," in *Speaking Minds*, eds. P. Baumgartner and S. Payr (Princeton University Press, 1995), 91.

27 Jerry Fodor, *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology* (2000; The MIT Press, paperback edition, 2001), 23.

28 John McCarthy, "What is Artificial Intelligence?" (McCarthy web page, 2007), http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html.

29 David Gelernter, "Artificial Intelligence is Lost in the Woods," *Technology Review*, (The MIT Press, July 2007).

David Deutsch, 2012:

> ...today in 2012, no one is any better at programming an AGI than Turing himself would have been ... The lack of progress in AGI is due to a severe log jam of misconceptions. I cannot think of any other significant field of knowledge where the prevailing wisdom, not only in society at large but among experts, is so beset with entrenched, overlapping, fundamental errors.[30]

Yet under any realized new paradigm, could the machine be genuinely intelligent? Maybe it couldn't. Perhaps there *is* a fundamental reason why the device must fail.

Kuhn pillories both theory *and equipment*. The conceptual *and instrumental* framework of normal science provoke crises. AI's instrumentation is the device it calls the computer. Perhaps AI does need to find new hardware.

But we don't know that. And we can only find out by adopting different paradigms and seeking to realize them – in the current machine. We need first to try to resolve the crisis of science using the existing instrumentation. We have to first assume the equipment is adequate, and that the only problem is the framework used to understand it.

A crisis of science, to those toiling on the bleeding edge, can be tremendously exciting. What if everyone else *is* wrong? They've been wrong before. Maybe everyone else is wrong again.

But where to start seeking a new paradigm? In response to particulate impacts of the proximate environment, sensors emit data streams which contain knowledge in some form. Given the symbolic understanding of the stream is kaput, what is that form? What new concepts and principles are needed to understand how sensory streams contain knowledge? And once extracted, how it can be stored, accessed and updated?

Traditionally, human use of the machine amounts to reacting to shapes displayed on exposed surfaces of peripheral attachments (on plastic key caps, in display screens, sprayed or melted onto sheets of paper). So naturally, unable to escape the conception of their own use, humans think sensory streams contain symbols.

But what do the streams contain *to the machine*? To answer this we use the same formula: what the machine reacts to. What it can react to defines the types of "thing" the stream "contains" – *to the machine*.

For a start, the substance of the impinging environment doesn't pass through the sensor into the inner world. Hence its intrinsic properties don't pass through either. And the process of sensory transduction doesn't copy those properties.

30 David Deutsch, "Philosophy Will Be the Key that Unlocks Artificial Intelligence," *The Guardian* (3 October 2012), https://www.theguardian.com/science/2012/oct/03/philosophy-artificial-intelligence.

So much for *substance* and its *properties*. What about the third item in the ultimate ontology – *relationships*. One internal relation is a duplicate of the causally antecedent external instance. Something does survive sensory transduction. Something on the inside is a duplicate, a copy, of something on the outside. Hence the internal instances are semantic. On this view, a sensory stream comprises a pattern of difference in which newly created inner substance and property are bound together by copies of instances of an *external* relation.

The machine can react to substance *per se*, to values of a property of the substance, or to the relation *per se* between units of the substance. Not one ontological type – but three. The current paradigm has just one type: symbol shape.

Streams by definition move, hence the relation of interest is temporal. In storage, however, the aspect of time is absent, and instances of the temporal relation in the stream take the form of structural connections via pointers in semiconductor storage, thus rendering permanent the temporal structure of the stream.

This rudimentary conceptual framework for understanding the stream from the machine's perspective is consistent with functionalism, just not the semantically vacant computational sort. It indicates it might be possible to adequately understand the workings of the "computer", so called, without using the concepts either of computation or of the symbol, and hence potentially to know how "computers" will acquire semantic knowledge.

The crisis of science exposed by John Searle's 1980 thought experiment of the Chinese room should inspire alternative understandings of the machine, ones not predicated on human use but rather on the abilities and potentialities of the device itself.

With that in mind, interested parties could start developing new conceptual frameworks. One place to begin is the acquisition of knowledge via the principle: all knowledge gained through sense perception is reducible to instances of the relation of temporal contiguity.