MEANING, UNDERSTANDING, SELF-REFERENCE

INFORMATICA 2/91

Keywords: Meaning, Chinese room argument, self-reference

Damjan Bojadžiev
Institute Jožef Stefan, Ljubljana

Arguments against the possibility of machine understanding as symbol manipulation tend to downplay the internal structure of the computational system. The case for genuine mechanical understanding can be made more appealing by considering the levels of (self)-representation and the self-referential mechanisms operating between them.

Pomen, razumevanje, samo-referenca: Argumenti proti možnosti mehaničnega razumevanja kot manipulacije simbolov običajno podcenjujejo notranjo strukturo programskega sistema. Možnost resničnega mehaničnega razumevanja se zdi bolj smiselna, če upoštevamo nivoje (samo)reprezentacije in samo-referenčne mehanizme, ki delujejo med njimi.

1. Introduction

Mainstream tradition in philosophy and more or less educated common sense maintains that there is (and will always be) a huge gap between the cognition, and in particular the linguistic abilities, of men and machines (computers as their most advanced representatives). It is said that, by manipulating symbols, computers might at most succeed in producing an illusion of understanding: appearing to understand what is being said/typed to them and what they themselves say or print, but without actually doing so. Although the current state of the art of artificial intelligence and natural language understanding does not make this a burning issue, many different positions have already been taken on it, presumably because the possibility of machine understanding challenges our current linguistic and cognitive supremacy, and because of the extent to which it affects our self-image as cognitive beings, our understanding of ourselves. The prospect of mechanical understanding apparently tends to have a negative effect on our conception of ourselves, though it is not obvious that it should do so. The indignant position that we are not "mere machines" would only go along with what is sometimes called weak AI, the position that AI programs might only provide superficial I/O simulations and need not be regarded as actual models of human cognitive performance.

The present paper is an overview of some typical arguments, notably yet another debunking of Searle's "Chinese room" argument (section 2) and a compilation of some programatic answers, centering on the notion of self-reference (section 3). The main thesis elaborated there is that the connection between meaning, understanding and self-reference goes through the subject of that understanding, which is construed as an effect of self-referential mechanisms in a meta-level architecture. If the objections against computational understanding are summarized by saying, as Searle does, that computers have syntax but not semantics, and that syntax alone is not sufficient for semantics [19:34], the answer can be summarized, following Casti [4:334-5], by saying that syntax and self-reference may add up to semantics.

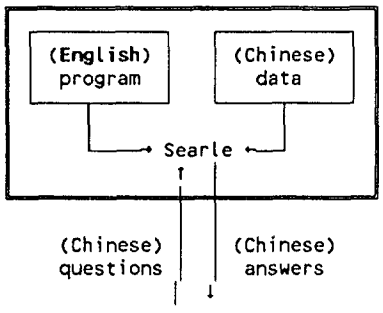
2. Real Understanding

The standard objection against computational understanding by means of symbol manipulation in a formal representational system says:

computers themselves don't mean anything by their tokens (any more than books do) - they only mean what we say they do [7:32-3]; symbols must mean something to the system performing the operations [6].

The second quotation comes from the philosopher F. Dretske, who helps us keep a cool head in such discussions by reminding us that computers are tools, and that we should not get carried away in attributing conceptual or cognitive capacities to them. According to Dretske, the meaning of internal signs consists in their correlation with external conditions, and affects (through learning) the way the system manages the signs. (His negative conclusions about computational cognition seem to rest on the premise that robots can't learn)

A much debated argument about machine understanding is the thought experiment suggested a decade ago by the philosopher J. Searle, known as the Chinese Room argument [19]. Searle imagines that he could pass a (Turing) test for understanding eg. Chinese in the following way, although he does not actually know the language:



Isolated in a room, Searle would carry out instructions (in English) for manipulating incoming strings of Chinese symbols using given strings of Chinese data, and producing answer strings such as would be given by a speaker of Chinese. It is important in this setup that the instructions only say how to manipulate formally characterized Chinese symbols, and thus do not interpret them (in English). Searle then claims that, just as he would not have understood a word of the Chinese, although he would appear to do so from the point of view of the external questioner, a computer executing the same instructions would also not understand anything[1]:

In the Chinese case, the computer is me, and in cases where the computer is not me, the computer has nothing more than I have in the case where I understand nothing [Minds, Brains and Programs].

It seems obvious that Searle would indeed not come to understand Chinese in this way, although even that might be allowed as an abstract possibility[2]. More importantly, a computer taking Searle's place would indeed also not understand anything, but only on Searle's peculiar usage of the word 'computer'. Searle seems to think of a computer as an "empty" processor, without programs and data: he almost seems to identify with the CPU; more accurately, he identifies with the interpreter (in the computational sense of the word) of the formal instructions for manipulating Chinese symbols. No one has ever claimed understanding (of Chinese or anything else) for a computer in this sense; it is the programmed computer (plus data) that might be said to understand - this is what Searle calls the systems reply to his argument[3]. Torrance aptly summarizes it by relating it to an everyday situation: it is not the washing machine's motor which washes clothes, but the washing machine [22:15].

The part I find interesting about the Chinese room thought experiment[4] is Searle's attempt to identify with an understanding computer. According to strong AI, which he is trying to refute, computational understanding models human understanding, so that some identification should be possible; Searle only makes a type error in the particular identification he imagines. The identification does not go through because we do not recognize a (formal) description of what it is for symbols to be meaningful to us in the work of the fast, dumb homunculus whose place Searle thinks he could take. The systems reply to the "experiment" suggests another identification, with the whole Chinese room (or, in a parallel processing version with some Chinese hotel). A variant on this would be identification with (the perspective of) the simulated speaker [10:378]. These possibilities correspond to the two senses in which we speak of symbols being meaningful to us: meaningful to our whole cognitive system or meaningful to some particular point within the system: consciousness, self, mind's eye or I, whatever; such labels only presume what they appear to explain. A short argument for the connection between understanding and "selfhood" can be found in Haugeland [8:240]: ego involvement may be integral to understanding, and ego involvement requires a self-concept. A more technical one can be built around the notions of levels of representation in a symbolic system and the relationships between such levels. An early, idealized attempt to discern some analogue of subjectivity within a formal system was made by Hofstadter, in his 'Gödel, Escher, Bach' [9]. He proposed an arithmetical, Pythagorean version of the interplay between levels as engendering the self: it is formed when it can reflect itself [9:709]. The

(meta)arithmetical basis of the slogan were the effects of sentential self-reference for particular predicates: the "strange loop" connecting the (un)-provability of a sentence and the sentence's assertion of its own (un)provability. In hard(er) computer science, ideas about the effects of incorporating the meta-level in a formal representational system were further developed under the label meta-level architectures[5].

3. Self-Referential Systems

If the symbols of a system are to be meaningful to the system itself, it must first be able to consider them and the way it manipulates them. Thus, Perlis [15] has suggested that a system would use symbols meaningfully if it could consider its own symbolic forms and distinguish between its symbols and what they symbolize. Both conditions require second-order representation (naming or otherwise representing the symbols themselves) and could thus be met in a system with quotation[6]. The suggestion may appear singularly inappropriate, since the alleged problem with computational understanding was that a computer has no access to what its symbols symbolize (for us), that it cannot get "down" from symbols to symbolized (things, meanings): quotation can then only get it one level further up, and disquotation (through a truth predicate) can only get it down to where it was in the first place[7]. Quotation would thus appear to offer no help with the original problem, but then it turns out that the suggestion also includes the position that there is actually no such problem: Perlis says that

> we are reduced to dealing with symbols and their meanings, whatever they are, via expressions or other internal forms ... we never get to the outer "thing in itself" ... expressions or other internal forms do all the work, but at least one is momentarily taken as the thing-in-itself [How can a Program Mean].

The first part of this is the point frequently emphasized by Wilks: there is no escape from the world of symbols to the real world, since the world only comes to a symbolic system through further (sub)symbols. The second, metaphysical part of Perlis' position (reality is projected representation) might at first appear more startling, though it is presumably only a consequence of the first; Jackendoff claims something similar when he talks of a mentally supplied attribute of "out-thereness" and of our being constructed so as normally to be unaware of our own contribution to experience.

Perlis's suggestion could be characterized in currently influential terms by saying that the system must have a meta-level architecture, capable of self-reflection, or self-reference: being able to move to the meta-level and take a stance towards its own symbolic structures (notably judge them true or false), and then descend back to the object level. But in such a system it would seem possible to discern some features of subjectivity in the possibilities of self-representation and mediation between levels of representation. A connection between subjectivity and meaningfulness to the system would seem to make methodological sense, since meaning and the instance to which it is present can only be correlative phenomena, to be explained simultaneously in a computational or any other model. The basic idea is that the self (subjectivity) can be computationally characterized by rules relating (at least) two levels of representation: a ground, primitive level at which we have no representation of ourselves, and a higher, general level, at which we have a neutral representation of ourselves as just another object. Thus, Perry [16] sees indexicals (token-reflexive expressions such as personal pronouns) as mediating between such levels:

> At the "bottom" level, we have cognitions that have no representation of ourselves (or the present moment), which are tied pretty directly to cognition and action ... Since [simple organisms] are always in the background of their perceptions and actions, they need not be represented in the cognitions that intervene between them [Self-Knowledge and Self-Representation].

According to Smith [20], self-referential mechanisms suggest a computational idea of the self as mediating between "blindly" efficient, indexical representations and generic, (more) explicit representation of its circumstances. Smith considers three self-referential mechanisms which vary the theme of self-recognition: paraphrasing roughly, these are: autonimy (recognizing one's own name), introspection (recognizing one's own (implicit) internal structure), and reflection (recognizing oneself in the world)[8]. These mechanisms correspond to different conceptions of self: as a unit, complex system, and independant agent, respectively[9]. Most of the literature on meta-level architectures is concerned with introspection, since many problems of independant interest to the AI effort naturally fall in this framework: learning, planning, default reasoning, truth maintainance, reasoning about control, reasoning about beliefs, incomplete and inconsistent knowledge, handling impasses in reasoning [2], [13]. The classical problems with self-referential paradoxes in mathematical logic and elsewhere would come under the heading of narrow, sentential self-reference (of a sentence to itself), as opposed to introspection in

general as reference of a system to aspects of itself.

In the area of natural language understanding, self-referential structures will presumably appear with the inclusion of meta-level representations of knowledge, linguistic and otherwise. A major effort at formalising the meaning of natural language expressions, Montague semantics, can also be seen to fit in the framework of meta-level architectures (incorporating the meta-level of intensions into the object language) [21]. In general, systems for computational understanding of language will have to include part of their own meta-level because of the range of self-referential phenomena in communication. Thus, in addition to information about (external) situations to which they refer, statements also convey information about the speaker: his other beliefs (about himself and others), his distance (spatial, temporal, ...) from what he is talking about [1:30]; linguistic actions - asking (for help or information), lying, threatening, promising - generally have a self-regarding nature [10:267]. It would thus seem, as Perlis [14] has contended, that a proper understanding of language will in the long run be found to depend upon self-referential abilities.

N o t e s

1. Searle still maintains these views, as can be seen from his recent exposition in Scientific American, titled 'Is the Brain's Mind a Computer Program?' [19]. The subtitle gives a curious answer: 'No. A program merely manipulates symbols, whereas a brain attaches meaning to them'. The notion of a brain attaching meaning to symbols seems rather strange; a mind (subject) would seem better suited for that role.

2. Searle doesn't mention this possibility, though he discusses a similar one, on which he might assign some other, arbitrary interpretation to the symbols he is manipulating (chess moves, stock market predictions, ...) [19]. The structure of the formal system may allow such interpretations, but then similar reinterpretations could also be applied to what Searle (or anyone else) ordinarily says (in English, outside the Chinese room). Usually, we think of such reinterpretation, on a smaller scale, only when what is being said fits very badly with what it is said about. But the abstract possibility of such reinterpretation of the whole language in general remains, and has been explored in arguments directed against the importance of the relation of reference in semantics (permutations of reference; see eg. Davidson's paper 'The Inscrutabilty of Reference' in [5].

3. Searle's reply to the systems reply is that, if understanding is not ascribed to him but to the system <program,data,Searle>, in which he is the agent, he will simply internalize (memorize) the other components:

> The individual then incorporates the entire system. There isn't anything at all to the system that he does not encompass ... All the same, he understands nothing of the Chinese, and a fortiori neither does the system, because there isn't anything in the system that isn't in him [Minds, Brains, Programs].

Searle rightly says that he feels somewhat embarassed to give this 'quite simple' reply: paraphrasing to bring out its curiously childish nature, Searle seems to be saying: "OK, if understanding is not in me but in (my relationship to) these other things outside me, then I'll just put them inside me". Searle seems to be confusing physical with functional containment: he would still be only the agent in this internalized (sub)system, through which he would now be speaking in tongues. The argument

> The whole doesn't understand
> ---
> So, the parts don't.

only seems to bug the question: that is exactly a situation in which the whole doesn't understand while parts do.

4. Searle's "experiment" does dramatize some issues concerning the formalization of meaning and understanding: where or how is meaning present in a formal system (program), and who or what is that meaning present to. Common sense answers invoke the programmer: it is his knowledge of meaning that goes into setting up a formal system, and the meaning captured in it is present to him (as attributed meaning). Other attempts rely on some kind of (w)holism: the meaningfullness of individual symbols is an effect of the whole system, deriving from the relationships of individual symbols with indefinitely many other symbols and with the procedures which manipulate them. Symbols would then be meaningful to the system if it considered not only their immediate, explicit, locally computable ... properties, but also their relationships with indefinitely many other symbols; cf. Hofstadter in [9:582].

5. A survey of three early attempts at computational introspection (Smith's introspective LISP, Weyhrauch's introspective first order logic, FOL, and Doyle's model of introspective action and deliberation) can be found in [2]. More recent efforts are collected in the proceedings of the 1986 workshop on meta-level architectures and reflection, sponsored by the COST-13 project Advanced Issues in Knowledge Representation [13].

6. Perlis uses a more general idea of quotatation, according to which not only internal tokens, but any internal (mental) object or process can be "quoted" (or reflected, as he also says, meaning something similar to what Hofstadter calls jumping out).

7. The ideas of language levels (use and mention (quotation)) and disquotation (truth) are basic to Davidson's semantics for natural language [5]; I hadn't seen Perlis's point in my estimate of the usefulness of Davidson's semantics for computational understanding of language [3].

8. The basic, low-level view of the world, at which we have no representation of ourselves, could be compared to the perspective of the "man with no head" in [10:23-33]. The higher level consists of additional representational structure, which is basic to what Smith has to say about the self-referential mechanisms of the self. In the arithmetical case, considered by Hofstadter, there is no additional representational structure on the higher level, and it is the Gödel code which provides "additional", meta-arithmetical interpretations.

9. The importance of introspective mechanisms in evolutionary biological engineering was pointed out by Lycan:

Parallel processing, time-sharing and hierarchical control, all vital to the fabulous efficiency of such complex sensor-cognition-motor systems as we human beings are, individually and together require a formidable capacity for internal monitoring ... As a matter of engineering, if we did not have the devices of introspection, there would be no we to argue about, or to do the arguing [Consciousness:72-3].

References

1. Barwise, J., Perry, J., Situations and Attitudes, MIT Press 1983

2. Batali, J., Computational Introspection, MIT AI Memo 701, 1983

3. Bojadžiev, D., Davidson's Semantics and Computational Understanding of Language, Grazer Philosophische Studien, Vol.36, 1989

4. Casti, J.L., Paradigms Lost: Images of Man in the Mirror of Science, W.Morrow & Co. 1989

5. Davidson, D., Inquiries into Truth and Interpretation, Oxford Univ. Press 1984

6. Dretske, F., Machines and the Mental, Am. Phil. Assoc. Presidential Address, 1985

7. Haugeland, J., Mind Design, Bradford Books 1981

8. ----------, Artificial Intelligence: The very Idea, Bradford Books 1985

9. Hofstadter, D.R., Gödel, Escher, Bach: An Eternal Golden Braid, Basic Books 1979

10. ----------, Dennett, D.C., The Mind's I - Fantasies and Reflections on Self and Soul (1981), Penguin 1982

11. Jackendoff, R., Semantics and Cognition, MIT Press 1983

12. Lycan, W.G., Consciousness, Bradford, MIT 1987

13. Maes, P., Nardi, D. (eds.), Meta-Level Architectures and Reflection, Elsevier 1988

14. Perlis, D., Languages with Self-Reference I: Foundations, Artificial Intelligence 25, 301-22, 1985

15. ----------, How Can a Program Mean?, in Proceedings of the 10.th IJCAI, Vol.1, Milano 1987

16. Perry, J., Self-knowledge and Self-representation, in Proceedings of the 9.th IJCAI, Vol.2, Los Angeles 1985

17. Searle, J., Minds, Brains and Programs, Behavioral & Brain Sc. 3, 1980; reprinted eg. in [7]

18. ----------, Minds, Brains and Science, The 1984 Reith Lectures, British Broadcasting Corporation 1984

19. ----------, Is the Brain's Mind a Computer Program?, Sc. Am. Vol.262 No.1, Jan. 1990

20. Smith, B.C., Varieties of Self-reference, in J.Y.Halpern (ed.), Theoretical Aspects of Reasoning about Knowledge, Proceedings of the 1986 Conf., Morgan Kaufmann 1986

21. Steels, L., Meaning in Knowledge Representation, in [13]

22. Torrance, S. (ed.), The Mind and the Machine - Philosophical Aspects of Artificial Intelligence, Ellis Horwood 1984