Wittgenstein on Language and Artificial Intelligence: The Chinese-Room Thought Experiment Revisited

Author(s): Klaus K. Obermeier

Source: *Synthese*, Sep., 1983, Vol. 56, No. 3, Ludwig Wittgenstein: Proceedings of a Conference Sponsored by the Austrian Institute, New York, Part II (Sep., 1983), pp. 339-349

Published by: Springer

Stable URL: https://www.jstor.org/stable/20115911

KLAUS K. OBERMEIER

WITTGENSTEIN ON LANGUAGE AND ARTIFICIAL INTELLIGENCE: THE CHINESE-ROOM THOUGHT EXPERIMENT REVISITED

> If a lion could speak we would not be able to understand him.

Artificial Intelligence, the study and emulation of intelligence, has brought forth a host of epistemological and ethical issues (Anderson 1964). In the wake of viable natural-language-understanding programs, the question, "What does it take to put together an understanding program?" has superseded the hypothetical query, "What if machines can think?" Philosophers who have been concerned with these issues have adopted two different points of view: (1) programs will never be able to emulate human intelligence, mainly because they lack consciousness and intentionality (Scriven 1953; Dreyfus 1979; Searle 1980); (2) programs already manifest intelligent behavior in limited domains and are potentially able to deal with larger domains in due course (Hofstadter 1980, 680; Schank 1975, 4; Winograd 1972, 2). Although both sides agree that the computer is a useful tool for simulation of behavior, the argument ensues over the issue if programs can go beyond mere imitation of human behavior, if they can be creative, if they can think. This paper investigates these questions from a Wittgensteinian point of view.

Wittgenstein's philosophy does not seem to be in the least amenable to "thinking" or "understanding" programs. It is centered around the human being, thus decidedly anthropocentric; along these lines, Wittgenstein's denial of credibility of "thinking" programs is categorical: "But machines can't think" (*PI*, 360). Therefore, it is surprising that he is considered to be the forerunner of "the concerns of modern AI" (Wilks 1976, 222). Even more problematic is the use of quotes from Wittgenstein to support advocates of "strong" AI (i.e., people who believe that a program *is* a mind) (Searle 1980), as well as contenders

Synthese **56** (1983) 339–349. 0039–7857/83/0563–0339 $01.10

of "weak" AI (i.e., people who consider the program as a useful tool only).[1]

In light of these controversial issues that are based on the function and notion of language, it will become obvious that a Wittgensteinian account of "understanding" does indeed support claims made by advocates of "strong AI" that pertain to the role and function of language in the understanding process. At the same time it will become evident that a linguistic theory in general, and a semantic theory in particular, does not necessarily include the notion of "intentionality" (Ziff 1960) or, more poignantly, does not have to deal with this notion at all (Putnam 1960).[2]

1.

Imagine a person sitting in a room and answering questions in Chinese. The person does not know any Chinese and has to rely solely on manuals that contain the appropriate instructions for him to be able to pass as a native speaker of Chinese. Imagine a machine answering the questions in Chinese and passing the Turing test (Turing 1950).

This scenario, contrived by John R. Searle (1980) and widely discussed by researchers in various fields of science,[3] is designed to lend credence to Searle's claim that "instantiating a program could not be constitutive of intentionality, because it would be possible for an agent to instantiate the program and not have the right kind of intentionality" (Searle 1980). The instantiation of a program is not a sufficient condition of understanding, either, since "the formal symbol manipulations by themselves don't have any intentionality; they are quite meaningless; they aren't even *symbol* manipulations, since the symbols don't symbolize anything" (Searle 1980). In brief, intentionality is a *conditio sine qua* for understanding.

Intentionality is for Searle a biological phenomenon; for Wittgenstein it is a linguistic one: "We use the words 'meaning', 'believing', 'intending' in such a way that they refer to certain acts, states of mind given certain circumstances" (*BB*, 147).[4] The circumstances help to determine the intention of an utterance; moreover, "there are a great many combinations of actions and states of minds which we should call 'intending'" (*BB*, 32). Therefore, the projection of a meaning onto an utterance is not the intention itself. Rather, it is an interpretation of the

utterance at hand: "For we could always have intended the opposite by reinterpreting the process of projection" (*BB*, 33). Suppose, as Wittgenstein does (*BB*, 147), I had made a move in chess. Someone asks: "Did you intend to mate him?" How do I know the answer since all I knew was what happened within me when I made the move (*BB*, 147). I could have intended to move my king from A4 to A5 or in any other odd way, or to aggravate my opponent, or to speed up the game. The meaning of the utterance (e.g., "checkmate") can be interpreted in different ways, too. The intention, however, is commonly inferred from the circumstances. Since I did not move in any odd way, or appear to be hostile towards my opponent, or choose to knock over the board, I *intended* to mate my opponent. From this example we learn that intentionality is a matter of inference from linguistic and nonlinguistic facts, not as Searle claims a biological process. In matters of language understanding, the question about the intention of an utterance gives way to the question of the function of a particular utterance. An utterance all by itself does not have "intentionality"; it is the immediate context in which it is used that establishes the "aboutness" or purpose of the utterance. Thus, "intentionality" is a possible, not a necessary by-product of a discourse. What is important for "understanding" is the function of an utterance in a given situation. Words are used for their meaning, and the meaning depends on the use in a given context. The meaning of the same word is subject to change according to the context, which is based on the function of the single words that make up the utterance. While we use words with a certain meaning qua use in mind, they acquire new meanings that depend on the specific use we put them to in "unheard-of ways" (*PI*, 133). However, "it is not our aim to refine or complete the system of rules for the use of our words in unheard-of ways" (*PI*, 133).

Intentionality is clearly not a concept that is necessary to deal with understanding; it is certainly not useful for a concise investigation of language due to its vagueness and biological implications. If we accepted a strictly biological foundation of "understanding", the usefulness of any scientific investigation would be reduced to a description of the mechanical aspect of "understanding" vis-à-vis an undescribable mental substrate. The reason for this chasm between mental and mechanical phenomena can be traced back to the mind–body problem. Searle seems to be hinting at just this dichotomy, hoping to show that mental states can be identified with physical states in humans, but not in machines. If this is

the case, there are two very good arguments against it: (1) the states of machines are comparable in the abstract (of course, neurons are different from electrochips) to the states of a human being: the logical and structural states correspond to the mental and physical states. Consequently, the solution to the mind–body problem would have an epistemological impact on the issue at hand (Putnam 1960) and would show at any rate a more humanoid disposition of the computer than Searle envisioned. (2) Even if the mind–body problem is solved, "it would not shed the slightest light on the world in which we live" (Putnam 1960, 96). Putnam is right to qualify the "identity" or "nonidentity" of logical and structural states (mental versus physical) as unimportant and "purely verbal" (*ibid.*). Searle enshrouded the central notion in his Chinese-Room Thought Experiment by making reference to the "specific biochemistry of its origin" (Searle 1980, 372), thus rendering "intentionality" an all-encompassing as well as a useless notion for the process of understanding.

2.

The process of understanding is primarily linguistic in nature. As Wittgenstein points out: "To understand a sentence means to understand the language. To understand the language means to be master of a certain technique" (*PI*, 199). This process of understanding depends on a viable representation and on the active participation of the listener: "A perspicuous representation produces just that understanding which consists in "seeing connections" (*PI*, 122). The process of understanding is contingent on the actual performance of the speech act and not on emotional by-products that might accompany it. The crucial point in Wittgenstein's remarks on understanding is his separation of understanding as performance from understanding as feeling (Wilks 1976, 231). Searle, as others before him (Dreyfus 1979), confuses at some point what we do versus what we feel when we do certain things (e.g., understand something). The understanding of an utterance depends on the meaning qua use of the words, and the context which is in part established by the specific use of the words:

> When someone says the word "cube" to me, for example, I know what it means. But can the whole *use* of the word come before my mind, when I *understand* it in this way?
>
> Well, but on the other hand isn't the meaning of the word also determined by its use? And

can't these ways of determining meaning conflict? Can what we grasp *in a flash* accord with a use, fit or fail to fit it? And how can what is present to us in an instant, what comes before our mind in an instant, fit a *use*?

What really comes before our mind when we *understand* a word? – Isn't it something like a picture? Can't it *be* a picture?

Well, suppose that a picture does come before your mind when you hear the word "cube", say the drawing of a cube. In what sense can this picture fit or fail to fit a use of the word "cube"? – Perhaps you say: "It's quite simple; – if that picture occurs to me and I point to a triangular prism for instance, and say it is a cube, then this use of the word doesn't fit the picture." – But doesn't it fit? I have purposely so chosen the example that it is quite easy to imagine a *method of projection* according to which the picture does fit after all.

The picture of the cube did indeed *suggest* a certain use to us, but it was possible for me to use it differently. (*PI*, 139)

Imagine that someone is sitting in a room and hears the word "cube". The word all by itself has four different "lexicon"-entries, each one of these has different meanings according to the context in which it is used. If we focus on Wittgenstein's original German example (*Würfel*) which does not have numerous lexicon entries, the person, upon hearing the word, might associate with it a linguistic definition (e.g., "a regular solid of six equal square sides"), or a pictorial representation, or even an anecdotal reference (e.g., "the sculpture in Greenwich Village in the middle of Astor Place"). Understanding a word in isolation is fortuitous since the representation and retrieval of "meaning" by the listener is not determined by either linguistic or nonlinguistic context.

The meaning of "cube" in a linguistic context differs considerably, thus giving rise to a "meaning conflict". How does the person in our imagined room interpret the following sentences?

(1) The cube fell in my lap.
(2) The cube destroyed Oscar's confidence.
(3) The cube turns if two people push it.
(4) The cube is enough for one cup of soup.

There is a common denominator for the meaning of "cube" in these examples which makes up our concept of a cube. This concept is instantiated upon hearing the word, then modified according to the context in which it is used. The overriding principle is the question of plausibility; if it is possible to use the picture of a cube that came to my mind differently, it has to have a meaning inherent in language that is congruent with its specific use in the given context. If we were to design a model of language processing that describes the physical world, at least

four processes should be involved (Waltz 1981): judging the plausibility; representing meaning; retrieving information from memory; taking appropriate linguistic or nonlinguistic action. Whenever the listener encounters a word, its meaning gets modified by the various uses it is subjected to.

The person in the room after hearing the word "cube" for the first time has managed to find a suitable semantic representation for it. When he hears sentence (1) he retrieves the meaning from memory and, assuming he is familiar with the other words in the sentence, tries to fit it into the context. Since "lap" refers to something tangible and concrete, these properties might be transferred to the new semantic representations of "cube"; moreover the size of a lap limits the representation of "cube" for the present context. This representation can depend on linguistic, logical or direct geometric modeling, using 2-D or 3-D systems (Waltz 1981, 2).[5] Besides the size, other qualities of the cube are important (e.g., color, weight) as well as how the cube fell into my lap. It seems to be clear from this example that visual representations of propositions seem to play a distinctive role in language understanding. Thus, the picture of a cube does indeed suggest a certain use which changes according to the context in which it is used.

The representation of "cube" is again modified after the person hears sentence (2): confidence, an abstract and nontangible concept is acted upon by a concrete object; a meaning conflict between sentences (1) and (2) is found. For a natural-language understander the concept of a cube entails further properties (e.g., ability to challenge, puzzle, bewilder someone). For the person in the room, however, whose domain of discourse is so far limited to "cube" and cubes that fell into my lap, the meaning of "cube" in (2) has to appear utterly opaque. Neither visual nor logical operations can prevent this. If knowledge of the outside world could help to determine the meaning of "cube" in (2) it would be added to the representation of the meaning of the language. This addition holds only for the situation at hand; Oscar might be exceptional when it comes to confidence-destroying cubes in that he had been confident that there were no things such as cubes, or that he was confident that he could solve the puzzle of Rubik's cube for good. Therefore, sentence (2) has added only a very limited property of cubes to its semantic representation.

The varying qualities of cubes become evident in (3) and (4). When it

comes to the size and weight of cubes, both qualities can differ greatly, not to mention the "purpose" of cubes. In (3) our subject in the room not only has to *imagine* the size of the cube but also in what position and in what direction it is pushed. There are three events or states that are important for the understanding of (3):

(A) the cube is static at t_1,
(B) the act of pushing at t_2,
(C) the turning of the cube at $t_{2/3}$.

To understand sentence (3), the relation of the two actions (i.e., pushing and turning) has to be represented; the representation of the cube itself in this particular sentence changes according to the linguistic context. The prototypical representation of a cube, which AI researchers would refer to as "default value" (Schank 1973), is altered depending on the specific scenario ("script") in which it is used. The semantic representation thus encompasses the meanings of an entity that depend on their uses in a particular context. The "disunity" of language arises from the "meaning conflicts" which in turn result from idiosyncratic uses of the words. By the same token the representation of "cube" in (3) and (4) consists of a common denominator with respect to shape and particular features in terms of weight, size, purpose, etc., that are instantiated in the particular sentences.

To be master of a technique, to understand a language, the person must have stored the meanings qua use of the words, and, more importantly, must be able to adapt and assimilate new meanings in unheard-of ways to the current use in a given context. This is a linguistic process than can be succinctly analyzed in terms of plausibility, representability and computability of a sentence. These properties in turn enable us to understanding that is guided by inferences and expectations. The storage and retrieval of information enable the human understander to function successfully in terms of language. The emulation of this ability by a computer program is contingent on the use it makes of information that has been presented to it. If a program can handle linguistic information, relying on the function of the individual words in context and acting according to established rules, it is capable of "understanding-as-performance". Imagine that some computer is sitting in a room and inputs the word "cube".

3.

The process of understanding, according to Searle, is inextricably connected with appropriate mental states. How can the program understand what pain is, i.e., emulate feelings, if it is biologically different from humans? A more basic question to be asked is, How does a human being understand the sentence "I am in pain"? Besides the possibility of uttering the sentence "I am in pain" after mental reflection, the human could have said it without prior act of judgment (Wittgenstein refers to this speech act as 'evincing'). As Putnam points out (1960, 81): "The difficulty is occasioned by the fact that the 'verbal report' . . . issues directly from the state it 'reports': no 'computation' or additional 'evidence' is needed to arrive at the 'answer'". Since a human being can make his mental states known without prior reflection, a computer program printing out the message, "I am in loop two", should not be required to provide evidence for this statement. By the same token a person upon being asked, "How do you know you are in pain?" would give the legitimate answer, "Because I am in pain". The crucial difference in stating versus justifying a certain state, be it mental or physical, consists in the evidence one has to provide while justifying a proposition. If a program ascertains being in a certain state, it is not required to provide evidence for the statement it is making.

Wittgenstein takes this query one step further: is it possible to tell from a person's behavior the kind of mental state he is in? His contention is that there is a logical/conceptual link between the linguistic and nonlinguistic behavior and an ascription of a certain mental state to a person. This relation is not inverse, since certain mental states have varying reasons. For researchers in AI the emulation of intelligence does not require equivalent mental states for computers; there is, however, the distinction between logical and structural states that corresond to the human mental and physical states, respectively. If a person says, "I am in state X", he makes reference to his physical state; if he says, "Y is in state X", he makes reference to his mental state that led him to believe that Y is in a certain state. Whereas the question, "How do you know you are in state X?" is redundant (for its tautological answer), the question, "How do you know Y is in state X?" requires a succinct line of reasoning to provide evidence for the claim. By the same token, programs can only be required to provide evidence for claims about other programs' states, not for claims about their own states (Putnam 1960, 81). From these remarks

it should be obvious that the notion of understanding consists of different criteria for an entity that understands itself, versus an entity that understands another entity.

The process of understanding is determined by the experience a person is subjected to. If a person never feels pain, the concept of pain might still be understood by him on a verbal level: under these conditions I know that "*X* is in pain". By the same token, a computer program can print out "*X* is in pain", and we can assume it understands this utterance on a linguistic and definitional level, not in an evincive sense. Therefore, understanding a state of being does not require a previous exposure to this state; otherwise, a person assuring a migraine sufferer by uttering the sentence, "I know what you feel", is either not telling the truth or speaking from personal experience. It would be equally nonsensical to say that a computer program does not understand pain because it could never be in pain. By the same token we understand perfectly well a person who says, "I feel like a machine", although we know he never has had first-hand experience what it is like to be a machine.

Words are tools of communication whose function/use in a given context determines their meanings, both literally and figuratively. Only if the domain of discourse is based on a commonly agreed upon set of rules can we arrive at understanding. In this sense the initial quote about the lion has to be interpreted as follows: although the lion would use the same mode of communication, his frame of reference, our understanding-as-performance, would not be congruent with the meanings we attach to his verbalizations. Since his frame of reference is not based on ours, we would not be able to understand him in terms of getting at the meaning of his utterances. His enunciations could be likened to those of a language learner who tries to transfer the meanings from his native tongue to his new medium. Although he speaks in well-formed sentences, he stays in his own system while his listeners stay in theirs. In the case of a computer program, the understanding is based on the frame of reference which the speech community has with which the program interacts. The act of understanding by a program is therefore not based on alien concepts, but on concepts common to the human user of the program.

"Understanding" and "thinking" programs have posed a number of philosophical questions, both epistemological and ethical in nature, to researchers of various fields of science. The question, "Can machines think?" might have been poignantly answered by Tarski, who after Paul Ziff had asked him replied, "Of course they can, it only depends on what you mean by 'think'".

NOTES

[1] An explanation of this problem might be that Wittgenstein's remarks, far from being aphoristic, are used and quoted in isolation and not considered to be only aspects of a coherent philosophical conviction. Dreyfus (1979, 57) claims as his thesis that what we are we never can explicitly know, "owes a lot to Wittgenstein"; whereas in *PI*, 485, Wittgenstein states, "Justification by experience comes to an end; otherwise it would not be a justification". Dreyfus also claims that his contention ("intelligence implies being human") is based on Wittgenstein; this does not take into account Wittgenstein's distinction between understanding-as-performance and understanding-as-feeling. As I will argue in this paper programs can have the former and therefore can be called "intelligent"; see (Wilks 1976, 231) for a similar claim.

[2] This does not lead Putnam or Ziff to the conclusion that "machines can think"; it only implies that a "behavioristic" semantics, as Ziff (1960) proposes, is possible, much to the chagrin of Chisholm who denies this possibility.

[3] Some twenty-eight responses to Searle's article were published. They were written by philosophers, psychologists, neurologists and AI researchers (Searle 1980).

[4] References are based on the Harper Colophon edition (1965).

[5] These four types of representation are in a hierarchical relationship, the 3-D representation being the "deepest". Linguistic and logical modeling have bounded, linear forms of representation, whereas geometrical modeling requires new representation schemes (Waltz 1981, 2).

REFERENCES

Anderson, A. R.: 1964, *Minds and Machines*, Contemporary Perspectives in Philosophy Series, Prentice-Hall, Englewood Cliffs, N.J.

Dreyfus, H.: 1979, *What Computers Can't Do* (Revised Edition), *The Limits of Artificial Intelligence*, Harper and Row, New York.

Hofstadter, D. R.: 1980, *Gödel, Escher, Bach: An Eternal Braid*, Vintage Books, New York.

Putnam, H.: 1960, 'Minds and machines', in Anderson (1964), 72–97.

Schank, R.: 1973, 'Identification of conceptualization underlying natural language', in R. Schank and K. Colby (eds.), *Computer Models of Thought and Language*, Freeman, San Francisco.

Schank, R.: 1975, *Conceptual Information Processing*, North-Holland, Amsterdam.

Scriven, M.: 1953, 'The mechanical concept of mind', in Anderson (1964), 31–42.

Searle, J. R.: 1980, 'Minds, brains and programs', *The Behavioral and Brain Sciences* Vol. 3, Cambridge University Press, Cambridge. Also in D. R. Hofstadter and D. C. Dennett (eds.), *The Mind's I*, Basic Books, New York, 353–73.

Turing, A. M.: 1950, 'Computing machinery and intelligence', *Mind* Vol. LIX, No. 236.

Waltz, D. L.: 1981, 'Toward a detailed model of processing for language describing the physical world', *Proc. 7th. IJCAL* Vol. 1, 1–6.

Wilks, Y.: 1976, 'Philosophy of language', in E. Charniak and Y. Wilks (eds.), *Computational Semantics: An Introduction to Artificial Intelligence and Natural Language Comprehension*, North-Holland, Amsterdam.
Winograd, T.: 1972, *Understanding Natural Language*, Academic Press, New York.
Ziff, P.: 1960, *Semantic Analysis*, Cornell University Press, Ithaca, N.Y.

Dept. of Linguistics
The Ohio State University
Columbus, OH 43210
U.S.A.