Searle's Experiments with Thought
Author(s): William J. Rapaport
Source: *Philosophy of Science*, Vol. 53, No. 2 (Jun., 1986), pp. 271-279
Published by: The University of Chicago Press on behalf of the Philosophy of Science Association
Stable URL: http://www.jstor.org/stable/187697
Accessed: 17/05/2013 16:28

DISCUSSION:
SEARLE'S EXPERIMENTS WITH THOUGHT*

WILLIAM J. RAPAPORT

Department of Computer Science
State University of New York at Buffalo

1. Introduction. John Searle's provocative essay, "Minds, Brains, and Programs" (1980), with its Chinese-room thought experiment (see the Appendix), has generated a great deal of interest and controversy among philosophers, computer scientists, and cognitive scientists.

Critics have disagreed not only about the validity and soundness of Searle's argument, but even about its very point. For instance, David Cole, in his recent "Thought and Thought Experiments" (1984), has summarized Searle's argument thus:

> since a *human* simulation of a *machine* . . . simulation of human behavior which in humans normally evidences understanding would not itself involve that understanding, neither does the machine simulation. (Cole 1984, p. 431)

But this is not quite the argument. Rather, Searle's argument is (at least in part) that if a "hand tracing" by a *human* of a particular *program* does not involve understanding, then a *machine* "trace" won't either.

Several critics (e.g., Sharvy 1983, Cole 1984) have likened Searle's story to Leibniz's story about the thinking machine—

> Perception [is] . . . inexplicable by mechanical causes. . . . Supposing that there were a machine whose structure produced thought, sensation, and perception, we could conceive of it as increased in size with the same proportions until one was able to enter into its interior. . . . [O]n going into it he would find only pieces working upon one another, but never would he find anything to explain Perception. (Leibniz [1714] 1902, sec. 17)

*Received September 1984.

†An earlier version of this essay was presented as comments on Cole (1984) at the American Philosophical Association Western Division Colloquium on Cognitive Science, 26 April 1984. I am grateful to Andrea Schnall and to my colleagues in the SUNY Fredonia Department of Philosophy—Jane Kneller, Marvin Kohl, Michael Losonsky, Kenneth G. Lucey, and Morton L. Schagrin—for their comments on earlier drafts.

Philosophy of Science, 53 (1986) pp. 271–279.

—often concluding that both commit a compositional fallacy. But Searle himself has faulted the Leibniz story on precisely these grounds (Searle 1983, pp. 267f.). Moreover, while both stories may be fallacious, they are not alike. Leibniz *assumed* that his machine *could* think and perceive, apparently *concluding* that its working could not be purely mechanical, whereas Searle argues *from* the purely mechanical, but non-biological, character of the running program *to* its inability to exhibit understanding: Thus, one man's *modus tollens* becomes another's *modus ponens*.

Cole offers several arguments against Searle, including: (a) that "the machine simulation of a human" is not analogous to "the human simulation of the machine," and (b) that the "simulation of a machine" *might* "produce understanding" (Cole 1984, pp. 431–32).

The first of these claims is weak, but sheds some interesting light on the nature of computer programs and on the nature of understanding. The second is much stronger. But both miss what I shall argue are more serious objections to Searle.

2. Humans and Machines. Let us consider, first, the alleged disanalogy between humans and machines; namely, that the human in the room is following rules, but a machine does not: a machine would only "act in accord with rules" (Cole 1984, p. 438). But suppose (as Hofstadter 1980 has suggested) that the human completely internalizes (or "compiles") the rules, so that they are not being (consciously and slowly) followed, but merely acted in accord with (at great speed). Even so, Searle could still respond that the human *still* doesn't understand Chinese *in the way that* he understands English, because of a missing *semantic* component—the lack of correlation between the symbols and the world. That is, Searle could admit that the human might have "syntactic understanding" but not "semantic understanding." (I shall return to these notions later.) *This* response on Searle's part—which needs to be countered more carefully—shows that the alleged disanalogy is not of central importance.

To support his claim of disanalogy, Cole also says that "computer programs are *not* a series of rules, instructions or commands" (Cole 1984, p. 439). But this depends on what's meant by 'program': If a program is understood to be an algorithm coded in a programming language, then Cole's claim is simply false. If a program is understood as a "running process," then Cole may be right. Now, the sense of 'program' needed, not only for the Chinese-room experiment, but for functionalism and "strong AI" (Searle 1980, p. 417) as well, is a combination of these, an algorithm being executed. And this is precisely Searle's use. Searle does not claim—*no* one claims (or should claim)—that a program *qua algorithm* understands. Only *running processes* could possibly understand. The issue is not whether the processes are *algorithmic*—Searle concedes that. The issue is whether they can "produce" understanding if run in *any* medium, *even a non-biological one,* and that is what Searle denies.

3. Syntactic and Semantic Understanding. Let me digress briefly to clarify the nature of what I have referred to as "syntactic" and "semantic" "understanding." What I have in mind can best be explained by means of an anecdote.

In school, I learned to solve algebraic equations as follows: To solve the equation

$$2x + 1 = 3, \tag{1}$$

I performed the following steps:

1. Move the '+1' from the left side to the right, changing it to a '−1', yielding

   $$2x = 3 - 1.$$

2. Move the '2' from the left side to the right, changing it to a '1/2', yielding

   $$x = \frac{3-1}{2}.$$

3. Simplify the right side, yielding the solution,

   $$x = 1.$$

Clearly, such steps can be generalized and made more precise, providing an algorithm for solving linear equations in one unknown. But notice that each step is *purely syntactic:* symbol manipulation in its purest sense. I gained a great deal of skill in performing these manipulations, and my teachers, fellow students, and I all agreed that I "understood" how to solve such equations. I had *syntactic understanding*.

Later that year, I watched an educational-television program whose subject was solving equations. The technique presented was, to me, radically different and quite eye-opening. To solve equation (1), the viewer was told to think of the equation as representing a balancing scale, with weights representing '$2x + 1$' and '3' on each pan of the balance. (An actual balance was used for demonstration.) There was one constraint: The scale must be kept in balance. The procedure for solving the equation by keeping the scale in balance was this:

1. Remove a 1-unit weight from each pan, leaving weights of $2x$ units and 2 units on each pan.

2. Halve the weights on each pan, leaving weights of x units and 1 unit on each pan.
3. The solution to the equation is thus $x = 1$.

I was quite excited by this new, *semantic understanding* that I, but not my fellow students, now had.

The question is: Is semantic understanding something qualitatively different from syntactic understanding? Searle's argument can be construed as answering yes; I shall argue that although I had a *better* understanding of algebra, it was not a qualitatively different *kind* of understanding.

4. Machine Understanding. Cole's other argument is that the machine *might* produce understanding. His Experiment 4 (Cole 1984, p. 435) envisages a "merger" of Searle's brain with that of Hao Wang. This is a nice elaboration of Dennett's suggestions about "two people, one of whom understands Chinese, inhabiting one body, or . . . one English-speaking person . . . engulfed within another . . . who understands Chinese" (Dennett 1980, p. 429). It brings out clearly that the merged person (or the human in the Chinese room) *considered as Chinese*-responder possibly *does* (or can *claim* to) understand Chinese even if the merged person (or the human in the Chinese room) considered as English-speaker *claims* not to.

(i) Simulation and Implementation. On the one hand, the important question is this: Even if the man's understanding of English is *distinct* from his understanding of Chinese, are they in any way *similar? Is real* understanding of English different from *simulated* understanding of Chinese? As Cole puts it, Searle "seems to hold . . . that the microstructure of the system will make all the difference between *being* mental and merely being a *simulation of* the mental" (Cole 1984, p. 433; my italics). But it is arguable that there *is* no relevant difference between the mental and an (appropriate)[1] simulation thereof. (Appropriate) simulations or implementations of mental phenomena *are* mental, just as simulated mathematical theorems and proofs—or implementations of mathematical theorems and proofs—*are* mathematical theorems and proofs (see, for example, Lenat 1982), just as simulated information—or an implementation of information—(such as data in a data base) *is* information, just as (perhaps) a simulation (or implementation) of a hurricane *is* a (kind of) hurricane (see Hofstadter 1981, pp. 73ff.), and (perhaps) just as a certain kind of simulation of urea (namely, synthetic urea) *is* urea. Perhaps *some* simulated X's, or implementations of X's, aren't X's (such as Searle's examples (1980, p. 424) of computer simulations of milk and sugar), but others are.

[1] That is, not fake, like midgets inside chess-playing machines, or science-fiction robots.

One way to make a case for this is to view a mental phenomenon, such as thinking, as something abstract that can be implemented in two different media; say in a human brain and in a computer. The *computer* implementation of thinking can be said to be a *simulation* of the *human* implementation of thinking, and the two kinds of thinkings can be distinguished by differences between the implementing media, yet they are *both* species of thinking.

(ii) Semantics as Syntax. On the other hand, *Searle's* response to the two-subsystems view (see Searle 1980, pp. 419–20) is that the simulated understanding of Chinese is not real understanding of Chinese because of the lack of a *semantic* component to the program, as I alluded to earlier. Such a component can be provided, but it can (I would say *must*) be *internal* to the program, and hence would not help Searle's argument. Let me explain:

Consider, first, Cole's Experiment 5, in which Searle

> has been programmed in a way which modified a heretofore unused part of his brain so as to enable him to understand Chinese. Searle laughs, says "Don't be silly," and leaves. The lab director's discounted explanation of what happened was given in Chinese. (Cole 1984, p. 442)

This is cute, but without any further detail it begs the question as to whether he really does understand Chinese. However, one *could* build an AI system that had *two* natural-language parsers (one from Chinese and one from English, both parsing into a neutral representational language, such as a semantic network) but that had only *one* natural-language generator (from the semantic component into English).[2] This system could also have a data base (or "knowledge" base) expressed in the neutral representational language, containing the built-in datum, "I do not understand Chinese." This datum (or "belief")[3] would be expressible, by hypothesis, only in English.

Now, we should believe the *cliam-in-English* of Searle-in-the-story *not* to understand Chinese if and only if we believed that only the English parser works; but we need *not* believe this claim, if we believe that *both* parsers do work (and that the datum is simply a false belief). Since, by hypothesis, *both* parsers do work, then there is just as much understanding *here* as there would be in the following modification of the Chinese room:

Suppose that there is a Chinese parser with an English generator, so

[2] Such a system can easily be built using techniques described in Shapiro 1982.
[3] See Maida and Shapiro 1982; Rapaport and Shapiro 1984.

that the human in the room gives all answers to his Chinese interlocutors in English: Input 'squiggle', he outputs 'hamburger'; input 'squoggle', output 'yes'; input 'squiggle squiggle', output 'delicious'; and so on. *If* the Chinese subsystem of the human has *no* link to the English subsystem (no common language of thought), then there is just as much—or as little—understanding here as there is in Searle's version. But if there *is* a link—as I believe there must be—then learning and understanding can take place, just as, for example, an English speaker can learn to translate French into Spanish, by way of English.

Now this way of providing a semantics assumes that the computer *does* understand some other language (in this case, a neutral language of thought), but, of course, there is a more obvious way of forging the semantic link: by means of sense data. Suppose the symbols aren't Chinese characters, but arbitrary symbols. The human in Searle's experiment does *not* then understand Chinese; he's merely manipulating symbols (though it might be said that he understands Chinese syntax). So, in Searle's story, the fact that the symbols are *Chinese characters* is irrelevant, the texts being presented are *not* necessarily Chinese, and there need be *no* understanding (at least, no *semantic* understanding) of Chinese. But suppose the human (or machine) could relate the symbols to *other* (external, or sensory) information. *Then* it *could* understand (either Chinese or the symbolic notational variant thereof) in a semantic sense. Yet, as what Searle (1980, p. 420) calls the "robot reply"—combined with a bit of healthy methodological solipsism—shows, all of this data can (must) be internally represented. Thus, this tactic reduces to the previous one.

In other words, the (so-to-speak) "external" semantics is *either* a link to the outside world—which cannot be put into a program any more than it can be put into a *human* mind—*or else* it is a link to internal representations of external objects, in which case it *can* be made part of a program. But such semantic links are really just more syntactic symbol pushing. So even if *Searle's* program *as is* is insufficient for understanding, a mere program that could do more *could* understand.

5. Causation and Realization. Finally, let me mention briefly what I take to be the crucial issue in Searle's argument; namely, to use Cole's phrase, "the very form of materialism which Searle" embraces (Cole 1984, p. 432). It is at once a *narrow* form and a *wide* form: It is narrow in that minds are not only material, but a particular *kind* of material; yet it is wide in that Searle accepts part of the functionalists' claim that mental phenomena are algorithmic, insisting only that the algorithm will only exhibit intentionality when running in the right kind of material; namely, one having the requisite causal powers to produce intentionality (Searle 1980, p. 422).

Now, what are these powers? Searle does not tell us in "Minds, Brains, and Programs," but he does elaborate on it in his book, *Intentionality* (1983, chap. 10), introducing the notions of phenomena being "caused by" and being "realized in" some medium. These notions demand careful analysis. Here, I shall only briefly—and programmatically—state that the relationships between these notions (intimately related to those between abstract data types and their implementations) are more subtle than Searle imagines and that a computer program for simulating human intentionality that is "realized in" some implementing device *can* thereby give that device the requisite "causal powers" to produce intentionality. And so, as Richard Sharvy (1983) has said, "It ain't the meat, it's the motion."

APPENDIX
SEARLE'S CHINESE-ROOM THOUGHT EXPERIMENT

Suppose that I'm locked in a room and given a large batch of Chinese writing. Suppose furthermore . . . that I know no Chinese. . . . To me, Chinese writing is just so many meaningless squiggles. Now suppose further that after this first batch of Chinese writing I am given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English, and I understand these rules as well as any other native speaker of English. They enable me to correlate one set of formal symbols with another set of formal symbols, and all that "formal" means here is that I can identify the symbols entirely by their shapes. Now suppose that I am given a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch with the first two batches, and these rules instruct me how to give back certain Chinese symbols with certain sorts of shapes in response to certain sorts of shapes given me in the third batch. Unknown to me, the people who are giving me all of these symbols call the first batch "a script," they call the second batch a "story," and they call the third batch "questions." Furthermore, they call the symbols I give them back in response to the third batch "answers to the questions," and the set of rules in English that they give me, they call "the program." . . . [I]magine that these people also give me stories in English, which I understand, and they then ask me questions in English about these stories, and I give them back answers in English. Suppose also that after a while I get so good at following the instructions for manipulating the Chinese symbols and the programmers get so good at writing the programs that from the external point of view—that is, from the point of view of somebody outside the room in which I am locked—my answers to the questions are absolutely indistinguishable from those of native Chinese speakers. . . . Let us also suppose that my answers to the English questions are . . . indistinguishable from those of other native English speakers. . . . From the external point of view—from the point of view of someone reading my "answers"—the answers to the Chinese questions and the English questions are equally good. But in the Chinese case, unlike the English case, I produce the answers by manipulating uninterpreted formal symbols. As far as the Chinese is concerned, I simply behave like a computer; I perform computational operations on formally specified elements. For the purposes of the Chinese, I am simply an instantiation of the computer program.

Now the claims made by strong AI are that the programmed computer understands the stories and that the program in some sense explains human understanding.

[I]t seems to me quite obvious in the example that I do not understand a word of the Chinese stories. I have inputs and outputs that are indistinguishable from those of the native Chinese speaker, and I can have any formal program you like, but I still understand nothing. For the same reasons, . . . [a] computer understands nothing of any stories. . . .

[W]e can see that the computer and its program do not provide sufficient conditions of understanding since the computer and the program are functioning, and there is no understanding. But does it even provide a necessary condition . . .? One of the claims made by the supporters of strong AI is that when I understand a story in English, what I am doing is exactly the same . . . as what I was doing in manipulating the Chinese symbols. . . . I have not demonstrated that this claim is false. . . . As long as the program is defined in terms of computational operations on purely formally defined elements, what the example suggests is that these by themselves have no interesting connection with understanding. . . . [W]hatever purely formal principles you put into the computer, they will not be sufficient for understanding, since a human will be able to follow the formal principles without understanding anything (Searle 1980, pp. 417–18).

REFERENCES

Cole, David (1984), "Thought and Thought Experiments", *Philosophical Studies 45*: 431–44.

Dennett, Daniel (1980), "The Milk of Human Intentionality", *Behavioral and Brain Sciences 3*: 428–30.

Hofstadter, Douglas R. (1980), "Reductionism and Religion", *Behavioral and Brain Sciences 3*: 433–34.

———. (1981), "The Turing Test: A Coffeehouse Conversation", in *The Mind's I,* D. R. Hofstadter and D. C. Dennett (eds.). New York: Basic Books, pp. 69–92.

Leibniz, G. W. v. [1714] (1902), *The Monadology*. In *Leibniz*. Translated by G. R. Montgomery. La Salle IL: Open Court.

Lenat, D. B. (1982), "AM: An Artifical Intelligence Approach to Discovery in Mathematics as Heuristic Search", in *Knowledge-Based Systems in Artificial Intelligence*, R. Davis and D. B. Lenat (eds.). New York: McGraw-Hill.

Maida, Anthony S., and Shapiro, Stuart C. (1982), "Intensional Concepts in Propositional Semantic Networks", *Cognitive Science 6*: 291–330.

Pylyshyn, Zenon W. (1980), "The 'Causal Power' of Machines", *Behavioral and Brain Sciences 3*: 442–44.

Rapaport, William J., and Shapiro, Stuart C. (1984), "Quasi-Indexical Reference in Propositional Semantic Networks", *Proceedings of the 10th International Conference on Computational Linguistics* (COLING-84). Morristown NJ: Association for Computational Linguistics, pp. 65–70.

Searle, John R. (1980), "Minds, Brains, and Programs", *Behavioral and Brain Sciences 3*: 417–57.

———. (1983), *Intentionality*. Cambridge: Cambridge University Press.

Shapiro, Stuart C. (1982), "Generalized Augmented Transition Network Grammars for Generation from Semantic Networks", *American Journal of Computational Linguistics 88*: 12–25.

Sharvy, Richard (1983), "It Ain't the Meat, It's the Motion", *Inquiry 26*: 125–34.