Ed Cottrell

PHIL 103

April 7, 1998

Programs Versus Reflections

John Searle's argument against strong Artificial Intelligence is based largely on his so-called "Chinese Room" thought experiment,[1] to which many replies exist, the best known of which is Hofstadter's "Reflections."[2] Searle's Chinese Room is a scenario in which a person is locked in a room with two batches of Chinese script and "a set of rules for correlating the second batch with the first batch."[3] The subject of the experiment is then given another set of symbols and rules for correlating these with the second batch. Given that these rules are in English (or the subject's native tongue), the subject of this experiment can take in "input" in the form of Chinese writing (the first and second batches of symbols), apply the set of rules, and produce a logical response, also in Chinese (some arrangement of the third set of symbols). Searle's thesis is that such a person has no more understanding of Chinese than before the experiment began, to which there are a number of replies, some of which he analyzes in turn. "Strong Artificial Intelligence," hereafter called strong AI, as defined by Searle, is the claim that the "appropriately programmed computer really *is* a computer," whereas "weak AI" is defined as the idea that a computer can be only a tool for studying the mind, and not a mind in itself.[4] Hofstadter's response is to dismiss Searle's work as a series of "illusions" and a "serious and fundamental misrepresentation," which is the "tremendous difference in complexity between two systems at different conceptual levels." Hofstadter does this primarily

---

[1] Searle, John R. "Minds, Brains, and Programs." The Behavioral and Brain Sciences. Cambridge: Cambridge UP, 1980. 417-424.

[2] Hofstadter, Douglas R. "Reflections."

[3] Searle, 418. Searle's actual proposal is that he himself be locked in the room, but any person who does not speak Chinese is satisfactory. Hofstadter calls this person the "demon" (Hofstadter, 3).

[4] Searle, 417. Searle's paper deals almost entirely with strong AI.

by analyzing what Searle calls the "systems reply," which is the idea that it is the whole system of subject, room, and symbols, which understands the Chinese.[5] Hofstadter claims that the sheer magnitude of the information required and the slowness of human beings in performing the necessary tasks make Searle's experiment not only impractical but also would prevent the system as a whole from passing a Turing test.[6] This argument overlooks the theoretical nature of the Chinese room and changes the central question of whether or not the system of the room can understand Chinese to whether or not it could literally pass the Turing test. Just as the impracticality of the infinite look-up tree in a "Blockhead" is not relevant to the principle it is illustrating, neither is the impracticality of a Chinese room whose subject could pass a Chinese Turing test.[7] In any case, Searle sees his experiment as a possible counterexample to the Turing test and sees "causal properties" or "intentionality" as a necessary element of understanding,[8] while Hofstadter does not. Hence, Searle's and Hofstadter's arguments form an intellectual stalemate. The purpose of this paper will therefore be to examine each philosopher's handling of the Chinese room thought experiment and the various replies to it, particularly the systems reply, and to suggest possible new arguments to demonstrate the flaws and merits of each opinion.

Searle's initial argument is simply that the subject in the Chinese room experiment does not understand Chinese, but instead is practicing "formal symbol manipulation."[9] The immediate reply to this is the systems reply, which is simply that the whole system must be considered as a candidate

---

[5] Hofstadter, 1.

[6] A Turing test consists of an interrogator posing questions to a subject, attempting to determine whether or not the subject is human. This is a simplification – see Turing, Alan M. "Computing Machinery and Intelligence" for the original scenario.

[7] A Blockhead is a machine designed to mimic human behavior by containing a look-up tree of every possible stimulus and reaction – an impossible situation in reality, but posed only as a thought experiment. The point it illustrates is, arguably, that it could pass a Turing test without any actual intelligence. See Block, Ned. "The Mind as the Software of the Brain."

[8] Searle, 421. Note that Searle questions the Turing test only in the context of the systems reply, and that he never defines "causal properties."

[9] Searle, 418.

for understanding, rather than just the human subject. Searle dismisses this by offering to have the subject "internalize" or memorize the rules and symbols, and claiming that the subject would still have no understanding of Chinese.[10] Hofstadter dismisses Searle's argument in turn as impractical and thus meaningless. Of course, this is contrary to the nature of a thought experiment, but Hofstadter continues. Hofstadter's next claim relies on transforming the Chinese room by a series of knobs on a hypothetical "thought experiment generator" into a very small and very fast "elf" inside a Chinese speaker's head. The basic idea behind this "transformation" is Hofstadter's claim that the set of properties describing the Chinese room can be changed into the properties describing a Chinese speaker's head. Hofstadter breaks these properties down into five categories: physical components, level of simulation of the human brain, physical size of the system, physical size of the "demon," and speed of the "demon." The properties Hofstadter assigns to the Chinese room are, respectively: "paper and symbols," "concepts and ideas," "room size," "human-size," and human speed. The properties assigned to the "elf" in the Chinese speaker's head are: "neurons and chemicals," "neuron-firings," "brain size," "eensy-weensy," and "dazzlingly fast." Hofstadter points out that Searle would accept that the speaker understood Chinese on the grounds of "causal properties," and claims that this constitutes a contradiction in Searle's argument. Admittedly, Searle never explicitly defines "causal properties;" he merely claims that they are inherent to the human brain. However, to avoid classifying Searle as a Cartesian dualist and negating the utility of this discussion, the most logical interpretation is that "causal properties" are best defined as "intentionality" and "meaning." In this case, Searle's claim makes some sense in that it is possible that the neurons of the Chinese speaker's brain could still contain meaning somehow and retain their original networking, although the synapses are damaged, as specified by Hofstadter.[11]

[10] Searle, 419.
[11] Hofstadter, 6.

Hofstadter does not directly address this possibility, and does not address whether or not the same intentionality and meaning found in the Chinese speaker's head can be found in the rules in the Chinese room. At this point, Hofstadter's argument begins to become somewhat paradoxical. Hofstadter's claim is that "a *true understanding* of [a] language" consists of "mixing the new language right in with the medium in which thought takes place."[12] Hofstadter continues this argument and claims that the issue at question is "level-mingling," the "ability of a higher level to loop back and affect lower levels – it's own underpinnings – … a kind of magic trick which we feel is very close to the core of consciousness."[13] The implication here is that, for a system to be capable of actually learning a language, it must be at least close to consciousness in that it must be able to alter itself. While the subject in the Chinese room might be able to learn the Chinese language eventually, doing so by "decoding" the rules and symbols in the Chinese room would take the greatest linguists an extremely long time.[14] However, Hofstadter maintains that the system of the room, the subject, and the rules somehow understands Chinese without having learned Chinese. The only response consistent with the systems reply is that it is the bits of paper which learned Chinese from the "programmer." Hence, the rules-programmer system understands Chinese. One objection that could be raised is that the programmer could die, yet his or her understanding would stay. However, this contradicts Hofstadter's claim above that "a *true understanding* of [a] language" consists of "mixing the new language right in with the medium in which thought takes place."[15] This would imply that there is some medium in which the written version of the programmer's understanding has thought; in other words, the written version is a sentient creation. This is something of an extreme claim. Note also that there is a change of systems here – even if the rules-

[12] Hofstadter, 7.
[13] Searle, 9.
[14] Note that I am not saying the subject would learn Chinese from his or her participation in the experiment – "decoding" the entire Chinese language from a set of symbol-replacement rules is a nearly impossible task.
[15] Hofstadter, 7.

programmer system understands, this does not mean that the room-subject-rules system understands anything in Chinese. The systems reply must then maintain that the combination of the rules and their application is what gives the system understanding. Note, though, that all understanding of Chinese must reside in the rules themselves – the subject works merely as an uncomprehending interpreter. This seems ridiculous and again contradicts Hofstadter's above claim. The following example should help demonstrate the implications of extending understanding from a programmer to a program.

If a set of rules alone can have understanding (as in a computer program), then my word processor's spell-checker understands English word structure. It can analyze the symbols I enter in the form of strings of ones and zeros, look for inconsistencies according to a set of rules (i.e., *x* must be an element of *y*, and *a* must never follow *b*, etc.), and produce a response. Of course, this is arguable, since an average human speaker can recognize that "dfhts" is not a possible word of the English language, yet my word processor can only tell me that it is not in any of its databases. Likewise, my computer's grammar checker must understand English sentence structure, so my word-processor as a whole (encompassing the spell- and grammar checkers) understands the structure of English language as a whole. Of course, this is also arguable on the basis that the computer accepts "she shoot the moon" as valid grammar, but this is merely a question of the complexity of the program. Still, as Searle would point out, the spell-checker has no concept of meaning, no "causal properties" or "intentionality." The difference between "dog" and "Doug" lies entirely in spelling for my computer. So, we add a dictionary into the word processor (an easy inclusion), and we have a program which "understands" English language structure and "meaning." Now, in what sense can my word processor attribute meaning to a word? Obviously, the dictionary component implies that it can attribute a semantic meaning to "dog," such as "a furry mammal weighing between twenty and one hundred fifty pounds, walking on four legs, having a tail, and

often kept as a house pet," etc. However, can such a clinical description constitute understanding of something as simple as the word "dog?" Such descriptions can never constitute a full understanding of a subject. For example, there is remarkably little consensus on what it means to be "patriotic" or "fair." If humans cannot agree that other humans understand the meanings of certain words, how can we attribute that understanding to a dictionary? One claim is that, since humans do not fully understand some concepts, it is sufficient for a program to understand only to the same level as humans do, and no more is required. However, this raises the issue of whether or not such a level of understanding is even possible for a non-human computer. For example, the human concept of "dog" inherently includes notions of "friendly," "mean," "cuddly," "smart," "stupid," etc. How can a dictionary definition constitute a full understanding of the adjectives "cuddly" or "smart," or, by extension, the noun "dog?" How could such a program understand what it "means" to be an American or to be "single?" The emotional baggage in such concepts is as essential to the accepted meanings as the literal definitions. It seems impossible, then, that even an advanced word processor, including a spell-checker, grammar checker, and dictionary, does not understand at least abstract concepts, and therefore has no conceptual understanding of concrete concepts associated with abstract concepts, such as "dog," which can be associated with "nice" or "smart." Questions of complexity do not apply here, since it is impossible to convey fully some such concepts in mere words. Thus, it seems that none of the features of my word processor, or even combinations thereof, can be said to understand English at any level. The "Robot Reply" Searle cites also seems ridiculous here.[16] It does not make sense that a dictionary definition alone is insufficient to understand "patriotic," while the combination of word processor, video camera, microphone, and pressure sensors, to simulate human sense, is sufficient to give meaning to such an abstract concept. I do not

[16] Searle, 420.

claim to know how the human brain attributes meanings to such concepts; I merely claim that the above systems are not matched to the task.

Although Searle's argument is not the clearest in places, particularly with respect to "causal properties," Hofstadter's reply largely misses the point. The biggest problem with Hofstadter's paper as I see it lies in his misinterpretation of the nature of the Chinese room thought experiment. The claim that complexity or speed play relevant roles in whether or not the Chinese room system understands Chinese does not hold much weight. For example, a person with a speech impediment may be very slow in responding to questions, yet this has nothing to do with whether or not he or she understands the question. Likewise, a student who is learning Spanish may have difficulty following a Spanish conversation, yet may be able to apply basic rules of Spanish grammar and contextual information to "fill in" gaps in his or her knowledge. This does not mean that the student has no understanding, just that it is not as complete as the understanding that a native speaker has. In fact, people frequently use basic rules to follow conversations in their native language. The other major problem with Hofstadter's paper is the digression on learning secondary languages, which, as I have shown above, becomes somewhat paradoxical and contradictory at times. Of course, Searle's paper is hardly perfect, either, particularly in terms of the undefined "causal properties," yet is consistent with itself, at least. The rebuttals offered by Searle to various replies made to his argument, such as the "combination reply" or the "many mansions reply," that these replies trivialize strong AI by dodging the issue somewhat, seem to be fairly sound.[17] These replies are worth analysis in their own right, but that was not the goal of this paper. With respect to the systems reply, Searle's argument holds more water than Hofstadter's does.

[17] Searle, 421-2.