GUNDERSON AND SEARLE: A COMMON ERROR ABOUT ARTIFICIAL INTELLIGENCE

Glenn C. Joy

Of course it all goes back to Turing. The paper, published in *Mind* in 1950, was entitled "Computing Machinery and Intelligence," and Alan Turing asks in the first sentence, "Can machines think?"[1] Instead of answering the question, he replaces it with one that he considers less problematic: "Are there imaginable digital computers which would do well in the imitation game?"

The account of the original imitation game has been garbled in the intervening years by constant retelling, so let me briefly summarize it.

The imitation game is played by three people (or two people and a digital computer): a man (or a computer) (A); a woman (B); and a third person, the interrogator (C). The interrogator, communicating with A and B via writing or teletype or some such media, may ask questions of "X" and "Y," and eventually must try to determine whether X or Y is B, the woman. The man (A) is to try to fool the interrogator (C), and the woman (B) is to try to help the interrogator (C) make the correct decision. The question of whether machines think is replaced by the question of whether the interrogator (C) will decide incorrectly who the woman (B) is just as often if a machine takes the place of the man (A) in the game.

For better or worse, the name has changed from "imitation game" to "Turing test," and most actual and hypothetical tests have been *imitation* games, usually involving an interrogator trying to decide if he or she is communicating with a person or a machine.

Keith Gunderson developed what he labelled a "parody comparison" to the Turing game, calling it the "toe-stepping" game.[2] The "toe-stepping" game is played by three people (or two people and a rock box): a man (or a rock box) (A); a woman (B); and a third person, the "interrogator" (C). The interrogator is in one room, and the other two are in the next room. The interrogator can place most of a foot through a hole in the wall into the other room, whereupon one of the two may step on a toe. The interrogator may specify whether X or Y is to

crush the toe, and after a suitable number of trials the interrogator tries to guess whether X or Y is the woman (B).

The question, "Can rocks imitate?" (analogous to "Can machines think?") is replaced by the question of whether C will decide incorrectly who the woman is just as often if a rock box takes the place of the man (A) in the game. A rock box is a box filled with rocks and rigged with an electric eye and appropriate pulleys so that when the toes appear in the hole in the wall they trigger the mechanism to smash the toes and then the pulleys move the box off the toes.

Gunderson's view of all this is that, since it is perfectly sensible to agree that while the rock box might play the toe-stepping game well, one can still ask, "Can rocks imitate?" Similarly, even though a machine might play the imitation game, one can still ask, "Can machines think?"[3] In other words, Gunderson's view is that playing the imitation game--no matter how well--does not have any bearing on answering the question, "Can machines think?"

But possibly the best known hypothetical game is that of John Searle, published in his article "Minds, Brains, and Programs"[4] and often referred to as Searle's "Chinese room." Searle imagines himself (knowing no Chinese) being locked in a room. He is given some Chinese writing, then given some additional Chinese writing along with a set of rules in English that tell him how the first writing is to be correlated with the second. Then, he is given a third batch of Chinese writing with instructions that tell him how to correlate the third with the first and second, and how to send back out of the room certain Chinese symbols for certain symbols that may be found in the third batch of writing. He is not aware of it, but outside the room the first Chinese writing is called a "script" (like Roger Schank's scripts, which are modules of information about a small segment of life or society or knowledge), the second is called a "story," and the third, "questions." The rules written in English are called a "program," and the symbols that are sent out of the room are called "answers."

Searle imagines that the "answers" sent out of the room are "absolutely indistinguishable from those of native Chinese speakers." Searle believes:

> 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.[5]

But Searle's claim is simply that he could make all these computations and still not understand Chinese. Searle believes that this shows that passing Turing-like tests does not require anything like human mental states--consciousness, understanding, or whatever.

Both the rock box and the Chinese room seem like impressive examples when they are first encountered. However, a little reflection shows that they are set up and described in such a way that they contain serious logical errors and that they in fact do not provide useful thought experiments with regard to the question, "Can machines think?"

The problem with Gunderson's example is that he concludes his analogy incorrectly. Notice that saying (as he does), "Of course a rock box of such-and-such a sort can be set up, but rocks surely can't imitate,"[6] is not a parallel to saying, "Yes, a machine can play the imitation game, but it can't think."[7] In the first case, the negative part of the sentence is the result of his shift from "rock box" (the total system) to "rocks" (a part of the system); and, in the second case, "machines" (meaning the total system) or the pronoun "it" appears twice. If we were to make the Chinese room example analogous to the rock box example, we would conclude that "The machine can play the game, but transistors (or central processing units or whatever) cannot think." This is true but uninteresting. On the other hand, if Gunderson had really made the rock box example analogous to the way the Chinese room example should be interpreted, he would conclude, "The rock box can play the game, but it (the whole system of the rock box) can't imitate." And this is clearly false; if it plays the game successfully, it *is* imitating.[8]

Searle's room is more persuasive than Gunderson's box because he is raising the question of understanding rather than

the question of mere imitation. Hence it gets right to the issue. Its persuasiveness also lies in the fact that we tend to put ourselves in the place of the person in the room and realize we would understand the English and would not understand the Chinese. So we "know" from this first-hand thought experiment that we could pass a Turing test and not understand a bit of Chinese.

Searle, however, has made a mistake that is similar to the one that Gunderson made since Searle's thought experiment requires one to imagine oneself processing information in the room using a "programmed" set of instructions. Hence it is the complete room that is analogous to the computer and not just the person. The person is analogous, at best, to the central processing unit of a computer. And of course no one claims that a programless CPU can think. Searle does not appear to be thinking about this as evidenced by the indented quotation above. In the course of two sentences he first says that he, the person in the Chinese room, is "an instantiation of the computer program" and then says that the AI claim is "that the programmed computer understands the stories." And of course just because a *computer program* does not have understanding has little or no bearing on whether the *programmed computer* does. Similarly, Searle says on the same page that "the computer has nothing more than I have in the case where I understand nothing." But the computer and its programs and its other inputs "have" a lot more than the person in the Chinese room.

It is unfortunate that this error is (apparently) so easy for philosophers to make, and it is unfortunate that is has not been more readily detected by other philosophers. Yet, it is easily detected by persons who have a little knowledge of the operation of computers and their programs.

I would like to point out how Searle responds to criticisms similar to mine. Searle tried to anticipate criticisms and to respond at the end of his article. Searle goes so far as to say that he is actually embarrassed (for his critic) because it is so "implausible" to claim "that while a person doesn't understand Chinese, somehow the *conjunction* of that person and bits of paper might understand Chinese."[9] He says, "Let the individual internalize all of these elements of the system."[10] One will still not have understanding. Whereas I

understand that "'hamburgers' refers to hamburgers, the Chinese subsystem knows only that 'squiggle squiggle' is followed by 'squoggle squoggle.'"[11] But of course there is absolutely no reason the Chinese *room*--there's Searle's mistake again--or the person who has internalized all the rules, cannot know that "squiggle squiggle" refers to "hamburgers" since the instructions are written in English. If a student studying a foreign language was only taught that "squiggle squiggle" means "squoggle squoggle," a hardworking student might memorize enough to pass the course without understanding the language, but certainly the programmer--the teacher--should be fired.

In his recent book *The Mind's New Science*, Howard Gardner suggests that Searle may have defined thinking and intentionality such that by definition only humans are candidates for these qualities.[12] If he or anyone else actually does this, then it renders the question of machine intelligence easy to answer but meaningless. But of course the question is not meaningless. However, even if we do not define away the possibility of machine consciousness or whatever, it will always be possible to maintain that computers do not have it. Reflect for a second on the question of "other minds" in humans. The lack of privileged access to another's mind makes it difficult to formulate an argument to establish other minds--which is nevertheless what everyone believes. It will for similar reasons be even harder to deal with the "problem of nonbiological minds."

It is possible to maintain that computers will never be able to do the things that those with the greatest faith and commitment to artificial intelligence maintain that they will be able to do. It will be possible to maintain that those abilities and mental qualities do not exist even if they in fact do exist. Yet, I'm afraid that we may be like an observer at Kitty Hawk who might say that a pretty good trick occurred that day, but for whom the idea that a grandchild might routinely fly across the country would be preposterous. We must remember that the history of artificial intelligence is extremely short, but in spite of that many things have been accomplished that scarcely seemed possible even ten years ago.

It is always possible to maintain that evolutionary changes have not occurred or that persons have not been to the

moon or that the earth is flat or whatever. But we must be watchful that our skepticism does not become unfalsifiable dogma. There comes a time, sometimes, when the rational person must decide that the weight of evidence is against him or her and give in as gracefully as possible. There comes a time, sometimes, when the view one has argued against finally falls of its own dead weight and its inability to stand up to experiment or analysis. But for the issue of machine intelligence and intentionality, neither time has yet come.

NOTES

[1]Alan Turing, *Mind* 59 (1950): 4-30, reprinted in Douglas Hofstadter and Daniel Dennett, *The Mind's I* (New York: Bantam Books, 1981), 53-67.

[2]Keith Gunderson, *Mentality and Machines*, 2d ed. (Minneapolis: University of Minnesota Press, 1985), 44.

[3]Ibid.

[4]John Searle, *The Behavioral and Brain Sciences* 3 (1980), reprinted in Hofstadter and Dennett, 353-73.

[5]Ibid., 356.

[6]Gunderson, 44.

[7]Ibid.

[8] My claim that, if it plays the game, it is imitating, is of course predicated upon my belief that Gunderson does not intend to include human-like intentionality in the notion of imitating. If he does, the rock box parody would itself be a parody of human versatility of intention. Since this is not his intention, Gunderson must be concerned only with the imitation of toe smashing. And as I said before, if it plays the game, it is imitating.

[9]Searle, 359.

[10]Ibid.

[11]Ibid.

[12]Howard Gardner, *The Mind's New Science* (New York: Basic Books, 1987), 175.