A Short Rebuttal to Searle

Nils Nilsson

November 6, 1984

In his draft paper entitled "Computers, The Mind, and Responsibility" prepared for the November 9, 1984 Office of Technology (OTA) meeting, John Searle concludes that machines of the sort being developed in artificial intelligence (AI) research are not ones that could ever be called *responsible*. While I agree that AI research has a long way to go (perhaps several decades) before it might produce responsible machines, I disagree with Searle's arguments for why AI will, in principle, *never* produce such machines. Although the outcome of our argument may not have any material effect on public policy before the end of this century, I fear that Searle's point of view, if unopposed, could grievously misinform intellectual consideration of this matter.

Searle argues against what is called in computer science the "symbol system hypothesis." Roughly speaking, this hypothesis states that the essential and criterial aspects of intelligent behavior are properties of *formal* symbol systems, that is of abstract systems that manipulate symbols. The hypothesis claims that, although such systems must (of course) be embodied in *some* physical apparatus in order to function, it is completely irrelevant to their function *which* embodiment they are given—except for important practical matters of speed, cost, and efficiency. As far as the symbol system hypothesis is concerned, the explanation of intelligent behavior is to be found in the *rules* for manipulating symbols, not in the properties of the silicon, protein, or mechanical relays whose functioning implements symbol manipulation. I, and many others, subscribe to this hypothesis.

Searle is one of those who seems to think that the properties of the physical stuff involved in symbol manipulation make an essential difference. He claims to have laid this matter to rest with his "Chinese Room" analysis contained in his paper entitled "Minds, Brains, and Programs, " appearing in *The Behavioral and Brain Sciences* (Volume 3, 1980). In that paper he imagines a room containing a (non-Chinese-speaking) person following written rules about how to answer questions submitted to the room in the form of written strings of Chinese characters. Since the person in the room doesn't really know any Chinese and is merely following formal rules, Searle thinks it implausible that anyone would say that either the person, the room, the rules, or the combination of any of these *understands* Chinese. He claims therefore to have demonstrated that *understanding*, at least, involves more than formal symbol manipulation.

Actually, the commentary following Searle's paper makes it abundantly clear that the symbol system hypothesis was not damaged in the "Chinese Room." On the contrary, most computer scientists would say that Searle's analysis suffers from an inadequate treatment of the differences between *programs* and their *interpreters*. Also, the Chinese Room did not have certain very important programs that all human language understanders have in addition to those that enable them to understand language. Give the Chinese Room appropriate programs for self-reflection, for example, as well as programs for understanding Chinese, and the room itself will claim that it understands Chinese. After a reasonable amount of dialogue with the room, I would have to treat such a claim with the same courtesy and belief with which I would treat Searle's claim that he understands English. (The fact that AI has not yet produced programs capable of self-reflection does not really weaken the symbol system hypothesis. Whatever the key to self-reflection turns out to be, it clearly will involve the processing of internal symbolic representations. Few would believe that it is somehow inextricably tied to the physical properties of protein or silicon.)

Notwithstanding Searle's opinion that believing that computers could think is "silly," I (and many of my colleagues) are quite willing to defend (in public) the view that a machine could think "solely in virtue of [having] the right computer program." After all, I would say that a *multiplier* can *multiply* solely in virtue of its having a *multiplication* program. Such a view is not *profoundly anti-electronic nor profoundly anti-scientific*. It does not *deny everything we*

have learned about the relation of multiplication to transistor physics over the past thirty years. Granted, the characteristics needed for the ascription of responsibility to machines are vastly more complex than those needed for multiplication, but they are *only* vastly more complex. There is no reason to suppose that somewhere along the spectrum from multiplication to the higher aspects of human thought it gradually (or suddenly) becomes more and more necessary to involve in an essential way the physical properties of the medium of that thought.

In contrasting *as if* behavior with *the real thing*, Searle sees several distinctions that just aren't there. Certainly there are distinctions between simulations and the things simulated. Meteorological programs are not hurricanes. But when one discusses *programs* and *their* simulations, one needs to have a more articulated and sophisticated view of these matters. (One needs a clear understanding of the computer science notion of an *interpreter*, for example.) For the purposes that Searle has in mind, it is difficult to maintain a useful distinction between programs that multiply and programs that simulate programs that multiply. If a program behaves *as if* it were multiplying, most of us would say that it is, in fact, multiplying. For all I know, Searle may only be behaving *as if* he were thinking deeply about these matters. But, even though I disagree with him, his simulation is pretty good, so I'm willing to credit him with real thought.