The Chinese Room

In introducing his famous ‘Chinese Room’ thought-experiment, Searle articulates a methodological maxim, which he proposes to follow: ‘A way to test any theory of mind is to ask oneself what it would be like if one’s own mind actually worked on the principles that the theory says all minds work on’. Unfortunately, the question that Searle in fact goes on to ask is not this. Rather, he asks himself what it would be like if he were part of a mind that worked according to the principles that strong AI says all minds work on—in particular, what it would be like if he were the central processing unit (or the ‘interpreter’, as it is called in some systems). In other words, he imagines himself as a homunculus inside the complete AI system, self-consciously doing the work that the central processor (i.e. a part of the implementation or ‘hardware’) would normally do.

More specifically, as everybody knows, Searle considers a hypothetical system that meets the usual fluency tests for understanding Chinese, and then asks whether he, as its central processor, would ipso facto also understand Chinese. Plausibly, he maintains that he would not. But neither the question nor the answer is very interesting, since no one would ever have argued otherwise. So what could he have been thinking?

The form of his argument is a modus tollens. Thus, from the premises that (i) the hypothetical system understands Chinese, and (ii) if the system understands Chinese, then Searle himself does too, it would follow that (iii) Searle himself understands Chinese. Since the conclusion is false, one of the premises must be
false. In order to show that it's the first premise that's false, Searle must argue that the second—the conditional—is true. How can he do that?

Well, in the first version of his thought-experiment, he simply slides from characterizing himself as a part of the system—namely, its central processor—to speaking of himself as if he were the whole. Thus, in the paragraph where he draws the moral from his initial example, he writes: 'I have inputs and outputs that are indistinguishable from those of the native Chinese speaker'; and, even more blatantly: 'in the Chinese case, the computer is me'. Of course, if Searle just were the system—all of it, not just part of it—then the above conditional would be trivial, as would the whole argument, by the indiscernibility of identicals. But that simply isn't the case that Searle initially described.

The argument resting on the slide from part to whole is an obvious fallacy. Thus, it's not surprising that Searle's early interlocutors pointed it out to him—a point that survives in his published text as what he calls 'the systems reply': it isn't the central processor that they say understands conversations or stories, but rather the system as a whole. But Searle's response to this obvious objection is very odd. He claims that he could internalize the whole system, by simply memorizing all of its code and data, and then carrying out all of the operations in his head.

Notice first that this still isn't a case of Searle asking what it would be like if his own mind worked as the theory says. Rather, what he now asks is what it would be like if he, in his own mind, were consciously to implement the underlying formal structures and operations that the theory says are sufficient to implement another mind. And to that question, he replies, again plausibly, that he himself wouldn't ipso facto understand whatever that other mind supposedly does—namely, Chinese. But this is again irrelevant, and for a similar reason: the claim was never that the implementing hardware and/or software would understand anything, but rather the systems implemented in them. In other words, Searle has simply shifted from a part–whole fallacy to a level-of-description fallacy.

In his response to the systems reply, however, Searle also offers another argument, one that might be thought to address the level-of-description issue. So we had better look at this response a little more closely. What we are to imagine in the internalization fantasy is something like a patient with multiple personality disorder. One 'personality', Searle, is fluent in English (both written and spoken), doesn't know a word of Chinese, and is otherwise perfectly normal (except that he has the calculative powers of a mega idiot savant). The other ostensible personality—let's call him Hao—is fluent in Chinese (though only written, not spoken), has no English, and, moreover, apart from seeming to be
able to read and write, is deaf, dumb, blind, and paralysed. Indeed, it's worse than that: for, were Hao's Chinese interlocutor to present the (written Chinese) characters for 'Could you please write a little larger?', Hao's response (also in written Chinese) would have to be something like 'Gee, so far as I can tell, I'm communicating with you by mental telepathy; I have no knowledge of any writing at all'.

Searle seems to think this would be pretty bizarre; and who could disagree? But why, exactly, should we conclude that Hao doesn't understand the Chinese that he appears to be reading and writing ('automatically', as it were)? According to Searle's new argument, it's 'because there isn't anything in the system [Hao] which isn't in him [Searle]', and he doesn't understand Chinese. Let's spell this out. From the premises that (i) whatever is in Hao is in Searle, and (ii) there is some understanding in Hao, it follows that (iii) that very understanding is in Searle too. But what does this conclusion mean? One all too easy way to take it is: Searle himself has that understanding—thus, that he himself understands whatever Hao does. In particular, if Hao understood Chinese, so would Searle—which is exactly the conditional he needs for his modus tollens refutation of the premise that Hao understands Chinese.

But taken in that way, the argument equivocates on the word 'in'. When we say that the understanding that a person has is 'in' that person, what we mean by 'in' is something like 'is a capacity or feature of'. (Surely we mean more than that; but all that's needed to establish an equivocation is enough specificity in the two senses to show their difference.) But, when we agree that whatever is in Hao is 'in' Searle, we mean by the latter 'in' something like 'is implemented in some of the internal processes in'. So, if the argument is not to equivocate, this is what the 'in' in the conclusion must mean. Since this second sense is manifestly not the same as the first, Searle is not entitled to conclude that, were Hao to understand Chinese, so would he. Hence, the modus tollens still fails, for lack of the needed conditional.

The only way to rebut this diagnosis (or to have overlooked it) would be to maintain that being 'implemented in some of the internal activities in' something entails being 'a capacity or feature of' that same something. But that's just the level-of-description fallacy all over again. It would be like inferring from the fact that some of the internal activities in a cockroach implement the ability to seek out and recognize hostile viruses, that the cockroach itself can seek out and recognize hostile viruses; or like inferring from the fact that some of the internal processes in an engine implement the feature of rotating, that the engine itself is rotating.