to be conscious. We may continue to think of this by positing a stream of consciousness of sorts, but we no longer endow that stream with all of its traditional properties. Now that the stream of consciousness has been reconceived as the operations of a virtual machine realized in the brain, it is no longer “obvious” that we are succumbing to an illusion when we imagine such a stream occurring in the computer brain of a robot, for instance.

McGinn invites his readers to join him in surrender: It’s just impossible to imagine how software could make a robot conscious. Don’t even try, he says. Other philosophers have fostered this attitude by devising thought experiments that “work” precisely because they dissuade the reader from trying to imagine, in detail, how software could accomplish this. Curiously, the two best known both involve allusions to China: Ned Block’s (1978) Chinese Nation and John Searle’s (1980, 1982, 1984, 1988) Chinese Room.¹ Both thought experiments rely on the same misdirection of imagination, and since Searle’s has been the more widely discussed, I will concentrate on it. Searle invites us to imagine him locked in a room, hand-simulating a giant AI program, which putatively understands Chinese. He stipulates that the program passes the Turing test, foiling all attempts by human interlocutors to distinguish it from a genuine understanding of Chinese. It does not follow, he says, from this merely behavioral indistinguishability that there is any genuine understanding of Chinese, or any Chinese consciousness, in the Chinese Room. Searle, locked in the room and busily manipulating the symbol strings of the program according to the program, doesn’t thereby gain any understanding of Chinese, and there is nothing else in the room that understands Chinese either (this is “just obvious,” as Frank Jackson would say).

This thought experiment is supposed to prove the impossibility of what Searle calls “strong AI,” the thesis that “the appropriately programmed digital computer with the right inputs and outputs would thereby have a mind in exactly the sense that human beings have minds” (Searle, 1986a). There has been a huge outpouring of reaction to Searle’s many versions of this thought experiment over the last dec-

¹. Try to imagine the state of mind of Ji Hu-Min, my graduate student from Beijing, whose introduction to Anglo-American philosophy of mind (while his English was still quite rudimentary) was sitting in a seminar where students and professors vigorously debated what would be the case if the entire population of China were somehow forced to participate in a massive realization of a putatively conscious AI program (Block’s example) and then went on to discuss, with equal obliviousness to the sensitivities of a Chinese observer, Searle’s Chinese Room.
telling, in which we suspend our disbelief in magic, and so forth. By the way, we could imagine a somewhat labored continuation in which the Irishman turned out to be “right” in his second wish after all — perhaps he’s planning to throw a big party, and one glass won’t refill fast enough to satisfy all his thirsty guests (and it’s no use saving it up in advance — we all know how stale stout loses its taste). We tend not to think of such complications, which is part of the explanation of why jokes work. Is that enough?

This conversation is not dazzling, but let’s suppose it was good enough to fool the judge. Now we are invited to imagine all these speeches by CR being composed by the giant program Searle is diligently hand-simulating. Hard to imagine? Of course, but since Searle stipulates that the program passes the Turing test, and since this level of conversational sophistication would surely be within its powers, unless we try to imagine the complexities of a program capable of generating this sort of conversation, we are not following directions. Of course we should also imagine that Searle hasn’t any inkling of what he is doing in the Chinese Room; he just sees zeros and ones that he manipulates according to the program. It is important, by the way, that Searle invites us to imagine that he manipulates inscrutable Chinese characters instead of zeros and ones, for this may lull us into the [unwarranted] supposition that the giant program would work by somehow simply “matching up” the input Chinese characters with some output Chinese characters. No such program would work, of course — do CR’s speeches in English “match up” with the judge’s questions?

A program that could actually generate CR’s speeches in response to J’s questions might look something like this in action (viewed from the virtual-machine level, not from Searle’s ground-floor level). On parsing the first words, “Did you hear about…” some of the program’s joke-detecting demons were activated, which called up a host of strategies for dealing with fiction, “second intention” language, and the like, so when the words “magic lamp” came to be parsed, the program had already put a low priority on responses complaining that there were no such things as magic lamps. A variety of standard genie-joke narrative frames (Minsky, 1975) or scripts (Schank and Abelson, 1977) were activated, creating various expectations for continuations, but these were short-circuited, in effect, by the punch line, which invoked a more mundane script (the script for “asking for seconds”), and the unexpectedness of this was not lost on the program. . . . At the same
time, demons sensitive to the negative connotations of ethnic-joke-telling were also alerted, eventually leading to the second theme of CR's first response. . . And so forth, in vastly more detail than I have tried to sketch here.

That fact is that any program that could actually hold up its end in the conversation depicted would have to be an extraordinarily supple, sophisticated, and multilayered system, brimming with "world knowledge" and meta-knowledge and meta-meta-knowledge about its own responses, the likely responses of its interlocutor, its own "motivations" and the motivations of its interlocutor, and much, much more. Searle does not deny that programs can have all this structure, of course. He simply discourages us from attending to it. But if we are to do a good job imagining the case, we are not only entitled but obliged to imagine that the program Searle is hand-simulating has all this structure — and more, if only we can imagine it. But then it is no longer obvious, I trust, that there is no genuine understanding of the joke going on. Maybe the billions of actions of all those highly structured parts produce genuine understanding in the system after all. If your response to this hypothesis is that you haven't the faintest idea whether there would be genuine understanding in such a complex system, that is already enough to show that Searle's thought experiment depends, illicitly, on your imagining too simple a case, an irrelevant case, and drawing the "obvious" conclusion from it.

Here is how the misdirection occurs. We see clearly enough that if there were understanding in such a giant system, it would not be Searle's understanding (since he is just a cog in the machinery, oblivious to the context of what he is doing). We also see clearly that there is nothing remotely like genuine understanding in any hunk of programming small enough to imagine readily — whatever it is, it's just a mindless routine for transforming symbol strings into other symbol strings according to some mechanical or syntactical recipe. Then comes the suppressed premise: Surely more of the same, no matter how much more, would never add up to genuine understanding. But why should anyone think this was true? Cartesian dualists would think so, because they think that even human brains are unable to accomplish understanding all by themselves; according to the Cartesian view, it takes an immaterial soul to pull off the miracle of understanding. If, on the other hand, we are materialists who are convinced that one way or another our brains are responsible on their own, without miraculous assistance, for our understanding, we must admit that genuine understanding is somehow achieved by a process composed of interactions between a host of subsystems none of which understand a thing by themselves. The argument that begins "this little bit of brain activity doesn't understand Chinese, and neither does this bigger bit of which it is a part . . . " is headed for the unwanted conclusion that even the activity of the whole brain is insufficient to account for understanding Chinese. It is hard to imagine how "just more of the same" could add up to understanding, but we have very good reason to believe that it does, so in this case, we should try harder, not give up.

How might we try harder? With the help of some handy concepts: the intermediate-level software concepts that were designed by computer scientists precisely to help us keep track of otherwise unimaginable complexities in large systems. At the intermediate levels we see many entities that are quite invisible at more microscopic levels, such as the "demons" alluded to above, to which a modicum of quasi-understanding is attributed. Then it becomes not so difficult to imagine how "more of the same" could amount to genuine understanding. All these demons and other entities are organized into a huge system, the activities of which organize themselves around its own Center of Narrative Gravity. Searle, laboring in the Chinese Room, does not understand Chinese, but he is not alone in the room. There is also the System, CR, and it is to that self that we should attribute any understanding of the joke.

This reply to Searle's example is what he calls the Systems Reply. It has been the standard reply of people in AI from the earliest outings of his thought experiment, more than a decade ago, but it is seldom appreciated by people outside of AI. Why not? Probably because they haven't learned how to imagine such a system. They just can't imagine how understanding could be a property that emerges from lots of distributed quasi-understanding in a large system. They certainly can't if they don't try, but how could they be helped along on this difficult exercise? Is it "cheating" to think of the software as composed of homunculi who quasi-understand, or is that just the right crutch to help the imagination make sense of astronomical complexity? Searle begs the question. He invites us to imagine that the giant program consists of some simple table-lookup architecture that directly matches Chinese character strings to others, as if such a program could stand in, fairly, for any program at all. We have no business imagining such a simple program and assuming that it is the program Searle is simulating, since no such program could produce the sorts of results that would pass the Turing test, as advertised. (For a similar move and its rebuttal, see Block, 1982; and Dennett, 1985).
Complexity does matter. If it didn’t, there would be a much shorter argument against strong AI: “Hey, look at this hand calculator. It doesn’t understand Chinese, and any conceivable computer is just a giant hand calculator, so no computer could understand Chinese. Q.E.D.” When we factor in the complexity, as we must, we really have to factor it in — and not just pretend to factor it in. That is hard to do, but until we do, any intuitions we have about what is “obviously” not present are not to be trusted. Like Frank Jackson’s case of Mary the color scientist, Searle’s thought experiment yields a strong, clear conviction only when we fail to follow instructions. These intuition pumps are defective; they do not enhance but mislead our imaginations.

But what, then, of my own intuition pumps? What of Shakey the robot, or the CADBLIND Mark II, or the biofeedback-trained blindsight patient, for instance? Are they not equally suspect, equally guilty of misleading the reader? I’ve certainly done my best in telling these tales to lead your imagination down certain paths, and to keep you from bogging down in complexities I deemed unnecessary to the point I was attempting to make. There is some asymmetry, however: My intuition pumps are, for the most part, intended to help you imagine new possibilities, not convince you that certain prospects are impossible. There are exceptions. My variation on the brain in the vat that opened the book was designed to impress on you the impossibility of certain sorts of deception, and some of the thought experiments in chapter 5 were intended to show that, unless there were a Cartesian Theater, there could not be a fact of the matter distinguishing Orwellian from Stalin-esque content revisions. These thought experiments proceeded, however, by heightening the vividness for the “opposition”; the examples of the woman in the hat at the party and the long-haired woman with glasses, for instance, were designed to sharpen the very intuition I then sought to discredit by argument.

Still, let the reader beware: My intuition pumps, like anyone else’s, are not the straightforward demonstrations they may seem to be; they are more art than science. (For further warnings about philosophers’ thought experiments, see Wilkes, 1988.) If they help us conceive of new possibilities, which we can then confirm by more systematic methods, that is an achievement; if they lure us down the primrose path, that is a pity. Even good tools can be misused, and like any other workers, we will do better if we understand how our tools work.