# The Chinese Room: Just Say "No!"

To appear in the *Proceedings of the 22nd Annual Cognitive Science Society Conference*, (2000), NJ: LEA

Robert M. French
Quantitative Psychology and Cognitive Science
University of Liège
4000 Liège, Belgium
email: rfrench@ulg.ac.be

**Abstract**

It is time to view John Searle's Chinese Room thought experiment in a new light. The main focus of attention has always been on showing what is wrong (or right) with the argument, with the tacit assumption being that somehow there could be such a Room. In this article I argue that the debate should not focus on the question "If a person in the Room answered all the questions in perfect Chinese, while not understanding a word of Chinese, what would the implications of this be for strong AI?" Rather, the question should be, "Does the very idea of such a Room and a person in the Room who is able to answer questions in perfect Chinese while not understanding any Chinese make any sense at all?" And I believe that the answer, in parallel with recent arguments that claim that it would be impossible for a machine to pass the Turing Test unless it had experienced the world as we humans have, is no.

## Introduction

Alan Turing's (1950) classic article on the Imitation Game provided an elegant operational definition of intelligence. His article is now exactly fifty years old and ranks, without question, as one of the most important scientific/philosophical papers of the twentieth century. The essence of the test proposed by Turing was that the ability to perfectly simulate unrestricted human conversation would constitute a sufficient criterion for intelligence. This way of defining intelligence, for better or for worse, was largely adopted as of the mid-1950's, implicitly if not explicitly, as the overarching goal of the nascent field of artificial intelligence (AI).

Thirty years after Turing's article appeared, John Searle (1980) put a new spin on Turing's original arguments. He developed a thought experiment, now called "The Chinese Room," which was a reformulation of Turing's original test and, in so doing, produced what is undoubtedly the second most widely read and hotly discussed paper in artificial intelligence. While Turing was optimistic about the possibility of creating intelligent programs in the foreseeable future, Searle concluded his article on precisely the opposite note: "...no [computer] program, by itself, is sufficient for intentionality." In short, Searle purported to have shown that real (human-like) intelligence was impossible for any program implemented on a computer. In the present article I will begin by briefly presenting Searle's well-known transformation of the Turing's Test. Unlike other critics of the Chinese Room argument, however, I will not take issue with Searle's argument per se. Rather, I will focus on the argument's central premise and will argue that the correct approach to the whole argument is simply to refuse to go beyond this premise, for it is, as I hope to show, untenable.

## The Chinese Room

Instead of Turing's Imitation Game in which a computer in one room and a person in a separate room both attempt to convince an interrogator that they are human, Searle asks us to begin by imagining a closed room in which there is an English-speaker who knows no Chinese whatsoever. This room is full of symbolic rules specifying inputs and outputs, but, importantly, there are no translations in English to indicate to the person in the room the meaning of any Chinese symbol or string of symbols. A native Chinese person outside the room writes questions — *any questions* — in Chinese on a piece of paper and sends them into the room. The English-speaker receives each question inside the Room then matches the symbols in the question with symbols in the rule-base. (This does not have to be a direct table matching of the string of symbols in the question with symbols in the rule base, but can include any type of look-up program, regardless of its structural complexity.) The English-speaker is blindly led through the maze of rules to a string of symbols that constitutes an answer to the question. He copies this answer on a piece of paper and sends it out of the room. The Chinese person on the outside of the room would see a perfect response, even though the English-speaker understood no Chinese whatsoever. The Chinese person would therefore be fooled into believing that the person inside the room understood perfect Chinese.

Searle then compares the person in the room to a computer program and the symbolic rules that fill the room to the knowledge databases used by the computer program. In Searle's thought experiment the person who is answering the questions in perfect written Chinese still has no knowledge of Chinese. Searle then applies the conclusion of his thought experiment to the general question of machine intelligence. He concludes that a computer program, however perfectly it managed to communicate in writing, thereby fooling all human

questioners, would still not understand what it was writing, any more than the person in the Chinese Room understood any Chinese. Ergo, computer programs capable of true understanding are impossible.

## Searle's Central Premise

But this reasoning is based on a central premise that needs close scrutiny.

Let us begin with a simple example. If someone began a line of reasoning thus: "Just for the sake of argument, let's assume that cows are as big as the moon," you would most likely reply, "Stop right there, I'm not interested in hearing the rest of your argument because cows are demonstrably NOT as big as the moon." You would be justified in not allowing the person to continue to his conclusions because, as logical as any of his subsequent reasoning might be, any conclusion arising from his absurd premise would be unjustified.

Now let us consider the central premise on which Searle's argument hangs — namely, that there could be such a thing as a "Chinese Room" in which an English-only person *could actually fool* a native-Chinese questioner. I hope to show that this premise is no more plausible than the existence of lunar-sized cows and, as a result, we have no business allowing ourselves to be drawn into the rest of Searle's argument, any more than when we were asked to accept that all cows were the size of the moon.

Ironically, the arguments in the present paper *support* Searle's point that symbolic AI is not sufficient to produce human-like intelligence, but do so not by comparing the person in the Chinese Room to a computer program, but rather by showing that the Chinese Room itself would be an impossibility for a symbol-based AI paradigm.

## Subcognitive Questioning and the Turing Test

To understand why such a Room would be impossible, which would mean that the person in the Room could never fool the outside-the-Room questioner, we must look at an argument concerning the Turing Test first put forward by French (1988, 1990, 2000). French's claim is that no machine that had not experienced life as we humans had could ever hope to pass the Turing Test. His demonstration involves showing just how hard it would be for a computer to consistently reply in a human-like manner to what he called "subcognitive" questions. Since Searle's Chinese Room argument is simply a reformulation of the Turing Test, we would expect to be able to apply these arguments to the Chinese Room as well, something which we will do this later in this paper.

It is important to spend a moment reviewing the nature and the power of "subcognitive" questions. These are questions that are explicitly designed to provide a window on low-level (i.e., unconscious) cognitive or physical structure. By "low-level cognitive structure", we mean the subconscious associative network in human minds that consists of highly overlapping activatable representations of experience (French, 1990). Creating these questions and, especially, gathering the answers to them require a bit of preparation on the part of the Interrogator who will be administering the Turing Test.

The Interrogator in the Turing Test (or the Questioner in the Chinese Room) begins by preparing a long list of these questions — the Subcognitive Question List. To get answers to these questions, she ventures out into an English-language population and selects a representative sample of individuals from that population. She asks each person surveyed all the questions on her Subcognitive Question List and records their answers. The questions along with the statistical range of answers to these questions will be the basis for her Human Subcognitive Profile. Here are some of the questions on her list (French, 1988, 1990).

<u>Questions using neologisms</u>:

"On a scale of 0 (completely implausible) to 10 (completely plausible):

- Rate *Flugblogs* as a name Kellogg's would give to a new breakfast cereal.
- Rate *Flugblogs* as the name of start-up computer company
- Rate *Flugblogs* as the name of big, air-filled bags worn on the feet and used to walk across swamps.
- Rate *Flugly* as the name a child might give to a favorite teddy bear.
- Rate *Flugly* as the surname of a bank accountant in a W. C. Fields movie.
- Rate *Flugly* as the surname of a glamorous female movie star.

"Would you like it if someone called you a *trubhead*? (0= not at all, ..., 10 = very much)"

"Which word do you find prettier: *blutch* or *farfaletta?*"

Note that the words *flugblogs, flugly, trubhead, blutch* and *farfaletta* are made-up. They will not be found in any dictionary and, yet, because of the uncountable influences, experiences and associations of a lifetime of hearing and using English, we are able to make judgments about these neologisms. And, most importantly, while these judgments may vary between individuals, their variation is not random. For example, the average rating of *Flugly* as the surname of a glamorous actress will most certainly fall below the average rating of *Flugly* as the name for a child's teddy bear. Why? Because English speakers, all of us, have

grown up surrounded by roughly the same sea of sounds and associations that have gradually formed our impressions of the prettiness (or ugliness) of particular words or sounds. And while not all of these associations are identical, of course, they are similar enough to be able to make predictions about how, on average, English-speaking people will react to certain words and sounds. This is precisely why Hollywood movie moguls gave the name "Cary Grant" to a suave and handsome actor born "Archibald Alexander Leach" and why "Henry Deutschendorf, Jr." was re-baptised "John Denver."

Questions using categories:

- Rate *banana splits* as *medicine*.
- Rate *purses* as *weapons*.
- Rate *pens* as *weapons*.
- Rate *dry leaves* as *hiding places*.

No dictionary definition of "dry leaves" will include in its definition "hiding place," and, yet, everyone who was ever a child where trees shed their leaves in the fall knows that that piles of dry leaves make wonderful hiding places. But how could this information, and an infinite amount of information just like it that is based on our having experienced the world in a particular way, ever be explicitly programmed into a computer?

Questions relying on human physical sensations:

- Does holding a gulp of Coca-Cola in your mouth feel more like having pins-and-needles in your foot or having cold water poured on your head?
- Put your palms together, fingers outstretched and pressed together. Fold down your two middle fingers till the middle knuckles touch. Move the other four pairs of fingers. What happens to your other fingers? (Try it!)

We can imagine many more questions that would be designed to test not only for subcognitive associations, but for internal physical structure. These would include questions whose answers would arise, for example, from the spacing of a human's eyes, would be the results of little self-experiments involving tactile sensations on their bodies or sensations after running in place, and so on.

People's answers to subcognitive questions are the product of a lifetime of experiencing the world with our human bodies, our human behaviors (whether culturally or genetically engendered), our human desires and needs, etc. (See Harnard (1989) for a discussion of the closely related *symbol grounding problem*.)

I have asked people the question about Coca-Cola and pins-and-needles many times and they overwhelmingly respond that holding a soft-drink in their mouth feels more like having pins and needles in their foot than having cold water poured on them. Answering this question is dead easy for people who have a head and mouth, have drunk soft-drinks, have had cold water poured on their head, and have feet that occasionally fall asleep. But think of what it would take for a machine that had none of these to answer this question. How could the answer to this question be explicitly programmed into the machine? Perhaps (after reading this article) a programmer could put the question explicitly into the machine's database, but there are literally infinitely many questions of this sort and to program them all in would be impossible. A program that could answer questions like these in a human-like enough manner to pass a Turing Test would have had to have experienced the world in a way that was very similar to the way in which we had experienced the world. This would mean, among many other things, that it would have to have a body very much like ours with hands like ours, with eyes where we had eyes, etc. For example, if an otherwise perfectly intelligent robot had its eyes on its knees, this would result in detectably non-human associations for such activities as, say, praying in church, falling when riding a bicycle, playing soccer, or wearing pants.

The moral of the story is that it doesn't matter if we humans are confronted with made-up words or conceptual juxtapositions that never normally occur (e.g., *dry leaves* and *hiding place*), we can still respond and, moreover, our responses will show statistical regularities over the population. Thus, by surveying the population at large with an extensive set of these questions, we draw up a Human Subcognitive Profile for the population. It is precisely this subcognitive profile that could not be reproduced by a machine that had not experienced the world as the members of the sampled human population had. The Subcognitive Question List that was used to produce the Human Subcognitive Profile gives the well-prepared Interrogator a sure-fire tool for eliminating machines from a Turing test in which humans are also participating. The Interrogator would come to the Turing Test and ask both candidates the questions on her Subcognitive Question List. The candidate most closely matching the average answer profile from the human population will be the human.

## The English Room

Now let us see how this technique can be gainfully applied to Searle's Chinese Room thought experiment. We will start by modifying Searle's original *Gedankenexperiment* by switching the languages around. This, of course, has no real bearing on the argument itself, but it will make our argument easier to follow. We will assume that inside the Room there is a Chinese person (let's call him Wu) who understands not a word of written English and outside the Room is a native speaker/writer of English (Sue). Sue sends into the Room questions written in English and Wu must produce the answers to these questions in English.

Now, it turns out that Sue is not your average naive questioner, but has read many articles on the Turing Test, knows about subcognitive questions and is thoroughly familiar with John Searle's argument. She also suspects that the person inside the (English) Room might not actually be able to read English and she sets out to prove her hunch.

Sue will not only send into the Room questions like, "What is the capital of Cambodia?", "Who painted *The Mona Lisa*?" or "Can fleas fly?" but will also ask a large number of "subcognitive questions." Because the Room, like the computer in the Turing Test, had not experienced the world as we had and because it would be impossible to explicitly write down all of the rules necessary to answer subcognitive questions in general, the answers to the full range of subcognitive questions could not be contained in the lists of symbolic rules in the Room. Consequently, the person in the Room would be revealed not to speak English for exactly the same reason that the machine in the Turing Test would be revealed not to be a person.

Take the simple example of non existent words like *blutch* or *trubhead*. These words are neologisms and would certainly be nowhere to be found in the symbolic rules in the English Room. Somehow, the Room would have to contain, in some symbolic form, information not only about all words, but also non-words as well. But the Room, if it is to be compared with a real computer, cannot be infinitely large, nor can we assume infinite fast search of the rule base (see Hofstadter & Dennett, 1981, for a discussion of this point). So, we have two closely related problems: First, and most crucially, *how* could the rules have gotten into the Room in the first place (a point that Searle simply ignores)? And secondly, the number of explicit symbolic rules would require essentially an infinite amount of space. And while rooms in thought experiments can perhaps be infinitely large, the computers that they are compared to cannot be.

In other words, the moral of the story here, as it was for the machine trying to pass the Turing Test, is that no matter how many symbolic rules were in the English Room they would not be sufficient for someone who did not understand written English to fool a determined English questioner. And this is where the story should rightfully end. Searle has no business taking his argument any further — and, ironically, *he doesn't need to*, since the necessary inadequacy of an such a Room, regardless of how many symbolic rules it contains, proves his point about the impossibility of achieving artificial intelligence in a traditional symbol-based framework. So, when Searle asks us to accept that the English-only human in his Chinese Room could reply in perfect written Chinese to questions written in Chinese, we must say, "That's strictly impossible, so stop right there."

## Shift in Perception of the Turing Test

Let us once again return to the Turing Test to better understand the present argument.

It is easy to forget just how high the optimism once ran for the rapid achievement of artificial intelligence. In 1958 when computers were still in their infancy and even high-level programming languages had only just been invented, Simon and Newell, two of the founders of the field of artificial intelligence, wrote, "...there are now in the world machines that think, that learn and that create. Moreover, their ability to do these things is going to increase rapidly until – in a visible future – the range of problems they can handle will be coextensive with the range to which the human mind has been applied." (Simon & Newell, 1958). Marvin Minsky, head of the MIT AI Laboratory, wrote in 1967, "Within a generation the problem of creating 'artificial intelligence' will be substantially solved" (Minsky, 1967).

During this period of initial optimism, the vast majority of the authors writing about the Turing Test tacitly accepted Turing's premise that a machine might actually be able to be built that could pass the Test in the foreseeable future. The debate in the early days of AI, therefore, centered almost exclusively around the validity of Turing's operational definition of intelligence — namely, did passing the Turing Test constitute a sufficient condition for intelligence or did it not? But researchers' views on the possibility of achieving artificial intelligence shifted radically between the mid-1960's and the early 1980's. By 1982, for example, Minsky's position regarding achieving artificial intelligence had undergone a radical shift from one of unbounded optimism 15 years earlier to a far more sober assessment of the situation: "The AI problem is one of the hardest ever undertaken by science" (Kolata, 1982). The perception of the Turing Test underwent a parallel shift. At least in part because of the great difficulties being experienced by AI, there was a growing realization of just how hard it would be for a machine to ever pass the Turing Test. Thus, instead of discussing whether or not a machine that had passed the Turing Test was really intelligent, the discussion shifted to the question of whether it would even be possible for any machine to pass such a test (Dennett, 1985; French, 1988, 1990; Crockett 1994; Harnad, 1989; for a review, see French, 2000).

## The Need for a Corresponding Shift in the Perception of the Chinese Room

A shift in emphasis identical to the one that has occurred for the Turing Test is now needed for Searle's Chinese Room thought experiment. Searle's article was published in pre-connectionist 1980, when traditional symbolic AI was still the dominant paradigm in the field. Many of the major difficulties facing symbolic AI had come to light, but in 1980 there was still little emphasis on the "sub-symbolic" side of things.

But the growing difficulties that symbolic AI had in dealing with "sub-symbolic cognition" were responsible, at least in part, for the widespread appeal of the connectionist movement of the mid-1980's. While several of the commentaries of Searle's original article (Searle, 1980) briefly touch on the difficulties involved in actually creating a Chinese Room, none of them focus outright on the impossibility of the Chinese Room as described by Searle and reject the rest of the argument because of its impossible premise. But this rejection corresponds precisely to rejecting the idea that a machine (that had not experienced the world as we humans have) could ever pass the Turing Test, an idea that many people now accept. We are arguing for a parallel shift in emphasis for the Chinese Room *Gedankenexperiment*.

## Can the "Robot Reply" Help?

It is necessary to explore for a moment the possibility that one could somehow fill the Chinese Room with all of the appropriate rules that would allow the non-Chinese-reading person to fool a non-holds-barred Chinese questioner. Where could rules come from that would allow the person in the Chinese Room to answer all of the in-coming questions in Chinese perfectly? One possible reply is a version of the Robot Reply (Searle, 1980). Since the rules couldn't have been symbolic and couldn't have been explicitly programmed in for the reasons outlined above (also see French, 1988, 1990), perhaps they could have been the product of a Robot that had experienced and interacted with the world as we humans would have, all the while generating rules that would be put in the Chinese Room.

This is much closer to what would be required to have the appropriate "rules," but still leaves open the question of how you could ever come up with such a Robot. The Robot would have to be able to interact seamlessly with the world, exactly as a Chinese person would, in order to have been able to produce all the "rules" (high-level and subcognitive) that would later allow the person in the Room to fool the Well-Prepared Questioner. But then we are back to square one, for creating such a robot amounts to creating a robot that would pass the Turing Test.

## The Chinese Room: a Simple Refutation

It must be reiterated that when Searle is attacking the "strong AI" claim that machines processing strings of symbols are capable of doing what we humans call thinking, he is explicitly talking about programs implemented on *computers*. It is important not to ignore the fact, as some authors unfortunately have (e.g., Block, 1981), that computers are *real machines* of finite size and speed; they have neither infinite storage capacity nor infinite processing speed.

Now consider the standard Chinese Room, i.e., the one in which the person inside the Room has no knowledge of Chinese and the Questioner outside the Room is Chinese. Now assume that the last character of the following question is distorted in an extremely phallic way, but in a way that nonetheless leaves the character completely readable to any reader of Chinese: "Would the last character of this sentence embarrass a very shy young woman?" In order to answer this question correctly — a trivially easy task for anyone who actually reads Chinese — the Chinese Room would have to contain rules that would not only allow the person to respond perfectly to all strings of Chinese characters that formed comprehensible questions, but also to the infinitely many possible legible *distortions* of those strings of characters. Combinatorial explosion brings the house down around the Chinese Room. (Remember, we are talking about real computers that can store a finite amount information and must retrieve it in a finite amount of time.)

One might be tempted to reply, "The solution is to eliminate all distortions. Only standard fonts of Chinese characters are permitted." But, of course, there are hundreds, probably thousands, of different fonts of characters in Chinese (Hofstadter, 1985) and it is completely unclear what would constitute "standard fonts." In any event, one can sidestep even this problem.

Consider an equivalent situation in English. It makes perfect sense to ask, "Which letter could be most easily distorted to look like a cloud: an 'O' or an 'X'?" An overwhelming majority of people would, of course, reply "O", even though clouds, superficially and theoretically, have virtually nothing in common with the letter "O". But how could the symbolic rules in Searle's Room possibly serve to answer this perfectly legitimate question? A theory of clouds contained in the rules certainly wouldn't be of any help, because that would be about storms, wind, rain and meteorology. A theory or database of cloud forms would be of scant help either, since clouds are anything but two dimensional, much less round. Perhaps only if the machine/Room had grown up scrawling vaguely circular shapes on paper and calling them clouds in kindergarten and elementary school, then maybe it would be able to answer this question. But short of having had that experience, I see little hope of an a priori theory of correspondence between clouds and letters that would be of any help.

## Conclusion

The time has come to view John Searle's Chinese Room thought experiment in a new light. Up until now, the main focus of attention has been on showing what is wrong (or right) with the argument, with the tacit assumption being that somehow there could be such a Room. This parallels the first forty years of discussions

on the Turing Test, where virtually all discussion centered on the sufficiency of the Test as a criterion for machine intelligence, rather than whether any machine could ever actually pass it. However, as the overwhelming difficulties of AI gradually became apparent, the debate on the Turing Test shifted to whether or not any machine that had not experience the world as we had could ever actually pass the Turing Test. It is time for an equivalent shift in attention for Searle's Chinese Room. The question should not be, "If a person in the Room answered all the questions in perfect Chinese, while not understanding a word of Chinese, what would the implications of this be for strong AI?"" Rather, the question should be, "Does the very idea of such a Room and a person actually be able to answer questions in perfect Chinese while not understanding any Chinese make any sense at all?" And I believe that the answer, in parallel with the impossibility of a machine passing the Turing Test, is no.

## Acknowledgments

The present paper was supported in part by research grant IUAP P4/19 from the Belgian government.

## References

Block, N. (1981) Psychologism and behaviourism. *Philosophical Review, 90*, 5-43

Crockett, L. (1994) *The Turing Test and the Frame Problem: AI's Mistaken Understanding of Intelligence*. Ablex

Davidson, D. (1990) Turing's test. In Karim A. Said et al. (eds.), *Modelling the Mind*. Oxford University Press, 1-11.

Dennett, D. (1985) Can machines think? In *How We Know*. (ed.) M. Shafto. Harper & Row

French, R. M. (1988). Subcognitive Probing: Hard Questions for the Turing Test. *Proceedings of the Tenth Annual Cognitive Science Society Conference*, Hillsdale, NJ: LEA. 361-367.

French, R. M. (1990). Subcognition and the Limits of the Turing Test. *Mind, 99*(393), 53-65. Reprinted in: P. Millican & A. Clark (eds.). *Machines and Thought: The Legacy of Alan Turing* Oxford, UK: Clarendon Press, 1996.

French, R. M. (2000). Peeking Behind the Screen: The Unsuspected Power of the Standard Turing Test. *Journal of Experimental and Theoretical Artificial Intelligence*. (in press).

French, R. M. (2000). The Turing Test: The First Fifty Years. *Trends in Cognitive Sciences, 4*(3), 115-122.

Harnad, S. (1989) Minds, machines and Searle. *Journal of Experimental and Theoretical Artificial Intelligence, 1*, 5-25

Hofstadter, D. (1985). Variations on a Theme as the Crux of Creativity. In *Metamagical Themas*. New York, NY: Basic Books. p. 244.

Hofstatder, D. & Dennett, D. (1981). *The Mind's I*. New York, NY: Basic Books.

Kolata, G. (1982) How can computers get common sense? *Science, 217*, p. 1237

Minsky, M. (1967) *Computation: Finite and Infinite Machines*. Prentice-Hall, p. 2

Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3, 414-424.

Simon, H. and Newell, A. (1958) Heuristic problem solving: The next advance in operations research. *Operations Research, 6*