The Chinese Room

An argument by John R. Searle
against Strong Artificial Intelligence

Strong AI and Weak AI

The philosopher of the mind John R. Searle makes a distinction between two main approaches to Artificial Intelligence:

Weak AI: It is possible to build a machine that will act *as though* it is intelligent.

Strong AI: It is possible to build a machine that will *actually* think and have a mind.

What Searle labels as Strong AI is in fact called **computer functionalism**, the view that the human brain is a biological computer and the mind is a computer program. In other words, *the mind is to the brain as the program is to the hardware*.

$$\frac{Mind}{Brain} = \frac{Program}{Hardware}$$

Logically, this implies that a correctly-built machine running the right computer program will in fact have a mind.

The Turing Test

In 1950, the British mathematician and logician **Alan Turing** proposed a simple way to tell whether or not a machine was really intelligent. If it was able to answer any question posed to it in such way that its interlocutor could not possibly tell its answer from a human being's, then the machine would *definitely* be intelligent. In Turing's words:

> *"If a machine acts as intelligently as human being, then it is as intelligent as a human being."*

As Turing saw it, just like we assume that our fellow human beings think and have minds judging by their behaviour as analogous to ours, the same rule should be applied when considering machines.

> *"Instead of arguing continually over this point, it is usual to have a polite convention that everyone thinks."*

In this respect, Turing had a behaviourist attitude very much in line with Strong AI.

Symbol processing

In 1963, **Alan Newell** and **Herbert Simon** proposed that the key to human and machine intelligence was symbol manipulation.

> *A physical symbol system has the necessary and sufficient means of general intelligent action.*

Necessary → The human mind must consist in symbol manipulation (otherwise it would not be intelligent).

Sufficient → A machine that operates with a symbol system can potentially be intelligent.

This statement can be understood as rephrasing the claim of computer functionalism.

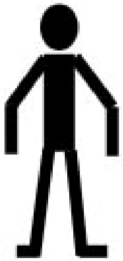
Searle’s Chinese Room

Suppose we built and programmed a machine that was actually successful in passing the Turing test. Would this really imply that our machine is intelligent?

Searle’s Chinese Room

Suppose we built and programmed a machine that was actually successful in passing the Turing test. Would this really imply that our machine is intelligent?

Imagine a man who didn’t speak a word of Chinese (this takes little effort).

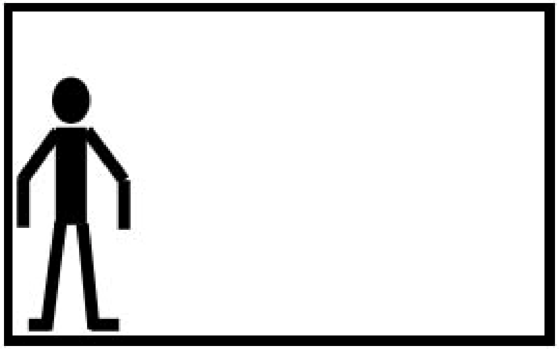


Searle’s Chinese Room

Suppose we built and programmed a machine that was actually successful in passing the Turing test. Would this really imply that our machine is intelligent?

Imagine a man who didn’t speak a word of Chinese (this takes little effort).

Lock him in a room…

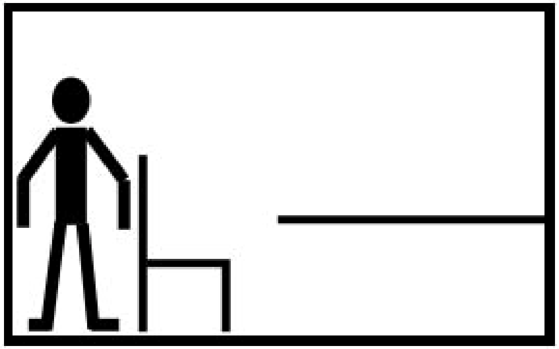


Searle’s Chinese Room

Suppose we built and programmed a machine that was actually successful in passing the Turing test. Would this really imply that our machine is intelligent?

Imagine a man who didn’t speak a word of Chinese (this takes little effort).

Lock him in a room…

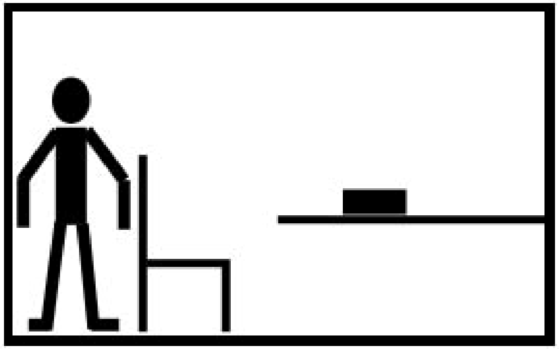


Searle's Chinese Room

Suppose we built and programmed a machine that was actually successful in passing the Turing test. Would this really imply that our machine is intelligent?

Imagine a man who didn't speak a word of Chinese (this takes little effort).

Lock him in a room with a very thorough rule book containing instructions on how to answer questions put to him in written Chinese (in effect, a computer program).



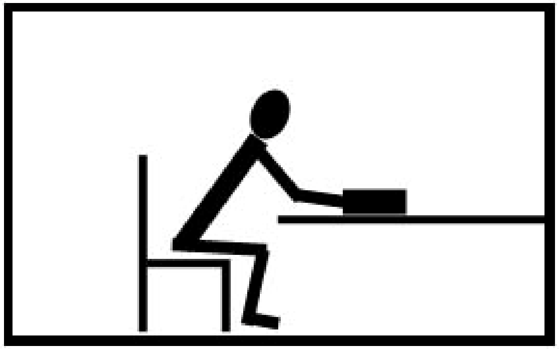
Searle’s Chinese Room

Suppose we built and programmed a machine that was actually successful in passing the Turing test. Would this really imply that our machine is intelligent?

Imagine a man who didn’t speak a word of Chinese (this takes little effort).

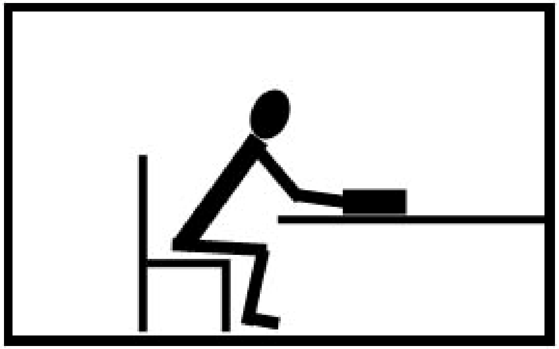
Lock him in a room with a very thorough rule book containing instructions on how to answer questions put to him in written Chinese (in effect, a computer program).

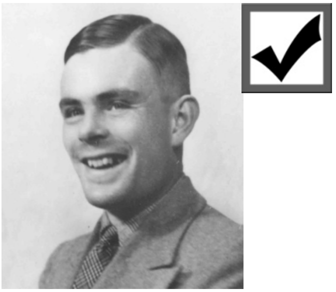
Send questions to him through a slot in the wall. He will look up the symbols and follow the appropriate instructions, writing down a few new symbols and sending them back to you through the slot.



Searle’s Chinese Room

If the rule book was properly made, you will interpret the result as a valid answer and our man will have passed the Turing test for understanding Chinese.

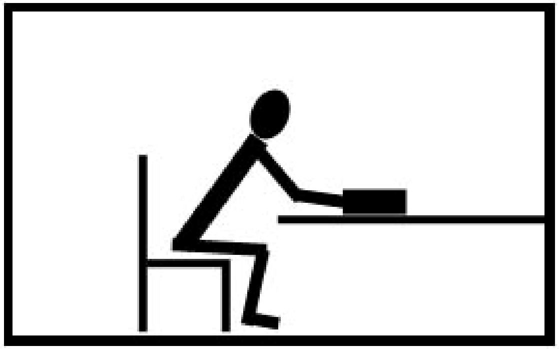


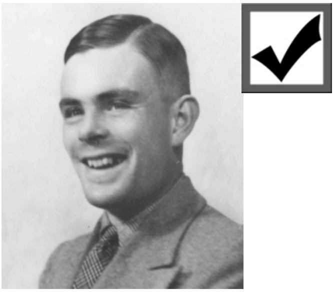


Searle’s Chinese Room

If the rule book was properly made, you will interpret the result as a valid answer and our man will have passed the Turing test for understanding Chinese.

But all the same, he does not understand a word of Chinese!





Searle’s Chinese Room

If the rule book was properly made, you will interpret the result as a valid answer and our man will have passed the Turing test for understanding Chinese.

But all the same, he does not understand a word of Chinese!

Therefore, neither could any computer, or formal system, understand Chinese on the basis on implementing the right program, no matter how successfully it may pass the Turing test.

☒

☑

Searle’s point

What Searle means by this argument is that there is more to the human mind that mere syntactical manipulations of meaningless symbols.

The key to understanding Chinese – or anything – lies beyond the reach of a formal system, because formal systems leave out ***meanings***. They are purely **syntactical**, but they lack the **semantics**. The human mind attaches meanings to the symbols.

Therefore, a formal system is an insufficient basis for intelligence.

What’s more…

Searle even goes further in developing his argument.

He points out that syntax and computation themselves are not even *intrinsic* to any machine. Rather, they are *observer relative*, in other words, they don’t exist objectively, but only “in the eye of the beholder”.

> “When I punch “2+2=” on my pocket calculator and it prints out “4” it knows nothing of computation, arithmetic or symbols, because it knows nothing about anything. Intrinsically, it is a complex electronic circuit that we *use* to compute with.”

You can never *discover* that the brain is a digital computer, because computation is not *discovered* in nature, it is *assigned* to it. In other words, computation in nature is *teleological*.

Paradoxically, according to Searle the only ***intrinsic digital computers*** are conscious agents thinking through computations, such as human beings.

Replies to the Chinese Room

Systems Reply

It is not the man that understands Chinese, but the whole *system*, including the man, the book and the room itself.

Searle answers that the reason the man doesn’t understand Chinese is that he has the syntax of Chinese, but not the semantics. The man has no way to get from syntax to semantics, and neither does the whole room.

Anyway, the man might just as well memorise the whole rule book and get rid of the room, and still he wouldn’t understand Chinese.

Replies to the Chinese Room

Robot Reply

If the “room” was placed inside a robot that could interact with its environment, a causal connection between the symbols and the objects they represent would be allowed.

As **Hans Moravec** puts it, “If we could graft a robot to a reasoning program, we wouldn't need a person to provide the meaning anymore: it would come from the physical world.”

Replies to the Chinese Room

Derived Meaning

The room *is* connected to the outer world through the Chinese interlocutor and through the programmers who designed the knowledge base in the rule book.

The symbols the man is manipulating are *already meaningful*, only not meaningful to *him*.

Searle complains that the symbols only have a derived meaning which is, again, *observer relative*. It depends on the conscious understanding of the Chinese speakers and the programmers outside, but the room itself has no understanding whatsoever.

Replies to the Chinese Room

Brain Simulator

Suppose the program in question was a simulation in fine detail of every neuron in a human brain. In this case, if it worked exactly like a virtual brain, wouldn’t it have a mind of its own and therefore be capable of understanding?

The point of this argument is that a formal system is a *language*, its purpose is not to constitute reality, but to describe it. This goes for describing minds too. Maybe the mind does not consist in symbol manipulations, but like any phenomenon it abides by certain rules, which apply at the level of neurons in the brain. These rules can be expressed formally, and thus the mind can be *simulated* by a computer program. Only, unlike other physical phenomena, a fine simulation of a mind would actually *be* a mind!

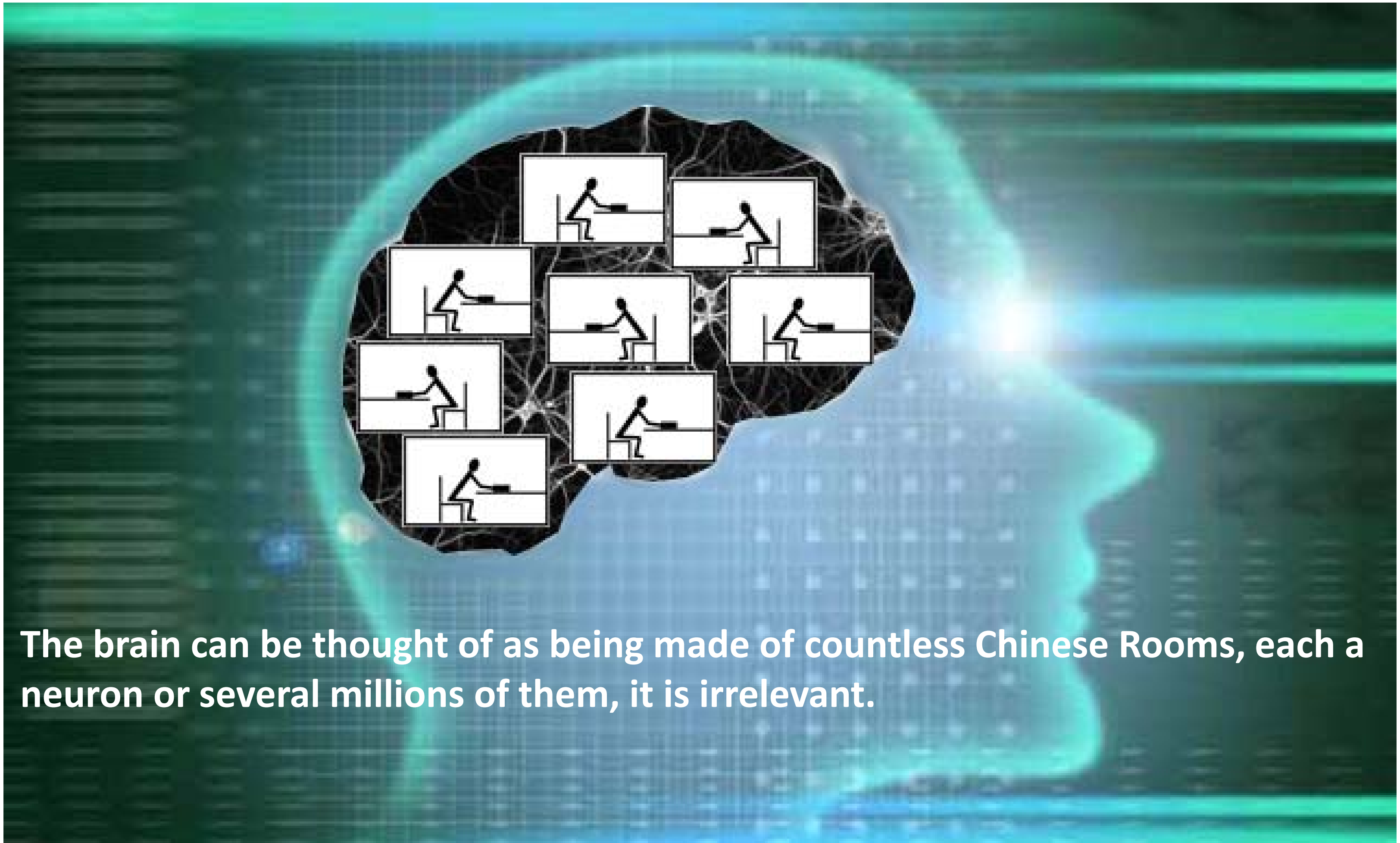
Conclusions

Searle’s Chinese Room Argument makes an important point: there is more to the mind than syntax, there has to be semantics *as well*.

However, this does not mean that the right computer programs (formal and syntactical as they may be) cannot generate a mind. Let’s see how this could be possible.

Conclusions

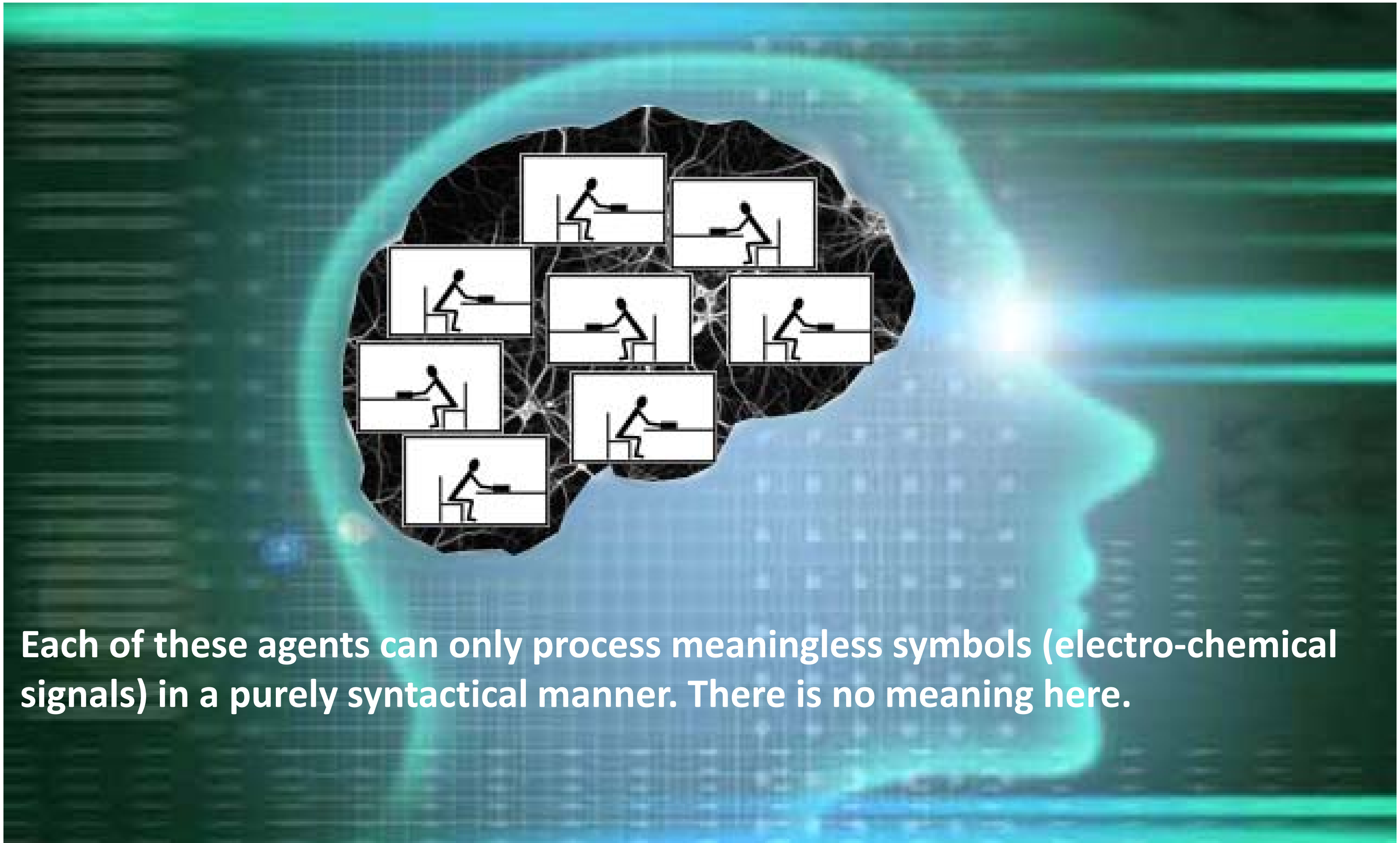
As a matter of fact, the Chinese Room is a very appropriate metaphor to describe how both the brain and the mind work. Consider this picture:



The brain can be thought of as being made of countless Chinese Rooms, each a neuron or several millions of them, it is irrelevant.

Conclusions

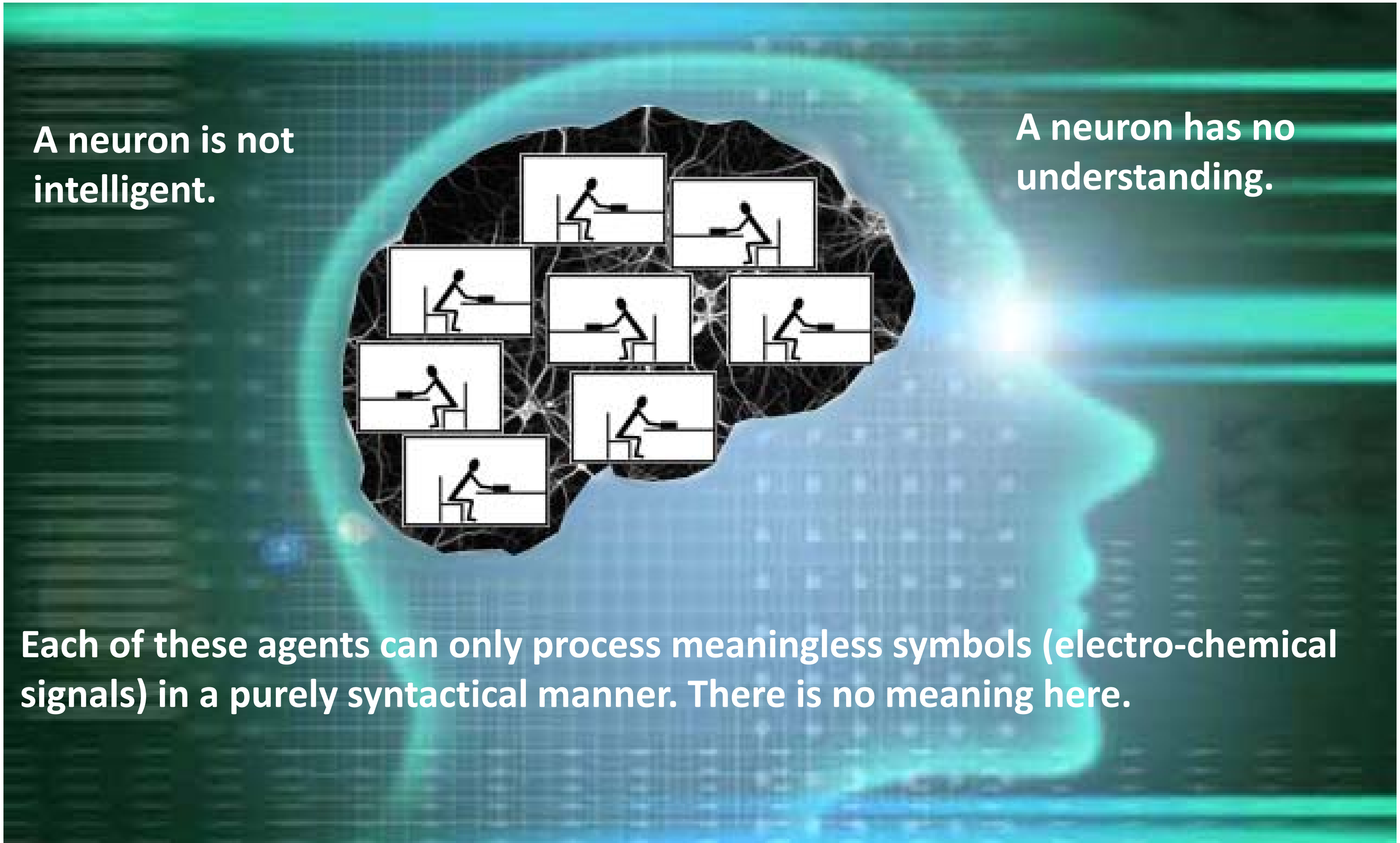
As a matter of fact, the Chinese Room is a very appropriate metaphor to describe how both the brain and the mind work. Consider this picture:



Each of these agents can only process meaningless symbols (electro-chemical signals) in a purely syntactical manner. There is no meaning here.

Conclusions

As a matter of fact, the Chinese Room is a very appropriate metaphor to describe how both the brain and the mind work. Consider this picture:



A neuron is not intelligent.

A neuron has no understanding.

Each of these agents can only process meaningless symbols (electro-chemical signals) in a purely syntactical manner. There is no meaning here.

Conclusions

However, the brain as a whole produces the mind, and the system does have *both* the syntax *and* the semantics.

So the question is how can semantics emerge out of the addition of purely syntactic agents (around one hundred billion of them, if it is neurons we are counting). Where do meanings appear?

The answer is very much related to the conception of the mind as an **agency** constituted by many lesser agents, explained by **Marvin Minsky** in *The Society of Mind*.

Intelligence can be achieved by joining non-intelligent elements, as long as they are connected in the right way. In other words, the key to intelligence is not only what the individual agents can do, but how they are *organised*. Intelligence lies in the **structure** within the system of agents.

Conclusions

Similarly, when we try to explain the origin of semantics out of purely syntactic elements, where can we look? Obviously, it *has* to be the connections.

Is this a plausible option? Indeed it is, with 7,000 synapses per neuron, an total estimate of 1,000,000,000,000,000 (one quadrillion) in a young human brain, it seems to be the only reasonable explanation for semantics in the mind.

Conclusions

Similarly, when we try to explain the origin of semantics out of purely syntactic elements, where can we look? Obviously, it *has* to be the connections.

Is this a plausible option? Indeed it is, with 7,000 synapses per neuron, an total estimate of 1,000,000,000,000,000 (one quadrillion) in a young human brain, it seems to be the only reasonable explanation for semantics in the mind.

Meaning:= the correct links between each complex symbol and the set of symbols to which it is related (some of them being the sensory signals stored in the memory), achieved through the connections between the agents that process them.