Can the Chinese Robot Think?

Dmytro Sepetyi — PhD, Assistant Professor
Zaporizhzhya State Medical University
(Zaporizhzhya, Ukraine)

E-mail: dmitry.sepety@gmail.com

In the paper, I discuss the Chinese room thought experiment, which was proposed by John Searle to show that executing by a computer of a program of data processing is not enough for genuine, i.e. understanding-based, thinking. The objection — advanced by Vadim Vasilyev — is also discussed that if the Chinese room is supplemented with devices to obtain external data, and so transformed into the Chinese robot, it may be attributed an understanding mind. Arguments are proposed that supplement those of Searle and give reason to turn down the objection. First, it is pointed out that the suggestion that the Chinese robot will have a specific program that has a quasi-semantic character and describes the relationship between linguistic signs and physical data is mistaken, because all the relations that programs assign are relations between units of inherently meaningless "data" — series of bits that have no inherent meaning but acquire their meaning only owing to conscious (human) interpreters. Second, it is argued that the robot's acquiring "knowledge" (data, as states of its memory cells) in the process of its interaction with the external world makes no relevant difference, because any such data can be just as well written in the same robot's memory cells from the very beginning.

Keywords: thinking, understanding, computer, algorithm, program, conscious mind, information, intentionality, semantics, epiphenomenal

Can a robot think? Or, to be more precise, can there be such a computer program that its execution by a computer will be genuine thinking, in the human sense of *understanding thinking*? Is execution of a program (algorithm) enough for thinking, in principle? Many philosophers of mind and cognitive scientists, as well as ordinary enthusiasts of computer technologies, believe that this is possible. John Searle advanced an influential argument to the contrary — the Chinese room thought experiment. If the argument is sound, then a computer that executes a program (algorithm) — no matter how perfect — can only imitate (simulate) thinking but cannot really think (understand). Opponents of the argument advanced several objections; Searle discussed the main objections and gave detailed and cogent replies. Recently, the Russian philosopher Vadim Vasilyev had "revived" one of these objections, which proposes to transform the Chinese room into the Chinese robot. Vasilyev advanced complementary arguments purported to show that Searle's reply was unsatisfactory. In this paper, I reconstruct the line of the discussion that proceeds from the Chinese room to the Chinese robot, and propose arguments that (I think) give good reasons to decline the Chinese robot counter-argument, taking into account Vasilyev's points.

The Chinese Room

The Chinese room thought experiment [Searle, 1980] has become a classical illumination of the principal difference between thinking and its computer imitation (Searle uses the

word “simulation”), that is, between the two kinds of what can be called “processing of information” − the one that is carried out *meaningfully, understanding-based* and the other that is carried out *automatically according to a certain algorithm (program)* — even if these processes are undistinguishable from the outside (by their input and output data).

Let us begin (following Searle) with outlining the problem-theoretical situation to which the experiment relates.

Suppose I am given a text to read. How can it be checked whether I have *understood* it or not? There is a proposition that this can be done by asking me such questions for which the text *does not contain direct answers*[1], but a person who has understood the meaning of the text and the question will easily guess what the correct answer should be. Consider, for example, the following story:

> A person has came into a restaurant and ordered a hamburger. He waited for a long time until he was brought a hamburger, which happened to be gravely burnt. He was enraged, shouted at the waiter and run out of the restaurant without paying.
>
> The question to answer is:
> Whether the person has eaten the hamburger?

I think that you will easily guess the right answer. Is it possible for a computer to give right answers to similar questions? If yes, would it mean that the computer *understands* the story and the question? The answer to the first question is “Yes”. The answer to the second is “No”.

How is it possible for a computer to give right answers to such questions? To see how, let us first think of what enables a person to answer correctly. It seems that he can answer so because he has some *additional information* about people and their behaviour in restaurants. However, if so, such information can be stored in computer memory, and a computer program can be developed to process it!

Let us suppose that scientists have developed a computer program called “Restaurant”. The program processes, besides the input data it gets from a user from the keyboard, the data from a database. The database contains (as a sequence of symbols, or electromagnetic states of the elements of computer memory) a computer “representation” of the information about restaurants and the behaviour of their clients that a person would use to give answers after listening to the story and the question like described above, — let us call it “the background data”. The program receives, as input, two sequences of symbols, a “story” and a “question”, then combines their symbols with symbols of the background data according to a very sophisticated system of rules (algorithm), and produces the result — “the answer”. Suppose that these answers coincide with those we would give. Does this mean that the computer *understands* the story and the question? John Searle contends that it does not, and explains this with the following thought experiment.

I do not know Chinese at all. I cannot distinguish the Chinese hieroglyphs from, say, the Japanese ones. Imagine that I am given a big book written with Chinese hieroglyphs, and explained “the rules of the game”: periodically, I will get two sheets of paper written with sequences of Chinese hieroglyphs, and my task is to combine the symbols (hieroglyphs) from these two sheets and from the book according to a certain system of rules (a certain algorithm), to write down the result on a clean sheet of paper, and to give this sheet out. I am

[1] If there is a direct answer, it may be found without understanding — as the combinations of symbols in the text which closely concurs with the combinations of symbols in the question.

given the handbook with the rules (algorithm) of combining symbols. Then I am placed in a special room; through its window, now and again, experimenters give me two sheets of paper written with Chinese hieroglyphs and take the sheet of paper that I have filled according to the rules (algorithm). Let us suppose that the book contains (in Chinese) the background information about restaurants and the behaviour of their clients, and the sheets of paper that I get contain stories and questions about restaurant affairs in Chinese. However, I do not even guess it, because I do not understand Chinese. Nevertheless, if I precisely execute the combining algorithm, and the algorithm itself is good enough, then the sequence of Chinese hieroglyphs that I write down on the paper and give out to the experimenters will be (though I do not even guess about this) "the answer" to "the question" asked.

Imagine that there is another room next to mine, and a literate Chinese sits in it. He receives sheets of paper with the same stories and questions as me and gives out sheets of paper with the same answers. From the point of view of an external observer who sees only the sheets of paper that come in and out of the two rooms, there is no difference. He cannot guess in which room the genuine Chinese sits. Nevertheless, there is a cardinal difference between what I do and what the Chinese does: for me, those combinations of symbols that I get, produce and give out are meaningless, — I do not understand them; whereas for the Chinese, they are the story, the question and the answer, — he understands them. I fulfil the functions of a computer program that *imitates (simulates) understanding* whereas he genuinely *understands*.

The Chinese Room Argument and the Turing Test

The great mathematician who was the developer of the mathematical foundations of computer technologies, Alan Turing had proposed a test purported to serve as the criterion of whether a computer *understands* information (and whether it, therefore, has consciousness).

The test is as follows: an expert-tester "communicates" *via* a keyboard and a monitor with, on the one hand, a person and, on the other hand, a computer that runs a program. The expert does not see and does not know with whom he communicates: which of his "interlocutors" is a human being and which a computer. If the expert cannot correctly distinguish which is which, or is as often mistaken as successful in his guesses, then the computer (program) has passed the Turing test. In this case, we are proposed to conclude that the computer *understands* those questions that were asked by the expert and gives meaningful answers (and, therefore, has consciousness).

No computer (computer program) has passed the Turing test as yet. Nevertheless, it seems that there are no weighty reasons to think that it is *in principle* impossible for a computer to pass the Turing test. It is possible to load any amount of information into a computer database. There are various algorithms and programs for a computer to generate answers to the questions of experts. Although at the present moment they are imperfect and incapable to mislead experts, they can be improved permanently. The developers of the programs will investigate why the experts have decided that they communicate with a computer rather than a person (which answers seemed to them non-human and what could be plausibly the human answers), will make corresponding changes-improvements into the algorithms, supplement the database; as a result, the answers of computers will become more and more (ever more often) like those that a person could give; so it will be more and more difficult to guess where is a computer and where a person. It is impossible to predict how far things will actually advance in this respect. But in principle, it seems that there are no weighty reasons to deny the possibility of such algorithms (programs) that their execution, given an access to a big enough store of information (commensurable with the whole amount of the

knowledge of an ordinary person, including what the person does not remember explicitly but retains somehow in his memory, so that he can recollect it in some circumstances or uses it unconsciously, *etc*.), can mislead the most inventive experts-testers.

However, Searle's thought experiment shows that even if a computer will pass the Turing test, this will not mean that it understands the information. Moreover, since we know that all the computer does is combining symbols according to a certain algorithm, we have weighty reasons to think that the computer, even if it passed the Turing test, does not understand information. If I, when in the Chinese room, have a large enough database in the form of a huge book written with Chinese hieroglyphs, and I have a correct algorithm and execute it precisely, *I would pass the Turing test on the understanding of Chinese, although I would not understand Chinese.*

"Aboutness" (Intentionality)

The Chinese room argument makes it clear that there is a principal difference between real understanding and its imitation — even if it is a perfect imitation (simulation) that is externally undistinguishable from the original. But what is the difference, precisely? What "the real understanding" means? We can understand this better if we address the concept *of intentionality* or *aboutness* (*ofness*).

This concept indicates the fact that our thoughts, ideas, statements, words are not merely sequences of sounds or signs (letters), — they mean something, are *about something*. For example, the word "tree" means a tree — some plant that belongs to the class of trees. And the statement "The foliage on the tree is green" is a statement about the tree and its foliage.

In the terms of linguistics, this property of language expressions is called "semantics". *Semantics* is the *realm of the meanings of language expressions* (what words, phrases, sentences mean, what they point at, what they are about); it must be distinguished from *syntax* as the system of rules of the organization of language expressions — linking words into phrases, phrases into sentences, sentences into more complex texts. Using these terms, Searle sums up the Chinese room thought experiment with the conclusion that computers "have only a syntax, but no semantics. Such intentionality as computers appear to have is solely in the minds of those who program them and those who use them, those who send in the input and those who interpret the output" [Searle, 1980: 422]. In the Chinese room case, the intentionality (meaning) that seems (from the outside) to be present in the sequences of Chinese hieroglyphs that I have produced (although I actually do not understand them; they *mean nothing for me*) is contained only in the minds of external observers who know Chinese.

Physical states and processes (including those that occur in human brains) do not have such a property as intentionality: they are merely what they are — certain physical structures and processes of their changes; they never are *about* something. However, *we can use* certain physical objects, properties, states, processes to designate by them something, so that they would represent-symbolize *for us* some *meanings*. This is exactly what we do when we write books or when we design and use computers. In the simplest case of oral speech we use some sound sequences (which physically are but specific vibrations of air caused by movements of our throats, tongues, lips) to represent and transmit various meanings. But such physical states and processes have no meanings *by themselves, as physical states and processes*; their meanings are *conventional* — they are *assigned* to them by *an agreement (convention)* between people. If wind caused by accident air vibrations that sound like some statements in Ukrainian, these sounds would not be meaningful statements. Statements in Ukrainian are not

meaningful for those people who do not understand Ukrainian, just as Chinese inscriptions are not meaningful for me. In fact, any physical objects may be used to represent certain meanings. For example, we can use stones or configurations of stones of certain sizes and forms to designate certain words. We can introduce certain rules of the organization of this stone-writing that will impart certain configurations of stones with the meanings of sentences and texts. Accordingly, some complexes of stones can represent meaningful statements, — conventionally, but not physically: for all those people who were not taught the meanings (semantics) and rules (syntax) of this stone language, they are merely heaps of stones that have no meaning, however much these people know about their physical properties.

Computer programs have nothing to do with understanding. As Searle says, computers do nothing else besides "the formal symbol manipulations" which "aren't even *symbol* manipulations, since the symbols don't symbolize anything" [Searle, 1980: 422] (mean nothing) for the computer.

The popular notion that the brain is sort of computer and that computers can think arise mainly from "a confusion about the notion of "information processing"" due to its ambiguity. It seems that human thinking is "something called "information processing", and analogously the computer with its programs does information processing". However, this kind of talk is misguiding:

"In the sense in which people "process information" when they reflect, say, on problems in arithmetic or when they read and answer questions about stories, the programmed computer *does not do* "information processing". Rather, what it does is manipulate formal symbols. The fact that the programmer and the interpreter of the computer output use the symbols to stand for objects in the world is totally beyond the scope of the computer. … The introduction of the notion of "information processing" therefore produces a dilemma: either we construe the notion of "information processing" in such a way that it implies intentionality as part of the process or we do not. If the former, then the programmed computer does not do information processing, it only manipulates formal symbols. In the latter, then, though the computer does information processing … it is up to outside observers to interpret the input and output as information in the ordinary sense. And no similarity is established between the computer and the brain in terms of any similarity of information processing." [Searle, 1980: 423]

The "Mind-of-the-System" Objection

Searle's opponents have offered a witty objection. They have admitted that if I execute the corresponding algorithm in the Chinese room, then I, indeed, do not understand Chinese and what is written in the book and on the pieces of paper. Nevertheless, they contend that if an external observer, after watching the movements of the sheets of paper and reading the inscriptions on them, would conclude that there is someone who understands Chinese in the Chinese room, he would be right too. From their point of view, the understanding of Chinese would reside not in *my* mind, but in the mind of the Chinese room. If I process the information (manipulate symbols) according to the correct algorithm (the same as that according to which the brain processes it), then *the system* that consists of me, the book and the auxiliary facilities (pencils etc.) has its own mind (distinct from my mind) that understands Chinese.

To this, Searle objected that even if I had such supermemory and superintelligence that I memorize the algorithm and the whole sequence of symbols in the book and execute all the data processing (symbol manipulating) in my mind, I would not understand Chinese all the same. To waive this objection, David Chalmers contended that because the situation is essentially the same as in the case when I use the book and the auxiliary facilities, but is

"internalized", so the mind that understands Chinese should emerge all the same: "this should … be regarded as an example of two mental systems realized within the same physical space" [Chalmers, 1996: 326]. That is, Chalmers suggested that my brain would be associated not only with my mind, but also with some second mind that understands Chinese, although I do not even suspect its existence. As for me, this contention is hugely implausible, for it is me, not some unknown to me second self of my brain, who processes the information. (Perhaps, it did not look so implausible from Chalmers' point of view because he adhered to the theory of epiphenomenalism according to which the mind does not influence anything anyway: I, as well as the Chinese in the next room, would do exactly the same if I (he) had no mind at all. If so, it is quite possible that a data processing by the brain generates some new mind that is unknown to me, or several such new minds...)

Although the idea that in the Chinese room (or even in my brain) a second mind emerges that understands Chinese seems very implausible, the argument is interesting in that it prompts to think in the opposite direction:

if the supposition that a system that consists of a person having a mind and some auxiliary facilities would, in virtue of processing data on a certain algorithm, have (or generate) its own conscious mind distinct from that of the person is very implausible,

then is not it just as implausible a supposition that a system that *consists only of elements that have no mind* (that interact purely automatically according to physical laws) would, in virtue of processing data on a certain algorithm, have (or generate) a conscious mind?

This pertains equally to both: the computer executing a program and the brain. It is just that we have got used to associate the brain with the mind; however, if we put this habit aside and look at the situation unbiasedly, there is nothing special about the brain to give it some principal advantage over other possible realizations of the same information-processing model — whether it as a Chinese room or a computer. This is well illustrated by Terry Bisson's science fiction story in which intelligent extraterrestrials with silicon-based bodies and brains, on having learned about humans, are unable to believe that *ordinary meat* (or grey jellylike substance of the human brain) may be conscious and intelligent. Our feeling of the incredibility of the idea that a silicon-based computer can have (not merely simulate) a conscious mind, or that a Chinese room can have it, is essentially the same.

From the Chinese room to the Chinese robot

Another interesting attempt at a rebuttal of the Chinese room argument consists in the supposition that a computer realisation of understanding is possible if the Chinese room, which realises *syntax*, is supplemented with *semantical* functional blocks. What is meant is as follows.

On Searle's view, the Chinese room thought experiment proves that however perfect a computer imitation of consciousness (understanding) is; it is *only simulation* all the same*, not realisation* (Searle uses the word "duplication"). In particular, it would be so even if a computer can answer questions so well that by these answers it would be undistinguishable from a human person. However, as the Russian philosopher Vadim Vasilyev notices, there is a problem: how to deal with such questions that a person does not know the answer but can answer after looking somewhere (or addressing some source for additional information in some other way)? Obviously, a computer cannot be programmed in advance to give such answers. Let us consider some examples in the model of the Chinese room.

If I am given a note with the Chinese hieroglyphs that mean the question "What time is now?", I cannot generate the answer by manipulating Chinese hieroglyphs according to a

given algorithm. On the other hand, if the algorithm contains the instruction to look at clocks and to process their figures according to further instructions, and as a result I generate some Chinese hieroglyphs, I can easily guess the meaning of the input sequence of hieroglyphs (the question "What time is now?") and of those hieroglyphs I generate for an output.

"Or let us imagine that the person who asks questions has beforehand brought in that room some thing — for example, a Coca-Cola can — and put it on the table … And now he asks: "What do you see on the table?"" [Vasilyev, 2009: 88] The question is asked in a note in Chinese. If the algorithm of processing of symbols contains the instruction to write down a hieroglyph depending on what there is on the table, I will probably guess the meaning of this hieroglyph and of the hieroglyphs in the question note.

Or suppose that I am given a note with the hieroglyphs that contain a fragment that a Chinese could read as the question: "What weather is now?" But I do not know about this; I just implement an algorithm of processing these symbols. The next instruction of the algorithm happens to be: "Look out of the window, if it rains, write down the hieroglyph  ; if there is sunny weather, write down the hieroglyph  ."

Obviously, while implementing such algorithms I will begin to understand Chinese.

But if so, let us imagine that a computer is connected with peripheral devices (such as a television camera, a sound recording device, a clock, a thermometer, *etc*.); these devices can on demand of the computer obtain external data and transmit data streams to the computer; then the computer processes these data on a certain algorithm. Will not such a system, while performing algorithmic manipulations with Chinese hieroglyphs and interacting with the external world, develop some understanding of Chinese? Let us imagine such a system as a robot having a computer instead of a brain, television cameras instead of eyes, sound recording devices instead of ears, electrically operated metallic legs and hands, *etc*. Some Searle's opponents suggested that such a robot would be capable to develop understanding of Chinese.

Searle have discussed this objection and given his reply [Searle, 1980: 420]. To take into account some specific features of Vasilyev's objection, I will adapt Searle's reply and work it out in a bit more details.

Let us imagine a Chinese robot with a Chinese room in its head; in the room, I manipulate Chinese hieroglyphs according to some algorithm. For this to be analogous to what a computer does, the algorithm I implement *should not* contain instructions of the sort: to look at a clock, or in a window, or elsewhere and write down hieroglyphs depending on what I see. Instead, I just transfer and receive through the room's input/output window some extra (as compared with the initial version of the Chinese room experiment) sheets of paper with hieroglyphs, or some other symbols which meaning is unknown to me (for example, numbers about which I do not know what they designate). Suppose that I execute my algorithm, and my next instruction is: "Write down the numbers 235, 245, 655 on a sheet of paper, put the sheet in the window, and wait for a reply, which also will be a sheet of paper with some numbers". I execute the instruction and, as a result, I get a sheet with the numbers "123, 545, 36789, 25". The next instructions of the algorithm tell what manipulations with Chinese hieroglyphs and figures I have to perform depending on the numbers in the list. Obviously, in such cases the execution of the algorithm will not provide me with any understanding of Chinese. However, this is just what is done by a computer that communicates with peripheral devices (video cameras, sound recorders, clocks, thermometers, *etc*.) to obtain external data. The computer changes electromagnetic states of some data cells (that are connected to some peripheral device) — these states can be described in the language of computer science as a sequence of "0" and "1" (bits) or, for a more compact representation, as a sequence of larger numbers (for example, a sequence of 8 bits — a

byte — can be represented as a number from 0 to 255) or (on a convention) any other symbols (for example, Chinese hieroglyphs). Such changes cause certain processes in the connected peripheral devices; as a result, these devices cause changes of some other data cells of the computer. Then the computer processes the values (sequences of bits, numbers, or Chinese hieroglyphs) of these cells according to its algorithm. As we see, nothing essentially changes in comparison with the initial experiment of the Chinese room, — at least as far as the computer and its (absent) thinking-understanding is concerned.

However, Vasilyev disagrees with such a conclusion:

> "…the modification of the Chinese room and … the transformation of the Room into the Robot … radically changes the conditions of the mental experiment … First, the robot is provided with its own program that is unknown to Searle, whereas it was originally supposed that Searle knows and executes the whole program that allows to emulate the language competence.[2] Secondly, the program installed in the robot has a rather specific, quasi-semantic character. For this program has to specify not only the relations between hieroglyphs, but also the relations of hieroglyphs to certain non-hieroglyphic, physical data that are obtained by its television camera; and these data, unlike hieroglyphs, would be understandable for Searle, if he could get acquainted with them. Accordingly, if we suppose that, in accordance with the initial assumptions, Searle would know the whole program, or the set of programs that allows emulating meaningful Chinese speech, then he would begin to understand Chinese. Really, let us suppose that Searle learns the robot's program. In that case, he would understand that a certain movement of its television camera corresponds to a certain Chinese command, that the presence before the television camera of things of a certain kind is designated by certain hieroglyphs, that another kind of things is designated by others, *etc*. But such knowledge would allow him to fill syntactic aspects of Chinese with semantic contents" [Vasilyev, 2007: 89-90].

I think that this objection is mistaken. In the situation of the Chinese robot, there are no other algorithms besides the algorithms that are executed by the computer; and I could execute them just as well in the Chinese room — I would just need to deliver through the input/output window extra sheets of paper with numbers generated on the algorithm and to receive other sheets with numbers. The peripheral devices that receive commands from a computer and transmit data to a computer do not need their own program (algorithm). Their interactions with the computer and the environment may be purely causal[3]: certain states of computer data cells cause certain physical processes in a peripheral device; those processes determine the mode of interaction between the peripheral device and the environment; this interaction, in turn, determines those electromagnetic-informational states that the peripheral device causes in the corresponding data cells of the computer. So, the whole algorithmic processing of information occurs in the computer.[4] And it can be executed just as well, in

[2] In this quotation, "Searle" stands for the person in the Chinese room who executes the algorithm of symbol manipulation.

[3] Searle, when discussing "the robot reply", has also remarked: "The first thing to notice about the robot answer is that it tacitly concedes that cognition is not solely a matter of formal symbol manipulation, since this reply adds a set *of causal relations with the outside world*." [Searle, 1980: 420] (Italics mine.)

[4] In fact, some peripheral devices *may* be programmable (or once programmed when produced) and to execute their implanted programs, but this *is not necessary*, for all such programs can be executed by the computer, so that all peripheral devices are relieved of all data processing, and all that is left for them to do are casual interactions.

principle, by a person in the Chinese room; and it can be "interiorised" (on the assumption of the supermind with supermemory). And, as I explained above, such knowledge and execution of algorithms cannot provide a person with any understanding of Chinese.

Despite Vasilyev's contentions, programs should not "specify … the relations of hieroglyphs to certain non-hieroglyphic, physical data that are obtained by its television camera", so that "these data, unlike hieroglyphs, would be understandable for Searle, if he could get acquainted with them". All that the program should specify, as far as the data obtained by the computer from the camera is concerned, are the relations of hieroglyphs to informational states (data) of the corresponding cells of the computer memory. And these data, just as well as hieroglyphs, would not be understandable for Searle, if he could get acquainted with them, for they, as well as all other computer data, are just sequences of zeros and units (values of bits) that can mean anything whatever. Also, "if we suppose that, in accordance with the initial assumptions, Searle would know the whole program, or the set of programs", in that case, he would *not* "understand that a certain movement of its television camera corresponds to a certain Chinese command, that the presence before the television camera of things of a certain kind is designated by certain hieroglyphs", since he does not know neither about the movement of the camera, nor about its construction, nor about things before it, nor even that the "command" (a sequence of bits, or numbers, or hieroglyphs) he generates while executing his algorithm is directed to a television camera rather than a scanner, or a sound recorder, or a clock, or a thermometer.

Admittedly, from an algorithm of processing the data that represents physical objects of a certain kind it is often possible, in principle, *to guess* what physical objects are represented by this data. However, such guessing is *an extra job* of *thinking and imagination* (*based on the understanding of meanings*), distinct from mere memorizing or performing the algorithm. And for such guesses to be possible, the guessing subject (person) must already have a good notion about the relevant kind of physical objects.

Vasilyev could object that even if my above reasonings are true, neither they, nor Searle's argumentation do not show that *the system consisting of a computer* (that manipulates symbols on a certain algorithm) *and peripheral devices* that interact with the computer and the environment (even if this interaction is purely causal, and peripheral devices do not execute any program of their own) do not think (in the sense of understanding-based thinking). I agree with it. Really, all that my previous reasonings (if they are true), as well as Searle's argumentation, have shown, is that *a computer* that executes algorithms does not, by virtue of this, understand — that the execution of algorithms by a computer is not sufficient for understanding-based thinking. To expand this conclusion on *a system that consists of the computer and its peripheral devices*, additional arguments are needed. In what follows, I provide them.

Let us suppose that we have a system that consists of a computer (that manipulates symbols on a certain algorithm) and peripheral devices that can interact with the computer and the environment and so supply the computer with external data. Let us suppose that these peripheral devices were never used as yet; the computer has never addressed them. The system undergoes the Turing test on the understanding of Chinese. The testers ask only such questions that the computer can answer them without addressing external sources (through peripheral devices) for additional data. As in the initial version of Searle's thought experiment, the computer just manipulates formal symbols on a certain algorithm. We know (from Searle's argument supplemented with the reasonings I have proposed above) that such data processing is insufficient for understanding-based thinking — the computer does not

understand the information it processes, as well as the meaning of the operations it executes. Because the peripheral devices stay idle, the data processing by the system in this case is equivalent to the data processing by the computer alone. Hence, the whole system, as well as the computer, does not really think, does not understand. Thus, the system is unable to understand any story and questions to such stories.

Now let us suppose that the system is asked such a question that the computer needs to address some peripheral device to obtain some external data. The computer executes its algorithm; at some step, it changes the informational states (a sequence of bits) of the cells that are connected with a peripheral device; this causes some processes in the device that, in their turn, cause changes of informational states (sequences of bits) in the corresponding memory cells of the computer; the computer processes these changed states according to its algorithm. As we know (from Searle's argument and the reasonings above), the computer does not understand the meaning of these happenings (it is not consciously aware of them at all); even more so, the peripheral device has no understanding (and no conscious awareness). Is it possible that the system consisting of these two understanding-and awareness-devoid devices will have awareness and understand? It seems very implausible, but let us leave the question about such a possibility open for a while.

What changes in the system as the computer communicates through peripheral devices with the external world more and more? Nothing that could be of significance for understanding. Some data (sequences of bits) in some memory cells of the computer get rewritten, change. But any such data could be written in the corresponding memory cells initially, before any interactions of the system with the external world.

Let us take two identical computers with identical memories connected with identical peripheral units. The first system intensely communicates through peripheral devices with the external world for a long period. The second system all this time is turned off. Before turning it on, a programmer writes in all its memory cells exactly the same data as the data contained in the corresponding cells of the first computer. Now these two systems are absolutely identical with respect to all informational states (let us suppose that peripheral devices have no memory of their own).

We have already shown that the second system (that was turned on for the first time a moment ago) does not understand the data stored in its computer memory, as well as stories and questions entered from the keyboard. But since the first ("experienced") system is absolutely identical with the second with respect to all informational states, this means that it does not understand all these things too, and all its "experience" of interactions with the external world does not provide it with any understanding.

The only possibility to avoid this conclusion is to suppose that (1) with the system, a new conscious mind emerges as something besides its computer data processing; (2) this mind gradually develops in the process of the system's interactions with the external world, and (3) this mind is epiphenomenal — it effects no influence on any actions of the system and the results of its data processing. After all, if the two systems are absolutely identical in their data processing algorithms and have the same data in all their memory cells, then in identical situations their data processing operations and results should be absolutely the same, irrespective of their history (the history of interactions with the world and of the formation of the data in their memory cells). The system with a newly emerged empty mind that has not yet began to develop processes data (and passes the Turing test) exactly as the system with a mind that has developed throughout a long period of interactions with the external world!

Besides, let us note that even if we suppose that with the system there really emerges such an epiphenomenal conscious mind, and that due to this mind the system (robot) becomes an understanding being, this not only does not refute Searle's argument, but directly supports it. After all, the gist of Searle's argument is not that is impossible that a conscious mind may emerge in a computer or a robot in some miraculous way (let us imagine, that God has reincarnated into a robot a soul of some Chinese). The gist of Searle's argument is that performing algorithms is not enough for understanding and understanding-based thinking; that understanding and thinking needs (in addition or instead of performing algorithms) a conscious mind that is something distinct from the informational states and processes (executing algorithms) of a computer. It is exactly what we would have in the case of a robot with an emergent epiphenomenal conscious mind: besides executing algorithms, there is a conscious mind, such that its current states and processes *are not determined entirely* by the current informational states and processes of the computer, as well as by any other current physical states and processes in the robot, but are dependent on the mind's own experience, its own non-physical memory.

🕮 References

Searle, John. Minds, Brains and Programs. In *The Behavioral and Brain Sciences*, Vol. 3. 1980: 417-424.

Chalmers, David. *The Conscious Mind.* New York, Oxford: Oxford University Press, 1996.

Vasilyev, Vadim. Coca-cola and the Chinese room secret. *Philosophy of Mind: Classics and Modernity*. Moscow: Savin S. A. Publisher, 2007: 86-94.

Vasilyev, Vadim. *The Hard Problem of Consciousmness*. Moscow: Progress-Tradition, 2009.