М

УДК 004.8; 001

A. V. Mishchenko
Institute for Research in Computer Science and Automation, Rocquencourt, France

AGNOSTIC EPIPHENOMENALISM AND QUALIA-READING PROBLEM: WILL ARTIFICIAL INTELLIGENCE BE A SUBJECT, HAVING CONSCIOUSNESS AND SUBJECTIVE EXPERIENCES?

А. В. Мищенко
НИИ информатики и автоматики, Роканкур, Франция

АГНОСТИЧЕСКИЙ ЭПИФЕНОМЕНАЛИЗМ И ПРОБЛЕМА СЧИТЫВАНИЯ КВАЛИИ: БУДЕТ ЛИ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ СУБЪЕКТОМ, ОБЛАДАЮЩИМ СОЗНАНИЕМ И СУБЪЕКТИВНЫМИ ПЕРЕЖИВАНИЯМИ?

The question of whether artificial intelligence can have subjective experiences and human-type on consciousness is important both for philosophy and the future of computer science.This paper replaces the question, posed by Frank Jakson, in his “knowledge argument” (also known as “Mary's room”) by a more formal “qualia-reading problem”, leading to a new branch of epiphenomenalism, called “agnostic epiphenomenalism”. Subjective world (for both human and artificial intelligence) is considered as constructed from “representational illusions”, which, according to the “theory of mind-matter”, are becoming more important then underlying reality.
Key words: artificial intelligence, subjective experience, qualia, knowledge argument, Mary's room thought experiment, RoboMary, physicalism, epiphenomenalism, illusions, mind-matter theory.

Вопрос о том, может ли искусственный интеллект иметь субъективные переживания и человеческое сознание, важен как для философии, так и для будущего информатики. В этой статье вопрос, поставленный Ф. Джексоном в его мысленном эксперименте “комната Марии“, заменяется более формальной “проблемой прочтения (считывания) квалии“, ведущей к новой ветви эпифеноменализма, называемой “агностическим эпифеноменализмом”. Субъективный мир (как для человека, так и для искусственного интеллекта) рассматривается как построенный из “репрезентативных иллюзий”, которые, согласно теории мыслящей материи, становятся важнее лежащей в их основе реальности.
Ключевые слова: искусственный интеллект, субъективный опыт, квалиа, аргумент знания, мысленный эксперимент «комната Марии», РобоМэри, физикализм, эпифеноменализм, иллюзии, мыслящая материя.

1 Introduction to knowledge argument

The knowledge argument (also known as “Mary's room”) is a philosophical thought experiment designed to stress differences between objective knowledge and subjective experiences inside human (or artificial) brain [1], [2]. This experiment describes Mary, color-scientist, living in a black and white room (with black and white television, black and white PC monitors, etc), isolated so that she never had her own perceptual experience of color. At the same time, Mary has the ability to study absolutely everything about color, including its physical properties, its effect on human retina and the neurophysiology of color-processing in human brain.

Frank Jackson, author of this thought experiment, poses his main question: what will happen when Mary leaves her black and white room and sees, for the first time, the real world in color? Will, at this moment, Mary know something new? Will she acquire any type of new knowledge from her personal experience of color [1]?

Jackson claims that Mary, once released from her room, will indeed get some additional, new knowledge.

This type of knowledge (which is possible to acquire only by personal subjective experience) is usually named “qualia”. The term "qualia" reflects that these properties are qualitative, as opposed to objective, quantitative measurements. Qualia is the subjective properties of experiences, which are not properties of objective facts. For example, sensations of taste (sweet, salty, etc) are not properties of sugar or salt; sensations of color (red, green, etc) are not properties of wavelengths of light.

Moreover, the sensations of taste/color are not properties of tongue/retina receptors. These receptors are simply activated by a specific chemicals or wavelengths, but there is nothing “red” or “sweet” in these receptors. These unique pleasing sensation of “sweet” loved by all children, appears subjectively in our mind, as qualia of taste.

If we support Jackson in his point of view that Mary cannot get this qualia of color by studying everything about color, then we agree that qualia is unreachable, isolated from the outside world. In this point of view qualia does not interfere with the world of objective facts and, therefore, it is not possible to somehow measure it from the outside. Such point of view is usually called “epiphenomenalism”. This term reflects the fact that qualia exists additionally to physical phenomena (Greek “epi” means "over", "on" or "near").

If we do not agree with Jackson and assume that Mary, from her studies, will know the subjective experience of color before leaving the room, we think that independent qualia does not exists. This point of view is called physicalism.

Physicalism claims that since Mary already knew "everything about color", that knowledge would include understanding the subjective sensations, the "qualia" of color. For example, that knowledge would include the ability to differentiate all colors. Mary would therefore already know exactly what to expect of seeing colors before leaving her black and white room.

Therefore, physicalism claims that this “simulation” or “modeling” knowledge is identical to the practical experience, and there will be no additional 'qualia' [3], [4].

Physicalism is opposing epiphenomenalism and assumes that everything is described by physics and there is nothing else than a physical world. The terms physicalism and materialism are sometimes used interchangeably. However it is possible to argue that there are phenomena, complying with materialism but not physicalism. For example, if we admit existence of something material, but, in principle, unreachable or indescribable by physics. We can fantasize, for example, that "dark matter" and "dark energy", which are

indescribable by nowadays physics, will be proved to be indescribable by physics in principle. At least, they can be proved to be unreachable. Similarly, as we cannot, in principle, reach inside of black holes, or other universes in multiverse theory, or even distant parts of our universe – all these are proved to be unreachable and non-observable (because any information can not travel faster than speed of light). Similar place, according to epiphenomenalism, is occupied by qualia: it exists, but is non-observable from outside – as, for example, the interior of black holes.

Paul Churchland [5] formulates the negation of physicalism in favor of epiphenomenalism as follows:

1. Mary knows everything there is to know about brain states and their properties.
2. It is not the case that Mary knows everything there is to know about sensations and their properties.
3. Therefore, sensations and their properties are not the same (≠) as the brain states and their properties [5].

Jackson formulates his interpretation similarly, but emphasizes the fact that Mary "does not know everything":

1. Mary (before her release) knows everything physical there is to know about other people.
2. Mary (before her release) does not know everything there is to know about other people (because she learns something about them on her release).
3. Therefore, there are truths about other people (and herself) which escape the physicalist story [2].

As usual with philosophical theories, both epiphenomenalism and physicalism has its own followers and opponents and it is impossible to prove and choose either of them:

It seems, however, that, in case of artificial intelligence, the choice between these two theories is clear: we expect AI to be entirely described by laws of physics, without any "qualia" inside its CPUs.

The AI-related modification of "Mary's room" experiment was proposed by Daniel Dennett and is described in the next section.

2 Robo-Mary Thought Experiment and Subjective Experiences of Artificial Intelligence

From the AI-scientists point of view, the most interesting modification of experiment (and the most interesting objection to the point of view that Mary will get a new knowledge) was invented by Daniel Dennett.

In order to justify physicalism and deny epiphenomenalism, he modifies the Jackson's thought experiment by replacing Mary by a robot.

Dennett begins with a "deliberately simple-minded version", where RoboMary is a standard robot without color vision: "her video cameras are black and white, but everything else in her hardware is equipped for color vision" [4]. Leaving black and white room corresponds, in this case, to changing these black and white cameras into the color ones.

In this thought experiment Dennett enables RoboMary to use her knowledge, to create a special software to colorize the input from her black and white cameras. Similar programs already exist – they are used to colorize old, black and white movies into their colorful versions.

Obviously, when RoboMary finally gets her color cameras, and disables her colorizing software, nothing is changed. So, RoboMary already knew everything not only about outside world, but about all anticipated subjective experiences as well.

Another version of RoboMary thought experiment is if she is prohibited from reprogramming herself. In this case, Dennett advises RoboMary to build a RoboMary-2, the model of herself, so that she will be able to observe everything what is going on inside RoboMary-2, both when this model is in a state with a black and white and in a state with a color cameras.

RoboMary notes all the differences between black and white state and color state of RoboMary-2 and makes the same adjustments to herself. This way he puts herself into color state as well. Therefore, RoboMary, also, can acquire any knowledge about all anticipated subjective experiences before getting her own color cameras installed.

Note, that J. Christopher Maloney has the similar reasoning for the human Mary: “If... Mary does understand all that there is to know regarding the physical nature of colour vision, she would be in a position to imagine what colour vision would be like. It would be like being in physical state Sk, and Mary knows all about such physical states. Of course, she herself has not been in Sk, but that is no bar to her knowing what it would be like to be in Sk. For she, unlike us, can describe the nomic relations between Sk and other states of chromatic vision...Give her a precise description in the notation of neurophysiology of a colour vision state, and she will very likely be able to imagine what such a state would be like” [6].

Similar to Maloney, Dennett himself creates this RoboMary thought experiment in order to claim that the same is true for the original, human Mary-experiment: "if materialism is true, it should be possible (“in principle!”) to build a material thing–call it a robot brain–that does what a brain does, and hence instantiates the same theory of experience" [4]. He notes that colorizing software and any other reprogramming is just a "robot version of... trans-cranial magnetic stimulation" [4]. Dennett claims that the logic “true to RoboMary” => “true to Mary” is correct, since "contemporary materialism... endorses the assertion that we are robots of a sort"[4].

In the section 4 we will use the same logic backwards: “true to Mary” => “true to RoboMary” to discuss possibility that artificial intelligence can have subjective experiences (qualia) and, therefore, be subjects in the very human sense. But before that, let's summarize the opinions of Jackson's opponents.

3 "Customs" of subjective experiences : Responses and objections to Jakson's knowledge argument

From my point of view, the popularity of Jackson's experiment is based on the fact that he takes a real, reproducible situation (deprivation from some sensation and then experiencing it). This situation seems easy to reproduce and familiar to everybody. Therefore, numerous discussions appear. But then Jackson speaks about "all possible knowledge", this practical situation becomes, on contrary, rather theoretic and non-reproducible. As a result, everybody wants to declare his personal opinion about what is "all possible knowledge" and what is not.

These discussions around Jackson's experiment resemble jokes about customs and smugglers: Jackson (and his followers) plays the role of customs, trying to disallow to smuggle the subjective experiences from outside world into Mary's brain. Jackson's opponents, as caught smugglers, try to explain to customs that what Mary had in her baggage, after leaving her room, was a legitimate item.

For example, some opponents explain, simply, that the color experiences will leak into these black and white room. For example, Evan Thompson notes that Mary will be able to see color in dreams, as well as in afterimages from light perception [7].

Others, such as Nemirow [8] and Lewis [9] invent the "ability hypothesis", explaining that what Mary gains is not a knowledge, but ability, which was not described in the original version of the "Jackson's customs". Similarly, Churchland distinguishes between two senses of knowing: "knowing how" and "knowing that" [10].

Earl Conee explains that, upon release, Mary is not getting any new knowledge, but is getting “acquainted” with previously known [11]. He formulates his "acquaintance hypothesis" as follows:

1. Qualia are physical properties of experiences (and experiences are physical processes). Let Q be such a property.

1.Mary can know all about Q and she can know that a given experience has Q before release, although—before release—she is not acquainted with Q.

2.After release Mary gets acquainted with Q, but she does not acquire any new item of propositional knowledge by getting acquainted with Q (in particular she already knew under what conditions normal perceivers have experiences with the property Q).

Owen Flanagan explains that Jackson's descriptions does not distinguish between "metaphysical physicalism" and "linguistic physicalism". And explains that "Mary knows everything about color vision that can be expressed in the vocabularies of a complete physics, chemistry, and neuroscience" – all that corresponds to linguistic but not a metaphysical physicalism: “Metaphysical physicalism simply asserts that what there is, and all there is, is physical stuff and its relations. Linguistic physicalism is the thesis that everything physical can be expressed or captured in the languages of the basic sciences…Linguistic physicalism is stronger than metaphysical physicalism and less plausible” [12].

A metaphysical physicalism can be something impossible to express in language, but nevertheless a fact about the physical world, such as Mary's experience after leaving her room [12]. Similarly, to Flanagan, Torin Alter agrees that Jackson mixes physical facts with "learnable" facts: “some facts about conscious experiences of various kinds cannot be learned through purely discursive means. This, however, does not yet license any further conclusions about the nature of the experiences that these discursively unlearnable facts are about. In particular, it does not entitle us to infer that these experiences are not physical events” [13].

In a similar way, Daniel Dennett in his RoboMary experiment, tries to “convince Jackson's customs” that what is allowed to RoboMary should be allowed to human-Mary as well.

4 Agnostic epiphenomenalism, qualia-reading problem and my answer to Jackson's knowledge argument

Putting aside the jokes about customs and smugglers, we should note that the discussions about “forbidden items”, described in the previous chapter is normal – we, at this moment, do not know the nature of such “items” as qualia, consciousness, subjective experiences. And, the good thing about this experiment is that it allows scientists and philosophers to discuss and try to better understand all these notions.

The bad thing about this experiment is that, during all these smuggling games, some essence of Jackson's question slips away from our attention.

In order to better highlight this essence, I can reformulate the Jackson's question to the "qualia-reading question". It can be formulated as follows: can Mary, by studying the neurophysiological, observable " from outside" effects of color vision (EEG, MEG, MRI and all precise measurements of brain activity) reproduce/mimic these observable effects and be sure that the state of mind that she experiencing during these imitations is the same as the state of mind when she will actually see colors?

I can formulate this more generally, as "qualia-reading problem" (without referencing Mary's experiment):

If the observable activity of two brains are similar - can we be sure that these two subjects have similar subjective experience?

Speaking about color vision, for example, can you be sure that I do not see red light at places where you see green? Can we be sure that the same activity of retina-receptors and the same activity of visual cortex, "mean" that we see the same color? Remember that retina-receptors do not have any color, they just react to some wavelength. The color is an invention of the brain, “representational illusion” (as it is called in the following section) making for us this beautiful representation of the outside world. Why our representations should look the same, but not, for example, depend on our DNAs or, at all, be random?

I can imagine Dennett's objection to this: "Really? Why such extravagant idea?" and he would be right – I do not know any reasons for this. And that is why I am not trying to prove that we have different color-picture of the world. I am trying to prove that it is impossible to prove that we have the same color-picture of the world. I am trying to prove that we cannot be sure that we are not mistaking about qualia of others. That means that we cannot undoubtedly "read" qualia of another person. Subsequently, for example, we cannot be sure that we can, someday, correctly copy our " internal worlds" to robots, clones or any other media.

So, my answer to Jackson's knowledge argument is that Mary will know everything about the color, including descriptions of subjective experience of color of other people. But, contrary to objective knowledge about the color, she, in principle, cannot be sure that, after leaving the room, her subjective experience of color will be as she expected. Note, that exactly the same can be said about RoboMary or about anything exiting the room (if, of course, this “anything” is having its own subjective experiences).

In relation to "qualia-reading problem" there is no difference between human and artificial intelligence – they are both a sort of mechanisms, which, at some stage of complexity, may (or may not) have their own, "unreadable to others" subjective experiences.

This viewpoint can be called agnostic epiphenomenalism

5 Objects and subjects in human and artificial consciousness

Essentially, all objections to existence of qualia from Daniel Dennett and other philosophers, close to physicalism, boils down to the affirmation that qualia is an illusion. Such viewpoint is a usual consequence of self-consistent materialism, as if material objects are the only existing reality, then, if something can not be observed by means of this objective reality (if it can only be perceived subjectively), than this something is an illusion.

More generally, all theories with objects as its base (such as physicalism) try, in some way, to eliminate subjects. They either ignore them (for example, science is dealing only with objects) or assert that subjects are illusory or abstract concepts. We can note that the contrary is also true: all theories with a subject as its base (such as subjective idealism) try to eliminate objects: they assert that objects (and the whole reality) may be our

illusions, kind of hallucinations, happening in our minds. Some philosophical and religious concepts, such as Zen-Buddhism, try to go “beyond the canonical distinction between subject and object”, claiming that “insight is that what does not need a subject ” [14].

In general, it is impossible to agree with any of these viewpoints before better understanding of what are illusions. I think that the role of what we usually call "illusions" is largely underestimated in psychology, philosophy and theory of Artificial Intelligence.

In the next section we discuss so called "representational illusions", which can be considered as one of main " building bricks " of human and artificial consciousness.

6 Representational illusions in Agnostic epiphenomenalism and Artificial consciousness

The above-mentioned idea (that qualia is an illusion) can be generalized even further. Strictly speaking, all existing in our internal world (including thoughts, feelings, mental constructions and the consciousness itself) – all these are illusions, as they do not exist in the real, objective world. For example, the EEG images of feelings and neural correlates of consciousness exist, whereas the feelings themselves and consciousness itself are illusions, generated by our brain. They appeared during evolution in order to help us to optimize our behavior. Feelings appeared to help us to behave adequately and adapt to behaviors of other people. Consciousness appeared to model the behavior of outside world to better plan our actions [15], [16].

The only difference between feelings and what we usually call "illusions" (dreams, hallucinations and other "uncontrolled" sensations) is that brain tries to create feelings in accordance with what happened in objective world, whereas dreams, hallucinations or, for example, artistic ideas – may happen without any correlation with what is happening outside. For example: brain creates feeling of disappointment inside if we didn't achieve something outside. The same way, it creates colors and musical harmonies, which are also illusions and do not exist in reality. What exists instead of color is repetitive change of electromagnetic fields (which we call waves). What exists instead of musical harmony is repetitive change of air pressure.

According to epiphenomenalism, all our subjective sensations are not causes but just representations of our actions. The simplest example is accidentally touching hot cooking plate: it only seems that you draw back your hand because you felt pain. If you will observe your sensations, you understand that you feel the sensation of touch first, then you draw back your hand, and often only after that you feel pain. The same is true for more complex examples, such as voluntary actions: it only seems that you fight because you feel anger. Epiphenomenalism asserts that your decision to fight is automatic, pragmatic and subconscious. Only after this decision, brain generates the " emotional image " of this situation (feeling of anger). This emotional image is made just to create a reason for your consciousness: why you are fighting [17]. The feeling of anger is created in less than a second and, therefore is hard to notice what caused what.

To understand how emotional images can be created afterwards, it is possible to remember numerous examples from other sciences, such as sociology.

For example, psychological decision to fight is similar to sociological situations when states begin wars. Such decisions are made also because of pragmatic reasons – such as to acquire territory (for example, access to the sea or other assets) or to improve positions in global politics, or to change the neighboring government to make from them a

marionette, satellite-state. This decision is usually made secretly and pragmatically by military-people, and only afterwards, other, culture-people create emotional image of this war, usually turning it over into fair and compelled war. It is also made just to create a reason for people: why they are fighting so far from their Motherland. In difference to psychological decision to fight (where emotional image is created in less than a second), the emotional images in society are created in days or even months and are easier to notice.

Another illustrative example from psychology is the personal feeling of sexual attraction. Pronounced sexual characteristics in a person of an opposite gender cause in your brain a sexual interest. Self-expectation of this sexual interest when you see pronounced sexual characteristics is exactly this emotional image, which we call "beauty", which is also subjective and exists only in our brains, being the same kind of illusion.

Such illusions (not only feelings, but all subjective sensations) are not similar to dreams or hallucinations, because they are created under control of sensory information, in order to somehow summarize useful information. In accordance with this, I call them "representational illusions".

They appeared evolutionary to have some simple representation of what is happening. This way, not only feelings were created, but, for example, colors as well : the brain of our herbivore ancestors managed to create the green color to represent all eatable parts of plants. Similarly, our more recent ancestors acquired ability to distinguish red from green to notice fruits among leaves. With black and white vision, wavelengths of 620–750 nm and 495–570 nm (the ones that are represented in our brain as "red" and "green" correspondingly) are indistinguishable, and therefore, it is impossible to notice fruits among leaves.

Note, that recent neurophysiological experiments already proved that many phenomena of our internal world are illusions. For example, as was proved in [18], what we call a free-will (conscious decision to do something) is an illusion. In the corresponding experiment, the recorded brain activity turned out to contain information about decision several seconds before the person, (whom brain activity was recorded) thought that he is consciously making this decision [18].

Finally, it is worth to mention, that the same "representational illusions" are becoming an important part of both emerging artificial consciousness and hybrid human-artificial consciousness. For example, what we see on the computer screen (such as words of this text or a computer game character playing football) are illusions. It seems that, if a leg of a footballer character hits a ball, then, according to physical laws, this ball will fly in an appointed direction. In reality there are no laws of physics in this case: pixels representing ball have no reason to be "pushed" by pixels representing a leg of a footballer. Nothing of these exists in reality – similarly as feelings, colors or musical harmonies do not exist. Nothing what we see in computer game exists in reality.

But, nevertheless, what we see on computer screen is more important than underlying reality. For example, words of this text are more important than underlying processes in CPU. Note that these words are more important not only for human or hybrid human-artificial consciousness, but for artificial consciousness as well: word-processing AI needs words, and not the electrical impulses in CPU.

As it was mentioned in [15], [16] the same is true for humans: from the moment of appearance of consciousness in humans, they associate themselves with their consciousness and not with their bodies or DNAs. When asked what they would prefer to be: a body without consciousness or a consciousness without a body, they obviously will choose the second option. Therefore, such "representational illusions" as consciousness and subjective

sensations are, in fact, more valuable for humans then initial biological reality. Nowadays, these illusions are becoming more valuable for AI as well (see above-mentioned example of word-processing AI) – this is one of indicators of appearance and evolution of mind-matter in computers [15], [16].

7 Final conclusions

In brief, we can answer the questions in the title of this article as follows:

Will artificial intelligence have consciousness? Yes, sure, it is acquiring consciousness right now. This consciousness is consciousness as it is defined in mind-matter theory (the model of the world, helping to optimize one's behavior in it [15], [16]).

Will artificial intelligence become subjects, having qualia and subjective experiences? We will never, in principle, know.

References

1. Jackson, Frank. Epiphenomenal Qualia [Text] / Jackson, Frank // Philosophical Quarterly. – 1982. – 32 (127). – 127–136. – doi:10.2307/2960077. JSTOR 2960077.
2. Jackson, Frank. What Mary Didn't Know [Text] / Jackson, Frank // Journal of Philosophy. – 1986. – 83 (5). – P. 291–295. – doi:10.2307/2026143. JSTOR 2026143. *There's Something*
3. Dennett, Daniel. Consciousness Explained [Text] / Dennett, Daniel. – Boston: Little, Brown and Co. 1991. – ISBN 978-0-316-18065-8. OCLC 23648691.
4. Dennett, Daniel. What RoboMary Knows [Text] / Dennett, Daniel // Alter, Torin (ed.). Phenomenal Concepts and Phenomenal Knowledge. – Oxford Oxfordshire: Oxford University Press. – 2006. – ISBN 978-0-19-517165-5. OCLC 63195957. Retrieved December 2, 2009.About Mary (2004)
5. Churchland, Paul M. "Reduction, Qualia, and the Direct Introspection of Brain States [Text] / Churchland, Paul M. // The Journal of Philosophy. – 1985-01-01. – 82 (1): 8–28.
6. Maloney, J. Christopher About being a bat [Text] / Maloney, J. // Australasian Journal of Philosophy. – 1985-03-01. –63 (1): 26–49. – doi:10.1080/00048408512341671. ISSN 0004-8402.
7. Thompson, E. Colour Vision [Text] / Thompson, E. – 1995.
8. Lycan, William G., ed.. Mind and cognition: a reader [Text] / Lycan, William G., ed.. – Cambridge, Massachusetts, USA: Basil Blackwell. – 1990-01-01. – ISBN 978-0631160762.
9. Lewis, David. Philosophical Papers Volume I - Oxford Scholarship [Text] / Lewis, David. – Oxford University Press. – 1983-08-18. – doi:10.1093/0195032047.001.0001. ISBN 9780199833382.
10. Churchland, Paul M. A neurocomputational perspective: the nature of mind and the structure of science [Text] / Churchland, Paul M. – Cambridge, Massachusetts: MIT Press. – 1989-01-01. – ISBN 978-0262031516.
11. Conee, Earl. Phenomenal knowledge [Text] / Conee, Earl //Australasian Journal of Philosophy. – 1994-06-01. – 72 (2): 136–150. – doi:10.1080/00048409412345971. ISSN 0004-8402.
12. Flanagan, Owen J. Consciousness reconsidered [Text] / Flanagan, Owen J. – Cambridge, Massachusetts: MIT Press. – 1992-01-01. – ISBN 978-0262061483.
13. Alter, Torin. A Limited Defense of the Knowledge Argument [Text] / Alter, Torin // Philosophical Studies. – 1998. –90 (1): 35–56. – doi:10.1023/a:1004290020847. JSTOR 4320837. S2CID 170569288.
14. Marcello Ghilardi. The Place of Subject and Absolute in Zen Buddhism [Text] / Marcello Ghilardi. – January 2015 Teoria. – 35(1):195-206.
15. Мищенко А. В. Тенденции развития планетарного интеллекта [Электронный ресурс] / Мищенко А.В. // Искусственный интеллект. – 2003. – No 4. – URL: http://iai.dn.ua/public/JournalAI_2003_4/Razdel5/02_Mishchenko.pdf
16. Mishchenko A. Computer Modeling of the Evolution of Civilization within the Futurological Theory of Mind-matter [Text] / Mishchenko A. // Problems of Artificial Intelligence. – 2021. – No 3(22). – URL: http://paijournal.guiaidn.ru/download_pai/2021_3/1_Мищенко.pdf
17. Мищенко А.В. Индульгенция людей [Электронный ресурс] / Мищенко А.В. – URL: https://AlesMishchenko.github.io/indulgencia
18. Unconscious determinants of free decisions in the human brain [Text] / Soon, Chun Siong; Brass, Marcel; Heinze, Hans-Jochen; Haynes, John-Dylan // Nature Neuroscience. 2008. – 11 (5): 543–5.

RESUME

A. V. Mishchenko

Agnostic Epiphenomenalism and Qualia-Reading Problem: Will Artificial Intelligence be a Subject, Having Consciousness and Subjective Experiences?

The question of whether artificial intelligence can have subjective experiences and human-type consciousness is important both for philosophy and the future of computer science. One of the most popular philosophical thought experiments, posing the question of whether subjective experiences can be modeled and simulated (for example, by AI), without actually feeling them, is the Frank Jackson's knowledge argument (also known as "Mary's room") [1], [2].

This experiment describes Mary, living in a black and white room, but studying everything about color, including its physical properties and the neurophysiology of color-processing in human brain. Will Mary know something new about color when she will leave this room? Will she acquire any type of new knowledge from her personal experience of color?

Although Jakson answers his question positively, a number of philosophers disagree, both including when Mary is a human (Christopher Maloney [6], Earl Conee [11], etc) and when she is a robot (Daniel Dennett [4]). In all cases, the vague formulation "everything about color" does not allow to prove any point of view.

The aim of this paper is to replace the knowledge argument by a more formal problem, allowing to distinguish, more formally, "the possible" from "the impossible" in studying the subjective experiences (either in human or, potentially, in AI).

This problem should allow both to formulate new philosophical viewpoint on subjective experiences and to reveal the structure of the entire "subjective world", according to the “theory of mind-matter” [16].

The proposed solution is to replace the knowledge argument by a more formal "qualia-reading problem”: if the observable activity of two brains are similar - can we be sure that these two subjects have similar subjective experience? The “agnostic epiphenomenalism”, formulated in this paper, suggests negative answer. For example, Mary will know everything about the color, including descriptions of subjective experiences of color of other people, but she, in principle, cannot be sure that, after leaving the room, her subjective experience of color will be as she expected.

This paper allows also to make the following conclusion about structure of subjective world (for both human and artificial intelligence): it is constructed from “representational illusions”, which, according to the “theory of mind-matter”, are becoming more important then underlying reality.

РЕЗЮМЕ

А. В. Мищенко

Агностический эпифеноменализм и проблема считывания квалии: будет ли искусственный интеллект субъектом, обладающим сознанием и субъективными переживаниями?

Вопрос о том, может ли искусственный интеллект иметь субъективные переживания и человеческое сознание, важен как для философии, так и для будущего информатики. Одним из самых популярных философских мысленных экспериментов, ставящих вопрос о том, могут ли субъективные переживания быть смоделированы и симулированы (например, в ИИ), без «самого их переживания», является эксперимент “комната Марии“, сформулированный Ф. Джексоном.

Этот эксперимент описывает Марию, живущую в черно-белой комнате, но изучающую все о цвете, включая его физические свойства и нейрофизиологию обработки цвета в человеческом мозге. Узнает ли Мария что-то новое о цвете, когда она покинет эту комнату? Приобретет ли она какие-либо новые знания из своего личного опыта восприятия цвета?

Несмотря на то, что Джексон отвечает на свой вопрос положительно, ряд философов не согласны с ним, в том числе, когда Мэри является человеком (К. Мэлони [6], Э. Кони [11] и т.д.) и когда она - робот (Д. Деннетт [4]). Во всех случаях, расплывчатая формулировка «Все о цвете» не позволяет доказать какую-либо точку зрения.

Цель данной статьи состоит в том, чтобы заменить эксперимент Джексона более формальным вопросом(проблемой), позволяющей отличить «возможное» от «невозможного» в изучении субъективного опыта (как у человека, так и, потенциально, у ИИ). Эта проблема должна позволить как сформулировать новую философскую точку зрения на субъективные переживания, так и раскрыть структуру всего «субъективного мира», в соответствии с теорией мыслящей материи [16].

Предлагаемое решение состоит в том, чтобы заменить эксперимент Джексона более формальной “проблемой прочтения квалии“: если наблюдаемая активность двух мозгов схожа - можем ли мы быть уверены, что эти два субъекта имеют сходный субъективный опыт? Сформулированный в статье «агностический эпифеноменализм» дает отрицательный ответ. Например, Мария будет знать все о цвете, включая описания субъективных переживаний других людей, но она, в принципе, не может быть уверена, что, выйдя из комнаты, ее субъективное переживание цвета будет таким же, как она ожидала.

Данная статья позволяет также сделать следующий вывод о структуре "субъективного мира" (как для человека, так и для ИИ): он построен из «репрезентативных иллюзий», которые, согласно «теории мыслящей материи», становятся более важными, чем лежащая в их основе реальность.

Статья поступила в редакцию