




<https://doi.org/10.22363/2313-2302-2022-26-4-740-754>

Research Article / Научная статья

## Philosophy and Science on the Way to Knowing and Making Consciousness

Igor F. Mikhailov  

Institute of Philosophy, Russian Academy of Sciences,  
12/1, Goncharnaya Str., Moscow, 109240, Russian Federation  
 ifmikhailov@gmail.com

**Abstract.** The latest progress in empirical studies of consciousness and spectacular advances in AI technologies kick philosophy out of the familiar comfort of uncontrolled proliferation of concepts and scholastic disputes. In the overview of the current state of empirical theories of consciousness, author reveals that those theories still find themselves in the pre-paradigmatic stage, therefore not yet posing an immediate existential threat to the philosophy of consciousness, though making it watch out. Author attempts to deal with the certain ambiguity of the term ‘consciousness’, stripping its meaning from parts already susceptible to science and technology and from parts still highly unlikely to be explained away. Besides, the relationship between philosophy and science is specified in general by analyzing them to their inner dynamics of theories and ontologies, showing that for science, the distinction between the two is substantially more important than for philosophy. From this perspective, philosophical schemas of consciousness claiming to be ‘experiential’ must have met recently formulated criteria for empirical theories of consciousness, otherwise failing to explain anything in the domain. Finally, author adds his pragmatic criterion that addresses the technological perspectives a theory provides. In the end, a winning competitive theory will have to let us produce and control artificial conscious devices.

**Keywords:** empirical theories of consciousness, intelligence, awareness, control, qualia, phenomenal consciousness, high order thought, information integration theory, global neuronal workspace, unconscious priming, intentionality, representation

### Article history:

The article was submitted on 15.07.2022

The article was accepted on 23.08.2022

**For citation:** Mikhailov IF. Philosophy and Science on the Way to Knowing and Making Consciousness. *RUDN Journal of Philosophy*. 2022;26(4):740—754. <https://doi.org/10.22363/2313-2302-2022-26-4-740-754>

### Introduction

These days, trying to ultimately cope with painful riddles of what makes us feel and think, we are met face-to-face with a whole lot of theories of consciousness (ToC), some of them relying on the brain circuitry, which constantly slips away

©



This work is licensed under a Creative Commons Attribution 4.0 International License  
<https://creativecommons.org/licenses/by-nc/4.0/legalcode>

from any final explanations, while others focusing on our phenomenal ‘insides’ in the certainty that this is the thing. (There is, however, a minority that is ready to add the subject to Emile Dubois-Reymond’s list of *ignoramus et ignorabimus* but let us constrain ourselves with what may be discussed any further.) One of the popular views is that theories of the first kind must be considered scientific or empirical<sup>1</sup>, while the second group is philosophical. This distinction is worth a quick look.

Two of the prominent students of mind and consciousness take on the distinction of scientific and metaphysical ToC: “the former typically at least tacitly assumes materialism and aims at explanation and mechanistic approaches to consciousness, whereas the latter is concerned with the ultimate nature of consciousness rather than with specifics about neuronal mechanisms” [4]. I suppose that this distinction is not quite accurate, as Hohwy and Seth attribute a tacit metaphysical stance to empirical students, which may not be universally acknowledged. On the other hand, there are explicit materialists among philosophers (*e.g.*, Paul Churchland) for whom there is nothing to consciousness except for the neural mechanisms. Probably, the ultimate difference between metaphysics and science is not in the ‘what’- but in the ‘how’-approaches. This difference is better grasped by Gerhard Schurz in saying: “while typical speculations postulate for each new phenomenon a new kind of theoretical cause, good science introduces new theoretical entities only if they figure as common causes of several intercorrelated phenomena” [5. P. 108]. But even this formula is not effective enough to guarantee the demarcation in all the cases, as Hegel-style speculative philosophies excel in inventing universal causes, while science is not barred from identifying specific causes for specific instances. I would propose an even more delicate analysis based on the distinction between a theory and an ontology [6]. Every theory, except formal ones in the proper sense of the word, needs an ontology, which is the same as a model for its interpretation. And this is what likens philosophy and science. The difference starts in attributing truth values and in what I would call the ‘finalizing’ of a study. A Hegel-style speculative philosopher would say that (1) there is an Absolute Spirit that generates all the rest of what there is, and (2) this proposition is true. A neighboring scientist, committed to the same ontology, would try to elaborate a system of measurements for the activity of the said Spirit to formulate its universal laws in their most precise form possible, for them to be testable against some measured observables. And only propositions of these laws and their corollaries may be evaluated as true or false. And only these propositions make up what may be referred to as a theory proper, unlike the ontological commitments serving as its interpretation.

Here starts another distinction, which is between scientific realism and instrumentalism. The former claims that ontological commitments somehow inherit the truth-value of a theory interpreted on them, which is to say that an ontology of

---

<sup>1</sup> For overviews of empirical theories of consciousness see Ned Block on some of the most important empirical theories [1], a perspective of neuroscientific advancements [2] and maybe the latest analysis of the situation in this realm [3]. Some of them will be approached further herein.

a theory considered being true is the true picture of the world. The latter, probably focusing on the real history of science, are more cautious in that respect.

As far as I can tell, this discussion implies that one of the main objectives of non-Hegel-style philosophy is to determine, if ‘consciousness’ refers to an ontological commitment at the risk of vanishing soon, like ‘caloric’ and ‘phlogiston’ did, under the pressure of well-measured and operationalized scientific models, or it stands for a real entity that threatens the causal closure of the physical world<sup>2</sup>.

I am inclined to let the second option alone, for I don’t know what to do with it. As for the first one, it presumes that to scientifically *explain* anything non-physical means to ‘expel it plain’ if I may say so, in plain coherence with Wittgenstein’s “<the solution of the problem of life is seen in the vanishing of the problem” [7. P. 250]. But, once the ghost is expelled from the machine, the question remains of what drives the machine and how it is set up. The actual science offers two principal ways. Physics, chemistry, biology, and genetics incline us to believe that mind and consciousness are emergent effects of the extremely complex interplay of multi-level natural forces, which gives us quantum, chemical and genetic conjectures on the nature of the phenomenon. Meanwhile, mathematics and computer science tend to elaborate a kind of simpler explanations presuming that parts of mental mechanisms impact each other with their abstract structural properties that line up in functional dependencies, which turns one part into a representation of another. This makes the whole thing an information-processing, *i. e.* computational, system.

The computational approach, while being simpler, rises a bunch of methodological problems discussed in thousands of publications nowadays. But its possible outcome is that what makes us capable of thinking, feeling, and taking account may be reproduced in other material systems with similar computational capacities, and we will be able to create as many conscious companions for ourselves as we probably need.

Based on this clarification of philosophy/science relations, I am inclined to pronounce in favor of methodological naturalism<sup>3</sup>. This approach consists, first, in understanding the task of philosophical research as the generation and adjustment of scientific (domain) ontologies. Since ontologies in this sense are conceptual schemes or abstract models on which the provisions of scientific theories are interpreted, meta-ontology [9], to which philosophical research is mainly dedicated, is equivalent to conceptual analysis. The latter was considered the only possible way of philosophical research by later positivists.

Methodological naturalism consists, secondly, in disputing the understanding of philosophy as being immune against empirical data and scientific concepts, *i. e.*, in proclaiming the not quite a priori character of philosophical propositions. The

---

<sup>2</sup> It is worth noting that ‘the physical world’ refers to some integrated ontology of the current physical theories considered to be true.

<sup>3</sup> For the distinction between physicalism and naturalism see, *e. g.*, [8].

most famous proponent of this view has been W. V. Quine. Philosophy, of course, does not have exact procedures for empirical verification or falsification, but it cannot ignore what is happening in the sciences.

With these observations in mind—and in consciousness as well—I will try to analyze some of the current tendencies and approaches to building successful empirical ToC.

### What is consciousness exactly?

Far from referring to an exact scientific concept, the term ‘consciousness’ rather stands for one of the few subject-matters resistant to any attempts of strict definitions. Highlighted by Wittgenstein, this class of ideas has not been labelled in any explicit manner, so I will speak of them as of *intractable notions*. They are usually fundamental for thought and culture and are very difficult or impossible to define discursively: ‘game’, ‘culture’, ‘consciousness’, ‘knowledge’, ‘time’, ‘space’, ‘number’, *etc.* The principal shortcoming of Wittgenstein’s ‘family resemblances’ concept [10. P. 32e] is that it doesn’t set any limits to the scopes of intractable notions: why, for example, does not everything in the Universe become a game when every instance of the ‘game’ is tangled with any other one only by some contingent non-universal property? It seems impossible to bound the concept of non-game in this case. One possible explanation may be that there is a universal common property of the instances of a class, although not of a substance-attribute kind, but of procedural (algorithmic) one: ‘time’ covers all possible actions with duration; ‘game’ is a term for any regular, though unconcerned, behavior, when the brain rewards practicing the rule-following with dopamine; ‘consciousness’ stands for all purposeful accountable actions combined with possibility of making non-standard decisions. If my assumption is correct, then ‘consciousness’ refers to some procedural properties of mind and behavior.

Still, we face the urge to define what exactly we mean by ‘consciousness’: (I) intelligence, (C) control, including the ability to account (awareness), or (Q) qualia, *a. k. a.* ‘phenomenal consciousness’. (I) does not cause serious problems in the theoretical and philosophical part of the cognitive sciences, while consistently and in many cases successfully finding technological embodiments. Besides, as Ned Block puts it, “<c>onsciousness and intelligence are on the face of it very different things. We all understand science fiction stories in which intelligent machines lack some or all forms of consciousness. And on the face of it, mice or even lower animals might have phenomenal consciousness without much intelligence” [1. P. 1112].

(C) is what is lacking while we are in deep sleep or with complete anesthesia and returns upon exiting these states. There are several competing approaches to the subject: the theory of the global workspace by Bernard Baars, the theory of integrated information by Giulio Tononi, and some others. Some day, one of these theories (or maybe another one still unknown) will win the competition and become paradigmatic. Anyway, this is what Chalmers could relate to as the ‘easy problems

of consciousness', so this issue can hardly be considered unsolvable for cognitive science either. Only (Q) claims the title of the 'hard problem' because, thanks to the scholastic efforts of philosophers like Frank Jackson, many are convinced that the existence of qualia (which is not contested by these philosophers, because, like Descartes, they believe in the a priori truth and reliability of their 'subjective experience') somehow contradicts the physicalist picture of the world. I see two possible ways out here: (Q<sub>1</sub>) scientific investigation of qualia using the 'small steps' approach — for example, starting with the biological sensitivity of the simplest organisms and reconstructing the evolutionary stairway from them to specialized 'modular' neurons and neural ensembles of a developed brain, etc., which does not guarantee success, but does not exclude it a priori as well, or (Q<sub>2</sub>) to recognize this problem as scientifically unsolvable.

But whatever is chosen, neither of these alternatives prevents science from studying C-consciousness independently from Q-experience, as the latter, while giving consciousness some unique and fascinating flavors does not account for what it is.

### ...And what is it not?

At the heart of the Cartesian *cogito*, which is mistakenly considered by some a starting point of the New Age philosophy of mind, lies the following logical move: my doubt in whatever existence is the guarantee of my own being there. However, this conclusion can be justified only if the following — mainly tacit — presuppositions are true.

*Assumption 1:* My consciousness is given to me as a thing-in-itself (my perception of my inner life cannot be false).

*Assumption 2:* Thinking (doubt) can only be a conscious act, and my account of my thought acts is necessarily reliable (unconscious distortions, biases, and attitudes are ignored).

Each of them can be the subject of reasonable philosophical suspicions.

The fallibility of introspection as a source for knowing consciousness, according to Kant, makes psychology as a science proper impossible [11. P. 8]. Nevertheless, classical phenomenology advocates the possibility of a transition from conscious experience to transcendental knowledge of mental properties. In Gallagher and Zahavi's interpretation, Husserl thought of phenomenology as an epistemological foundation for the sciences. Per his original project, it implied bracketing of the 'natural attitude' and rejection of the opposition of 'internal' and 'external', as well as the possibility of inferring from the 'givenness' of an object to the transcendental (or, at least, pre-experiential) structure of consciousness [12. P. 24—29].

Even though phenomenologists, starting from Husserl himself and including the likes of Merleau-Ponty, have insisted on the fundamental difference between phenomenological and psychological introspection, Gallagher and Zahavi define phenomenology as the study of the world in which we live from a new reflective

point of view, namely from the point of view of its meaning and manifestation for consciousness [12. P. 25]. More than once in their book, they speak of phenomenology as a theory based on experience [12. P. 6—7].

From my point of view, one of the main faults of phenomenology is the failure to distinguish the empirical from the experiential. The former refers mainly to semantics and justification, while the latter to the psychology of perception. For me, the proposition ‘Amazon Delta is situated in the Brazilian states of Pará and Amapá’ is an empirical one, although I never experienced the sight myself phenomenologically. If I did, my phenomenological impression could be taken as another supporting evidence for this empirical truth only as long as I knew in what way my perception would differ, had it been otherwise. This is how we learn whatever facts of the world. But phenomenologists claim that starting from the same simple experiences and applying a series of ‘reductions’ thereto, we may learn something of consciousness. This could be possible if we had the faintest idea of how those experiences would look if our consciousness was not the way it is. But we don't. There is a substantial epistemic asymmetry between a torch and what is seen in the spot of its light.

In short, being ‘based on experience’ is equivocal in the case of phenomenology. If it uses experience as a case for some sort of transcendental deduction, then it must demonstrate virtually Kantian rigor and the inevitability of its implications<sup>4</sup>, which it doesn't. Otherwise, if it claims to be an empirical theory proper, it must meet, firstly, the general criteria for empirical theories of consciousness and, secondly, the pragmatic criterion, all of which are discussed further. But, anyway, one gets an immediate impression that phenomenology begins with an ‘attitude’ reminiscent of the Kantian transcendental approach but proceeds to share all the faults of Hegelianism, such as optional conclusions and the claim to extra-scientific knowledge.

As for the phenomenal aspect of consciousness, the issue with it may be somewhat reminiscent of the Flying Arrow paradox attributed to Zeno of Elea. We have a continuum of neural activity, described by a dynamic state-space model operating with real numbers. Our communication system, based on the static semantics of discrete symbols, snatches arbitrarily fractional snapshots thereof. The remainder, not covered by fixed points of linguistic meanings, is perceived as something ineffable. Such an explanatory model excludes a special 'qualitative' side of consciousness, which, upon closer examination, turns out to be equal to those parts of the stream that, as it were, overflow the edges of the discrete linguistic structure with its systematicity, compositionality, and productivity [13].

### **Empirical theories of consciousness**

There are quite a few ToC listings in various reviews. Thus, Hohwy and Seth point out Global Neuronal Workspace Theory, Integrated Information theory,

---

<sup>4</sup> Which, in his case, had put him under severe critique with the further developments of science not envisaged in his ‘Critique’. But that is how honest intellectual enterprises work.

Recurrent Processing theory, Higher-Order Thought theories coupled with Metacognitive theories, Radical Plasticity thesis, Virtual Reality theories, Attention-based theories, Heterophenomenology, Core Consciousness theory, Orchestrated Objective reduction, and Electromagnetic theory [4]. Ned Block in an earlier review had focused on Higher Order Thought, Global Workspace, Integrated Information, and what he calls The Biological Theory dating back to the 1950ies, which occurs to be his preference [1. P. 1111—1113]. A more recent attempt to review and analyze ongoing approaches in empirical studies of consciousness has been made in [3], which I will touch on a bit later.

One of the common ideas brought about by experimental studies of conscious and unconscious perceptions is that perceptions that are beyond the threshold of conscious processing, nevertheless, affect other cognitions and behaviors of the subject. The following events were identified as objective neural correlates of conscious states (NCC is now a widely accepted abbreviation for these): a late increase in the corresponding sensory activity, cortical-cortical synchronization at beta and gamma frequencies over long distances, and the ‘ignition’ of a large-scale prefronto-parietal network. These data agree with the well-known Global Neuronal Workspace theory [2]. It has also been found that the global neuronal workspace (GNW) hypothesis can become the basis for the synthesis of empirical data regarding conscious access, attention, and working memory [14]. Studies of patients under anesthesia have shown that consciousness disappears when anesthetics cause a functional shutdown in the posterior parietal region, interrupting cortical communications and causing loss of integration; or when they lead to bistable stereotyped responses, causing a loss of informational capacity. Thus, it seems most likely that anesthetics cause unconsciousness when they block the brain's ability to integrate information [15]. Studies of returning to consciousness have demonstrated that neural activity on the way to conscious states is limited to a low-dimensional subspace. In this subspace, neural activity forms discrete metastable states that persist for minutes. The network of transitions that links these metastable states is structured so that some states form hubs that connect groups of otherwise unrelated states. Although there are different paths through the network, to eventually enter a state of activity compatible with consciousness, the brain must first pass signals through these nodes in an orderly manner [16]. In general, the available literature on empirical studies of conscious states suggests that conscious states are characterized by a greater degree of connectivity between different parts of the brain and neural ensembles. That is, conscious and unconscious states do not form a dichotomy, but are characterized by gradual quantitative transitions, for which a rheostat might be a proper metaphor [17]. And even if some information has not become the subject of broadcasting and, therefore, a fact of consciousness, it can still be processed by the cognitive system and causally affect the state and behavior of the subject.

According to the said Global Neuronal Workspace theory (GNW) [18—20], unconscious processes and mental states compete for the center of attention from

which global information is broadcast throughout the system. Consciousness is identified with this global broadcasting and, according to Baars, is an important means of functional and biological adaptation. The closest competitor to GNW is the integrated information theory (IIT) by J. Tononi and C. Koch [21—25]. It is one of the few modern theories that offer a measurable indicator for the degree of consciousness in a system that is labeled as  $\Phi$  (phi), which shows the degree of physical integration of information. According to Tononi, this indicator can be measured even in relatively simple physical systems, which, of course, leads to a kind of panpsychist conclusion. Nevertheless, the theory is said to be well interpreted on some neurophysiological data coming from anaesthetical practices and other situations of transition between conscious and unconscious states.

Both theories are, in a sense, models, since they point to possible neurobiological realizations of the functions under consideration. However, there are also purely theoretical ideas that are not tied to specific implementation models. These include, for example, the theory of ‘Higher Order Thought’ (HOT) by David Rosenthal [26, 27]. According to it, the mental states of the lower order (LO) include sensations and perceptions caused by the impact on the sense organs of objects of the external world. Higher-order mental states (HO) have any other mental states as their object. The LO mental states become conscious only when they become the object of the HO mental states. This theory has some similarities with various approaches stating the recursive nature of individual consciousness [28—30].

### Criteria for ToC

In a comparatively recent paper entitled “Hard Criteria for Empirical Theories of Consciousness” [3] the following criteria for ToC have been proposed:

1. Whether a theory sticks to paradigm cases of consciousness and the unconscious alternative.

As is clarified therein, “<p>aradigm cases with an unconscious alternative ensure that consciousness is the dependent variable in experiments, and contrast with approaches where only conscious states are investigated” [3. P. 5].

2. Whether a ToC is free from being subject to the ‘unfolding argument’, which equals its adhering to a falsifiable causal structure.

Doerig et al. give an example of the Recurrent Processing Theory [31—33], according to which visual consciousness emerges when stimuli having passed through an initial feed-forward network (FFN) of the visual tract start being broadcasted via some recurrent networks (RN) that connect the visual processing regions to other parts of the brain. The problem with this and similar approaches is that, as is stated mathematically, any RN may be unfolded into an FFN, thus being functionally identical thereto. Moreover, patients whose damaged RN-regions have been replaced with FFN-implants experience no faults in their consciousness. Therefore, according to Doerig et al., the kind of theories under consideration are not falsifiable, which puts them outside the scope of science.



3. Whether a ToC is free from the small (and large) network argument.

This argument boils down to the problem of panpsychism and the unity of consciousness. If a theory lacks a quantitative criterion of a system being conscious, then it may imply that a network of, say, ten neurons may be conscious. If a theory implies that a small-size network is enough to produce consciousness, it is subject to two issues: one is panpsychism, and the other is the problem of the unity of consciousness in large-scale networks, such as the human brain.

4. Whether a ToC is resistant to the multiple realizations' argument.

Resorting to a computer metaphor of an application being implemented within different operational systems, the authors claim that “ToCs that explain consciousness by pointing to certain brain regions or characteristics claimed to be sufficient for consciousness need to explain why they are also necessary for consciousness. Hence, our fourth criterion asks whether ToCs can make clear-cut and specific predictions about which other systems are conscious, apart from humans” [3. P. 8].

The paper is interesting in its conclusions as well. Comparing ToC issues with those biologists are faced with regularly, the authors point out that the latter lack any rigid definition of life, which doesn't prevent them from knowing what life is as they associate it with a set of necessary processes, like homeostasis, reproduction, etc. The current situation with ToC, in their view, is like the one with magnetism in the science of ancient times. Researchers just look for known or unknown 'things' to identify them with their subject of interest. While, as Doerig et al. put it, it may well be that “consciousness is a ‘solution’, a by-product, or a core component of a computational challenge that information processing systems need to solve — and that we have not discovered yet” [3. P. 16].

I would add that in a future sequel to their seminal paper, somebody should examine how diverse species of the computational approach meet the four criteria.

### **The pragmatic criterion added**

From the point of the previous discussion, to know is to obtain a reliable theory coherent with some criteria that make a theory properly explanatory and predictive. From the point of view of social pragmatics, to know is to be able to make. But there may be a substantial difference in various kinds of making. We can make a clock to know what time is now— but do we know then what time is as such? We would, but only if we made a time machine. Being able to measure is not the same as being able to alter.

These days, we can make AI devices, primarily, in the form of neural networks, but we still cannot determine the level of their consciousness, let alone provide them with it. The present course of the AI development in a natural way leads to the idea of making not only just intellectual but rather in some sense conscious devices, which could make more effective decisions choosing certain circumstances and intentions to be taken into account and which would fit better within human communication.

There is a reason to believe that we are close to the scientific revolution, when — in Thomas Kuhn's terms — one or a small number of intensively proliferating and developing empirical theories of consciousness will win the competition to become paradigmatic and the successive period of 'normal science' will lead to the occurrence of pursued technologies.

Such a goal implies a certain renewal of philosophical and scientific approaches to the very idea of Artificial Intelligence. We, people, grounded on introspection and interpersonal empathy (also known as a 'theory of mind') traditionally judge from the fact that consciousness plays a critical role in searching and making optimal decisions. And it is natural to think that artificial intelligence devices, once turned conscious, will be more effective in some relations thereby. But as the very term 'consciousness' has historically fallen victim to various confusions and puzzles, scrupulous conceptual and methodological — *i. e.* philosophical — analysis must precede the related scientific research and technological development of conscious devices.

It was traditionally believed that consciousness is a privilege of humans, which, elevated them to the top of evolution and the food chain. However, the latest research in animal ethology and psychology suggests that consciousness is characterized not by absolute presence or absence, between which there is a clear qualitative threshold, but by quantitative gradations. And this consideration significantly complicates the task for both philosophers and experimental researchers: after all, no matter what metaphysical position they take — eliminative materialism or functionalism or any other — they must recognize either the identity of consciousness and its material correlates (the structures of the brain or body as a whole), or some form of causal dependence of the former on the latter, at least in the form of supervenience. And this means that it is hardly possible to identify natural objects and, accordingly, to technically reproduce a certain special isolated mechanism that is responsible for consciousness itself and ensures it. Rather, we can talk about complex distributed multifunctional structures, some (re)configuration of which produces consciousness as one of the systemic emergent effects, possibly quantifiable. And while the thinking of neuroscientists and AI specialists does not require a radical restructuring to approach this problem, since they all, to a certain extent, master the tools for probability and statistics, networks, and multi-dimensional state-spaces, philosophers, many of whom have lingered mainly in the realm of traditional Aristotelian and Boolean logic, are faced with the need to significantly update their conceptual analytical tools.

The very technological possibility to create artificial conscious devices has both pragmatic and scientific significance. The former is obvious: conscious devices are expected to become more efficient at finding and making optimal decisions. In addition, the possibilities of their communication with people and animals will expand significantly, and so will the scope of their applications: these can be services, care, education, bureaucratic formalities, etc. The scientific significance may be summed up as follows: to make is to understand. If we are

making atomic bombs, we do not doubt that we have a substantial understanding of the mechanism of a nuclear reaction. The same can be projected onto the problem of consciousness: the mass production of conscious devices will mean that there are no philosophical puzzles left in this area.

### Conclusion

Several problems are directly or indirectly determined by the foregoing, the solution of which is expected from any future philosophical attempts at the issues of Artificial Consciousness (AC).

#### *A. Are intellectual operations possible without awareness?*

The current abundance of intelligent albeit not conscious devices and programs speaks in favor of a positive answer. However, this is not obvious in the case of the human psyche. One needs to turn to the data of cognitive psychology, which would confirm or refute the initial assumption about the fundamental independence of these cognitive capacities. If this assumption is not confirmed, it will be to the benefit of AC skeptics, who, at the very least, strongly doubt the possibility of technically reproducing the human mind as it is.

#### *B. Is awareness possible without qualia?*

One of the standard approaches in the philosophy of mind, including analytic philosophy, is that qualia constitute the core feature of consciousness that, therefore, turns out to be unformalizable and, therefore, incomputable ‘residue’ of the activity of cognitive systems — the experience of ‘what it is like to be [in a certain state]’, which in this case probably cannot be technologically reproduced. This is another point of productive application of the conceptual competencies of philosophy given the future technological breakthrough: if the initial hypothesis about the fundamental independence of conscious control from the qualitative states of the subject is correct, then the problem of consciousness in the correct sense of the word is easily separated from Chalmers's ‘hard problem’ and becomes the subject of feasible technological solutions.

#### *C. Is the computational approach to consciousness justified?*

Computationalism in the cognitive sciences has a long history. The very project of unified cognitive science (CS) owes its existence to the so-called ‘computer metaphor’. Later discovery of the limitations of the classical symbolic CS caused many to become disillusioned with the computational approach as such, giving rise to numerous variants of anti-computationalism and post-cognitivism. If the computational approach may be saved by turning to some up-to-date concepts of computation, then we need yet to determine which algorithms can provide the functions of consciousness that we have already tentatively defined as control and reporting. That is, a conscious cognitive system, natural or artificial, must be able not only to present input data in a certain form and automatically perform algorithmic actions with them, but also to provide feedback on the results, the content of previous experience, and other available resources, so that, if necessary, change the algorithm. Presumably, of particular interest to the problem are recursive algorithms, parallel distributed computations, and statistical algorithms, especially

those related to predictive processing. Demonstrating the feasibility of this task at a conceptual level, if successful, will open the way to the actual design of artificial conscious agents.

*D. How do 'artificial consciousness' (AC), 'artificial intelligence' (AI), 'artificial life' (AL), 'artificial emotions' (AE), and 'artificial societies' (AS) compare?*

It is highly likely that, at the computational algorithmic level, AC and AI are different technologies, which, however, have some points of intersection. 'Artificial life' and 'artificial societies' are historically and genetically related concepts, originally based on the theory of cellular automata. However, AS has evolved into the technology of multi-agent systems (MAS), a subtype or reincarnation of which is the more modern research in the field of 'agent-based systems' (ABS). AE, being originally an aspect of AI and AL, gradually grows into an independent field of study, where true neurobiological mechanisms and functions of emotions are studied, as well as their impact on cognitive processes proper. If intelligence, consciousness, and inter-agent interactions are computational processes aimed at optimizing life processes, then this optimization task can be solved by them alternatively, depending on which algorithms — recursive control or parallelization of computations — are more efficient at a certain stage of life or under certain circumstances. Simply put, society and consciousness seem to compensate each other, reducing the computational load of each of them if necessary. This, to some extent, corresponds to what our everyday experience tells us: namely, that less cognitively loaded individuals, as a rule, are easier and more readily socialized. Confirmation of this hypothesis would mean that the technological developments in the field of AC, which claim to be successful, must be — at least in the long term — technologically compatible with all these related areas.

A positive answer to question (C) would significantly shorten the path from scientific theories to technological solutions. The discovery of ontological and methodological intersections with other 'artificial' studies, as in (D), would save the research on AC from starting from scratch. The proven generic identity of artificial and natural consciousness would allow for a prompt implementation of the results obtained here into the fields of psychology and neurobiology. Separating the problem of consciousness from the problem of qualia would significantly increase the chances for the feasibility of AC. And finally, the formulation of criteria and a philosophical assessment of the comparative prospects of AC architectures and algorithms would facilitate the mission of empirical researchers, to whom philosophy should eventually pass the baton.

## References

- [1] Block N. Comparing the major theories of consciousness. In: Gazzaniga MS, ed. *The Cognitive Neurosciences*. IV. Cambridge, MA: MIT Press; 2009. P. 1111—1123.
- [2] Dehaene S, Changeux JPJPP, Dehaene S, Changeux JPJPP. Experimental and Theoretical Approaches to Conscious Processing. *Neuron*. 2011;70(2):200—227. <https://doi.org/10.1016/j.neuron.2011.03.018>

- [3] Doerig A, Schurger A, Herzog MH. Hard criteria for empirical theories of consciousness. *Cognitive Neuroscience*. 2021;12(2):41—62. <https://doi.org/10.1080/17588928.2020.1772214>
- [4] Hohwy J, Seth A. Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philosophy and the Mind Sciences*. 2020;1(II). <https://doi.org/10.33735/phimisci.2020.II.64>
- [5] Schurz G. Structural correspondence, indirect reference, and partial truth: Phlogiston theory and Newtonian mechanics. *Synthese*. 2011;180(2):103—120. <https://doi.org/10.1007/s11229-009-9608-7>
- [6] Mikhailov I. Has Time of Philosophy Passed? *Voprosy filosofii*. 2019;(1):15—25. <https://doi.org/10.31857/S004287440003613-9>
- [7] Wittgenstein L. Tractatus logico-philosophicus. In: *Tractatus Logico-Philosophicus*. Anthem Press; 2021. P. 56—250. <https://doi.org/10.2307/j.ctv22d4t7n.8>
- [8] Vintiadis E. Why a Naturalist Should Be an Emergentist about the Mind. *SATS*. 2013;14(1):38—62. <https://doi.org/10.1515/sats-2013-0003>
- [9] Van Inwagen P. Meta-Ontology: A Brief Introduction. *Erkenntnis*. 1998;48(2/3):233—250. <https://doi.org/10.5840/wcp201999235>
- [10] Wittgenstein L. *Philosophical Investigations*. 3rd ed. Oxford: Blackwell Publishers Ltd; 1986.
- [11] Kant I. *Metaphysical Foundations of Natural Science*. Cambridge: Cambridge University Press; 2004.
- [12] Gallagher S, Zahavi D. *The Phenomenological Mind*. [2nd ed]. Routledge; 2013. <https://doi.org/10.4324/9780203126752>
- [13] Tacca MC. Syntactic Compositionality, Systematicity, and Productivity. In: Tacca MC. *Seeing Objects: The Structure of Visual Representation*. Paderborn: Brill, mentis; 2010. P. 37—52. [https://doi.org/10.30965/9783969751190\\_005](https://doi.org/10.30965/9783969751190_005)
- [14] Mashour GA, Roelfsema P, Changeux JP, Dehaene S. Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron*. 2020;105(5):776—798. <https://doi.org/10.1016/j.neuron.2020.01.026>
- [15] Alkire MT, Hudetz AG, Tononi G. Consciousness and anesthesia. *Science*. 2008;322(5903):876—880. doi:10.1126/science.1149213
- [16] Hudson AE, Calderon DP, Pfaff DW, Proekt A. Recovery of consciousness is mediated by a network of discrete metastable activity states. *Proceedings of the National Academy of Sciences of the United States of America*. 2014;111(25):9283—9288. <https://doi.org/10.1073/pnas.1408296111>
- [17] Arp R. Consciousness and Awareness. Switched-On Rheostats: A Response to de Quincey. *Journal of Consciousness Studies*. 2007;14(3):101—106.
- [18] Baars B. J. *Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press, 1993. 446 p.
- [19] Baars BJ, Franklin S, Ramsoy TZ. Global workspace dynamics: Cortical “binding and propagation” enables conscious contents. *Frontiers in Psychology*. 2013;4(200). Accessible from: <https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00200/full>. <https://doi.org/10.3389/fpsyg.2013.00200>
- [20] Boly M, Seth AK, Wilke M, et al. Consciousness in humans and non-human animals: Recent advances and future directions. *Frontiers in Psychology*. 2013;4(625):1—20. <https://doi.org/10.3389/fpsyg.2013.00625>
- [21] Tononi G, Boly M, Massimini M, Koch C. Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*. 2016;17(7):450—461. <https://doi.org/10.1038/nrn.2016.44>


- [22] Tononi G. Consciousness as integrated information: A provisional manifesto. *Biological Bulletin*. 2008;215(3):216—242. <https://doi.org/10.2307/25470707>
- [23] Tononi G. Integrated information theory of consciousness: an updated account. *Archives italiennes de biologie*. 2012;150(4):293—329. <https://doi.org/10.4449/aib.v149i5.1388>
- [24] Mayner WGP, Marshall W, Albantakis L, Findlay G, Marchman R, Tononi G. PyPhi: A toolbox for integrated information theory. *PLoS Computational Biology*. 2018;14(7): e1006343. <https://doi.org/10.1371/journal.pcbi.1006343>
- [25] Edlund JA, Chaumont N, Hintze A, Koch C, Tononi G, Adami C. Integrated information increases with fitness in the evolution of animats. *PLoS Computational Biology*. 2011;7(10):e1002236. <https://doi.org/10.1371/journal.pcbi.1002236>
- [26] Rosenthal DM. A theory of consciousness. In: Block N, Flanagan OJ, Guzeldere G, eds. *The Nature of Consciousness: Philosophical Debates*. Cambridge MA: MIT Press; 1997.
- [27] Rosenthal DM. Consciousness and its function. *Neuropsychologia*. 2008;46(3): 829—840. <https://doi.org/10.1016/j.neuropsychologia.2007.11.012>
- [28] Vergauwen R. *Consciousness, recursion and language*. In: Lowenthal F, Lefebvre L, editors. *Language and Recursion*. New York: Springer; 2014. [https://doi.org/10.1007/978-1-4614-9414-0\\_13](https://doi.org/10.1007/978-1-4614-9414-0_13)
- [29] Corballis MC. The Recursive Mind: The Origins of Human Language, Thought, and Civilization. *Journal of Multilingual and Multicultural Development*. 2011;33(3):319—321. <https://doi.org/10.1080/01434632.2012.656976>
- [30] Baryshnikov PN. Language, brain and computation: from semiotic asymmetry to recursive rules. *RUDN Journal of Philosophy*. 2018;22(2):168—182. <https://doi.org/10.22363/2313-2302-2018-22-2-168-182>
- [31] Lamme VAF, Roelfsema PR. The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*. 2000;23(11):571—579. [https://doi.org/10.1016/S0166-2236\(00\)01657-X](https://doi.org/10.1016/S0166-2236(00)01657-X)
- [32] Lamme VAF. Visual Functions Generating Conscious Seeing. *Frontiers in Psychology*. 2020;11:e83. <https://doi.org/10.3389/fpsyg.2020.00083>
- [33] Lamme VAF. How neuroscience will change our view on consciousness. *Cognitive Neuroscience*. 2010;1(3):204—220. <https://doi.org/10.1080/17588921003731586>

#### About the author:

Mikhailov Igor F. — Doctor in Philosophy, Senior Research Fellow, Institute of Philosophy, Russian Academy of Sciences, Moscow, Russia (e-mail: ifmikhailov@gmail.com). ORCID 0000-0001-8511-8849

## Философия и наука на пути к познанию и созданию сознания

И.Ф. Михайлов  

Институт Философии, Российская академия наук,  
Российская Федерация, 109240, Москва, ул. Гончарная, д. 12/1  
 ifmikhailov@gmail.com

**Аннотация.** Последние достижения в эмпирических исследованиях сознания и впечатляющие достижения в области технологий искусственного интеллекта выбивают философию из привычного комфорта бесконечно множющихся концепций и схоластических споров. В обзоре современного состояния эмпирической теорий сознания показано,

что эти теории пока находятся на до-парадигмальной стадии, поэтому ещё не представляют непосредственной экзистенциальной угрозы для философии сознания, хотя и заставляют её быть начеку. Автор предлагает попытку совладать с хорошо знакомой многозначностью термина «сознание», отделив те аспекты его значения, которые уже покоряются науке и технике, от тех, которые пока ещё не поддаются объяснению. Кроме того, в статье предлагается анализ внутренней динамики теорий и онтологий, который позволяет уточнить отношения между философией и наукой в целом и показывает, что для науки различие между этими элементами существенно важнее, чем для философии. С этой точки зрения философские концепции сознания, претендующие быть «основанными на опыте», должны соответствовать недавно сформулированным критериям эмпирических теорий сознания, чтобы действительно что-то объяснять в этой сфере. Наконец, автор добавляет свой прагматический критерий, учитывающий технологические перспективы, предлагаемые теорией. В конце концов, победившая в конкуренции теория должна позволить нам производить искусственные сознательные устройства и управлять ими.

**Ключевые слова:** эмпирические теории сознания, интеллект, осознание, контроль, квалиа, феноменальное сознание, мышление высшего порядка, теория интегрированной информации, глобальное нейронное рабочее пространство, бессознательный прайминг, интенциональность, репрезентация

**История статьи:**

Статья поступила 15.07.2022

Статья принята к публикации 23.08.2022

**Для цитирования:** *Mikhailov I.F.* Philosophy and Science on the Way to Knowing and Making Consciousness // Вестник Российского университета дружбы народов. Серия: Философия. 2022. Т. 26. № 4. С. 740—754. <https://doi.org/10.22363/2313-2302-2022-26-4-740-754>

**Сведения об авторе:**

*Михайлов Игорь Феликсович* — доктор философских наук, старший научный сотрудник, Институт философии Российской академии наук, Москва, Россия (e-mail: [ifmikhailov@gmail.com](mailto:ifmikhailov@gmail.com)). ORCID 0000-0001-8511-8849