

(Column host LIU Tingting)

DOI: 10.20098/j.cnki.1001-5132.2024.1121

Towards conscious artifacts: ChatGPT is the key, but to open the door, one must find the lock

SAMSONOVICH Alexei V^{1,2}

(1.National Research Nuclear University MEPhI, Moscow 115409, Russian Federation;

2.George Mason University, Fairfax, Virginia 22030-4444, USA)

Abstract: The idea to replicate in a machine human consciousness, understood as an immediate first-person subjective experience, becomes one of the greatest challenges of our time. Scientific understanding of the nature of phenomenal consciousness may require an expansion of the scientific method itself. While this resolution is postponed, the task of reproducing the consciousness phenomenon in a computer should be approached within the framework of functionalism. Proposed so far theories of this sort appear problematic. An approach to implementing humanlike consciousness is proposed here based on BICA (biologically inspired cognitive architectures) combined with LLM (large language models). The concept is illustrated by example of a virtual tutor. It is argued that final solutions can be obtained in a purely neuromorphic form, which will enable their further adaptation, growth, and evolution. We start from discussion of the ontological problem of consciousness and its temporary nonreductive solution, which makes a scientific study possible. The main hypothesis for the proposed study is formulated based on the BICA Challenge in terms of the eBICA cognitive architecture. Only the limited case of socially emotional consciousness is analyzed as example. It is explained how LLM, being not conscious on their own, accessible via ChatGPT, can be connected to eBICA, used to build a conscious prototype and validate its claimed property. Selected example of a virtual tutor further illustrates how this technology will allow the tutor to develop human-level

Received date: 2024-11-22.

JOURNAL OF NINGBO UNIVERSITY (NSEE): <http://journal1g.nbu.edu.cn/>

Foundation item: Supported by Russian Science Foundation (22-11-00213).

Biography: SAMSONOVICH Alexei V, professor, specializing in: cognitive architectures and semantic technologies. Email: AVSamsonovich@mephi.ru

teacher-student relationships that will increase the efficiency of tutoring. Future prospects for the proposed direction of research include a variety of sentient agents to be created for the many practical human needs with the possibility of their further autonomous evolution.

Key words: machine consciousness; the Hard Problem; BDI; BICA; LLM; virtual tutor

CLC number: TP18

Document code: A

Article ID: 1001-5132 (2025) 02-0001-16

1 Introduction

Reasoning about consciousness seems easiest, but in fact, it is difficult. Moreover, it is a thankless task, because only very few are rewarded for it. For most researchers, the question of consciousness lies beyond the scope of their professional capacities. Nevertheless, interest in this topic is notably growing. This fact is primarily due to the development of Artificial Intelligence (AI). Today, artifacts are emerging that are approaching humans in terms of overall levels of social intelligence and behavior. For example, new models of ChatGPT exhibit the ability to perceive and understand complex human emotions in nontrivial social situations, respond to them adequately, and prove helpful in practical tasks^[1]. Should we treat such agents as conscious beings? If not, then when and on what basis can we say that they have acquired this status^[2]? And most importantly, how can it be achieved, and how can it be prevented? As pointed out by many, these questions have moral and ethical implications for the society and therefore require a serious analysis.

In this context, three interrelated topics are of interest: the ontological problem of consciousness, functional models of consciousness, and the task of embodying consciousness in a computer. They are considered below in order.

2 The ontological problem of consciousness and its temporary solution

There are many approaches to defining the concept of consciousness. Accordingly, there are many notions

corresponding to this word. The most interesting among them is consciousness as the immediate first-person experience, as the “view from within” the direct sensation of qualia. This is the notion of consciousness that will be considered further in this work. It was detailed by Chalmers^[3].

Understood in this sense, consciousness is a phenomenon that is (believed to be) familiar to every normal adult; as a phenomenon, the presence of which we experience daily, and therefore there should be no doubt about its existence. Moreover, consciousness is the only thing accessible to us directly: everything else is perceived through the prism of consciousness. However, we encounter difficulties when trying to prove its presence in a particular object, including ourselves. The problem is that in reasoning, we try to proceed from objective facts, and they turn out to be insufficient.

Chalmers^[3] showed that the phenomenon of consciousness, understood as first-person subjective experience, can only be accessible to the subject associated with it and cannot be registered by any physical measurement device. Moreover, there is no way to prove or disprove its existence in a particular case based on objective data. This does not mean the impossibility of “reading” thoughts or feelings using neurophysiological measurements. Indeed, such capabilities exist today^[4]. It is only impossible to prove the status of the read thoughts and feelings as authentic first-person experiences.

To prove his assertion, Chalmers used the well-known “zombie argument”, namely, the logical possibility of the existence of a physical copy of a person devoid of consciousness understood as a “view from within”. Consider an example. We would not

attribute consciousness to a book we are reading, even though it reflects thoughts and feelings. We believe that the book (i.e., a stack of paper sheets) is incapable of experiencing anything. Similarly, in the brain of Chalmers' zombie, thoughts and feelings may be represented and not experienced by the zombie. In zombie brain the same physiological processes may occur as in the brain of a conscious person, causing the same behavior, including verbal reports, and yet lacking consciousness.

Some philosophers object to Chalmers, saying that the ability to imagine something proves nothing. But this is not an empty fantasy. There is a well-known analogy known in psychiatry as dissociative identity disorder^[5]. In this case, the same processes occur in the brain as in a normal conscious state, but the subject is not aware of anything and subsequently remembers nothing. Of course, this is not a pure case of Chalmers' zombie, since the two states can still be physiologically distinct, although hypothetically, we could assume for the sake of the argument that they are not. Then we would again come to the conclusion that the fact of the subject's consciousness (understood in Chalmers' sense) cannot be established using physical measurements, or any approach based on a third-person perspective. The point here is that the same thought processes in the brain can be conscious or unconscious: from a third-person perspective their status is ambiguous.

This ambiguity has been termed the "Hard Problem"^[3]. To resolve it, the principle of supervenience was proposed, which boils down to postulating that Chalmers' zombies do not exist in nature, although they could exist. This is done without discussing any mechanisms ensuring the constraint. At the same time, the question remains open as to which substrates, besides the human brain, can support consciousness. To resolve it, another postulate is introduced: the principle of organizational invariance (POI), according to which systems with the same functional and informational organization have identical characteristics of their consciousness regardless of how and on what basis these systems are implemented^[3]. It remains unclear how to

deal with systems that are lower than humans in their level of organization, e.g., lower animals. While eschewing this question, note that one of Chalmers' proposed answers to it was panpsychism. According to this doctrine, any physical object or system possesses consciousness, but to varying degrees. This answer raises even more questions. Therefore, we limit consideration here to human-level consciousness only.

Chalmers' viewpoint has been criticized many times since Dennett (e.g., by Dubrovsky^[6]) and is currently not recognized as an accepted scientific theory, although many well-known philosophers (Block, Goldman, Nagel, Palmer, Rosenthal, Searle, Tye, Van Gulick, to list alphabetically only a few) share certain aspects of Chalmers' view or reason similarly^[7]. The main reason for rejecting Chalmers' viewpoint as a scientific theory is that the modern scientific method is based on materialism, while the principle of supervenience takes us beyond materialism^[8-9]. The scientific method requires reasoning about consciousness from a third-person perspective, while the phenomenon itself is only observable from a first-person subjective perspective. Consequently, the scientific method turns out incompatible with Chalmers' theory. This conclusion does not solve the problem, but merely postpones its resolution. It calls for an extension of the scientific method itself^[10], which does not look feasible in a near future.

The materialistic view of the nature of consciousness is essentially represented by the identity theory or reductionism^[11-12], according to which all mental phenomena are nothing but physical phenomena occurring in the brain^[8]. From this perspective, the "hard" question about their authenticity as mental phenomena loses meaning. Functionalism^[11] goes even further in this direction and abstracts from the physical substrate, being interested only in the functional organization of the system and its interactions. The concepts discussed above, including consciousness and POI, still can be used within a functionalist framework without problems, if defined functionally and treated as

abstract notions.

This approach can be a temporary solution to the problem, acceptable from a scientific point of view. Indeed, for a scientific study of consciousness based on this approach, we do not need to appeal to intuition, suggesting that each conscious being has a “soul” or a “view from within”. It is sufficient to identify those definitive aspects of the human brain’s functionality that correlate with the clinical fact of clear consciousness and look for their equivalents in other systems. The task of implementing an artificial conscious system then reduces to the task of implementing this identified functionality on a non-biological basis.

Probably this is the best we can do today: postpone solving the Hard Problem for the future. For now, a temporary solution is to take for granted the fact of the existence of non-reductionist phenomenal consciousness, as well as the principles of supervenience and POI applicable to it. Then, without abandoning the concept of phenomenal consciousness, but based on generally accepted documented cases of its presence or absence, seek the necessary and sufficient conditions for a substrate of consciousness within the framework of functionalism.

Physics does not explain the fact of the world’s existence^[13] but proceeds from it. Similarly, a scientific theory of consciousness should proceed from the fact that some systems possess human-level phenomenal consciousness in the sense of subjective first-person experience, which is not reducible to a set of physical characteristics. The task of the science of consciousness is to describe objectively the general properties and laws of such systems in terms of their functional and informational organization.

Thus, the task of implementing a conscious artifact can be formulated and solved according to the following plan.

(1) Identify general functional requirements for the substrate of consciousness based on the analysis of functional correlates of documented examples of systems capable of possessing consciousness.

(2) Build a minimal functionalist model of a

system sufficient as a substrate of consciousness based on the criteria obtained in (1).

(3) Implement a prototype of this system. Through empirical study of its behavior in realistic social environments, confirm its ability to gain general recognition in the community as a conscious subject.

Regarding point (1), the rich baggage of knowledge and facts should be borrowed from experimental psychophysiology, cognitive neuroscience, clinical psychiatry, and other studies of consciousness. This knowledge makes it possible to formulate functional models of conscious systems and to identify correlates of consciousness expressed in terms of these models. In contrast, modern approaches in data science and AI usually disregard scientific domain knowledge, which seemingly makes modern AI approaches incompatible with the above plan. Therefore, it looks like currently popular large language models (LLM), including the GPT line, do not possess consciousness^[2]. Can they still help us in making machines conscious? We shall return to this question later.

Concerning the task of building a minimal functionalist model of consciousness, as stated in point (2), there are a variety of theories intended as solutions today, most of which are unfortunately unacceptable as an answer. Consider two popular examples. The first one is Global Workspace Theory^[14], which is based on Dennett’s^[12] philosophical position. Mathematically it is not different from the well-known blackboard architecture, that was around in software engineering for a while. Software tools based on this architecture usually are not claimed to be conscious.

As a second example, consider the popular attempt recently made in the direction of a functional reduction of the concept of consciousness by Giulio Tononi: his theory of integrated information^[15], which sets an approach to formulating generally necessary and sufficient informational requirements for any system possessing consciousness in terms of the internal system’s organization only. Despite the beauty and ambition of this proposal, its particular form has many

problems. This approach ignores key factors that traditionally are used in behavioral identification and characterization of consciousness, namely, interaction of the system with the environment in which it is embedded, and the process of system's development. Also, Tononi's measure of consciousness Φ is not additive, but is defined by the "most conscious" component for any given system, which raises ethical questions. Finally, it is not clear how the proposed measure Φ can be used to practically guide the development of artificial conscious systems.

Further in this work, an alternative approach to constructing a minimal functional model of consciousness based on biologically inspired cognitive architectures (BICA) will be considered, as well as the possibility of its LLM-based implementation for practically important tasks. To address the point (3) of the above plan, it would be necessary not only to build a convincing demonstration, but also to have a confirmation of its acceptance by the community of practitioners, as the scientific method requires^[16].

The conclusion for this part is that solving the question of the nature of human consciousness, understood as an immediate first-person experience, requires expanding the very concept of the scientific method, which today is not urgent and can be postponed. Nevertheless, the task of reproducing the phenomenon of human consciousness in a computer (or another non-biological substrate) is important and can be posed within the framework of functionalism. The required criterion should be sought using the language of cognitive modeling and informatics. Solutions proposed so far^[14-15] fall short of solving the problem. An approach to construct an acceptable solution is presented below.

3 Constructing a functionalist model of social-emotional consciousness based on BICA

The task of implementing a model of human-like consciousness in a computer has been posed numerous

times. In one of the variants, it was called the BICA Challenge^[17]. The challenge is formulated as follows:

Create a computational equivalent of the human mind, including the entire complex of higher cognitive functions manifested in real life, based on a biologically inspired approach.

The mentioned cognitive functions include:

(a) perception, comprehension, reasoning and decision-making, action control, volitional act, agency, goals and desires, goal-directed behavior, planning, basic forms of learning and memory, the ability to communicate in natural language, the theory-of-mind ability;

(b) goal setting and management, a humanlike concept of "self" and personality, dreams and imagination, complex higher emotions, motives, a humanlike value system, social-emotional competency, ability to develop personal relationships, humanlike autobiographical episodic memories, metacognition, self-reflection and self-regulation, humanlike active learning, creativity and cognitive growth.

This truncated description of required functionality can serve as a first approximation for the definition of functional correlates to be taken as the criterion for the presence of human-level consciousness in a system. It should be noted that this is not a set of requirements for intelligence, and not a complete set of requirements for consciousness. This is only one half of the criterion, which should be complemented with certain requirements for the internal functional organization of the system, unfolding the words "biologically inspired", as explained below.

The right biologically inspired approach that allows for enabling the above functionality in an artifact is BICA, with many of the functions listed above mapped to certain structural or functional components of the architecture. The first group of functions (a) require a minimal cognitive architecture^[18], including basic memory types: sensory, motor, working, procedural, and semantic. Functions (a) can be considered already replicated in AI systems and in cognitive models. The second group (b) still represents the challenge^[17] and

requires additional functional components in the architecture: humanlike episodic memory (including autobiographical, both retrospective and prospective), emotional and metacognitive extensions to working memory^[19-21].

BICA as a theoretical paradigm is known under several names. Arguably, almost all known cognitive architectures are biologically or psychologically inspired, therefore, in this sense BICA is a synonym of a cognitive architecture, which is in turn a synonym of the notion of a cognitive agent^[22-24].

Definition 1 A *cognitive architecture (cognitive agent)* is a cognitive system embedded in a certain environment with which it interacts, in which it has sensors and actuators, certain needs and a means of their satisfaction.

The agent performs the complete *cognitive cycle*, including perception, understanding, decision making, and action control. A *cognitive system* is a dynamical system endowed with semantics related to the environment with which it interacts, such that elements and rules of semantics map to those of the system dynamics. By tradition, functional components of the system are called *memory systems*. For a cognitive system to be a cognitive architecture, it must implement a complete agent with its cognitive cycle closed through the environment.

Analogies of the cognitive cycle (or cognitive loop) can be found in theoretical models used in educational sciences (self-regulation^[25]), in military sciences (OODA Loop^[26]), in developmental psychology, and elsewhere. Another notion similar to BICA is the functional system theory developed by Anokhin^[27]. Internal semantics and a closed cognitive cycle are the common features that distinguish all BICA from other systems.

The overarching goal of BICA as a theoretical paradigm is to mathematically describe the workings of human thought, including consciousness. Indeed, BICA is the only widely known computational framework, within which the notion of a voluntary action is defined.

Another important concept in the theory of BICA is the notion of a minimal architecture of its memory systems (minimal BICA) known as the Common Model of Cognition^[18]. At the same time, extended BICA models^[28] allow for describing the entire set of cognitive functions, including (a) and (b). This last statement can be taken as the *working hypothesis* about BICA. In other words, the hypothesis is that the BICA Challenge is solvable.

The main problem with this approach is that the development and implementation of new BICA for each new application paradigm need to be done manually. Furthermore, the multimodal interaction of the system with the user, including the ability to communicate fluently in natural language, can be more efficiently implemented based on other tools. For instance, building a natural language processing system based on BICA is possible, but would be highly inefficient in terms of the cost-benefit ratio, compared to using LLM.

Below we shall focus on the key question of how, based on BICA, a human-level social-emotional conscious AI can be implemented, in particular, enabling agents that can build long-term differentiated relationships with a group of users, achieve mutual understanding, trust and empathy with each of them based on the appropriately individualized social behavior, and ultimately be efficient in collaboration with human partners across many paradigms.

Today, the task of modeling emotions based on BICA receives significant attention^[19,29]. There are various approaches to creating emotionally intelligent social agents^[29-30]. As a candidate approach, we consider the cognitive architecture eBICA^[20-21] (the acronym stands for “emotional BICA”). Its framework is based on two key concepts: semantic spaces and moral schemas. Both are briefly described below.

3.1 Defining the formalism and key terminology

There are many formalisms used today for agent modeling, developed based on varieties of modal logics. Their examples are: situation calculus, fluent calculus, event calculus, a family of action languages, and more^[31-32]. Many models of cognitive systems were built

based on these calculi. For example, modal-logic-based BDI (belief-desire-intention) cognitive architectures were used to implement the OCC^[33] (Ortony, Clore, Collins) theory of emotions as computational models^[32,34-36]. However, the complexity of formal logic requirements of these approaches make them not very suitable for practical applications. On the other hand, cognitive architectures built as rule-based systems, e.g., Soar^[37], although can implement elements of emotional reasoning^[19,37], are not suitable to handle the complexity of human-level emotion logic^[33]. How, then, should we practically build human-level social-emotional intelligence in a cognitive architecture? Several approaches exist^[29-30], and one more is described below.

It is not the intent here to build a formal-logic approach, or a simple-minded rule-based system, although certain elements and ideas will be borrowed from both. The proposed alternative here is based on one general building block: a moral schema, that is defined in the next section. The whole cognitive system can then operate via moral schema action, with possibly multiple moral schemas acting in parallel. To explain this concept, it is necessary to introduce its elements first, including such notions as an actor, perspective, fluent, appraisal, attitude, and intension, that are used here as terms defined for the present purpose within the scope of this work.

The subject of consideration here is an agent interacting socially with other agents in some world. More precisely, we are interested in a model of the social environment (world), represented in the agent's memory system. Following the tradition^[31], we shall use the term *fluent* to refer to any truth-valued variable defined by a predicate, describing any fact about the current state of the world and its dynamics: events, properties, relations, etc. Main elements of the world are agents represented as *actors* and their actions with respect to each other represented as discrete events. Consideration is limited to discrete time. This is a reasonable simplification for building internal representations in the agent.

The notion of an actor captures only relevant characteristics of an agent, who can be a virtual character or a person. Internally, an *actor* is characterized by its identity and internal state of awareness. Externally, an actor is characterized by a set of fluents that determine its position, status, observable conditions.

Actors and actions are characterized by their *appraisals*^[38-39] derived from observable facts based on the agent's value system. In our case, an appraisal of an actor is the social *attitude* of the actor towards another actor, derived from the history of actors' behavior. There can be twice as many different actor attitudes as there are pairs of instances of actors; however, in many cases one attitude would be sufficient to characterize an actor.

An appraisal of an action is its *intension*. In mathematics and in humanitarian sciences, *intension* (not to be confused with *intention*) is a synonym of connotation, or meaning^[40]. In our case, intension of an action is defined by the effect that the action produces on the appraised attitude of the actor. These notions are summarized by the following equation defining the updating rule for the appraisals of actors upon commitment of an action^[21]:

$$A(\text{actor}) := (1 - w) \cdot A(\text{actor}) + w \cdot A(\text{action}, \text{role}). \quad (1)$$

This equation defines the linear appraisal model. Here $A(\text{actor})$ is the appraisal of actor's attitude before or after the action, and $A(\text{action}, \text{role})$ is the appraisal of the action's intension, evaluated for the specified role of the actor (author or recipient). Therefore, an action involving two actors, the author and the recipient, is characterized by two intensions. Finally, w is yet another characteristic of the action: its weight, which can be taken as a small positive scalar. Attitudes and intensions are not scalars but elements of an affine space (could be vectors). It is assumed that appraisals of agents' attitudes do not change when no actions are performed, and appraisals of action intensions are constant.

It follows from Eq. (1) that action's *intension* can be characterized by the target value of the actor's appraisal (attitude). Here "target value" refers to the limiting value of actor's appraisal, when Eq. (1) is

applied repeatedly. In contrast, the action's *intention* is characterized by the target values of fluents that the action intends to change.

Thus, an action has two distinct characteristics. One of them, *intention*, tells what the action is aimed at directly: e.g., close the door. Another one, *intension*, tells what message the action sends to the actor. One can slam the door, expressing anger, or close it gently, showing respect. Closing the door from inside could mean a call for privacy, and so on.

Mathematically speaking, appraisals, including intensions and attitudes, are objects of one kind, taking values on a semantic map, as explained below. A *semantic map* is an overloaded term. Here it is used in connection with the notion of a semantic space^[4,41-43], sometimes also called a "conceptual space"^[44].

Definition 2 A *semantic map* S is an embedding of a set C of semantic categories in an abstract geometrical space X , called in this case a *semantic space*:

$$S: (C, \mu) \rightarrow (X, \nu), \quad (2)$$

such that certain semantic characteristics μ of C are mapped to certain geometric and/or topological characteristics ν of X .

Stated differently, (2) is a morphism, relating a certain set of semantic features μ of C to the corresponding set of geometric features ν of X . For example, if μ is a set of analogical relations like " p to q is like u to v " defined for elements of C , then ν can be the set of corresponding equations " $p-q=u-v$ " for images in X (cf. Word2Vec). A simpler example is a geometrical representation of rating scales.

Here the notion of a semantic category is understood broadly and may refer to arbitrarily general categories (e.g., "good" and "bad"), as well as to arbitrarily specific notions (objects, instances, events, features, etc.) with the restriction that C does not include its subsets as its elements. Terms "semantic space", "space" will also be used to refer to S .

Semantic spaces can be continuous or discrete. In particular, they can be indexed by a mixture of continuous and discrete coordinates. Two mutually

nonexclusive classes of semantic spaces are of interest here^[42-43]:

(1) The semantic space based on a metric space, in which the distance between points is a measure of semantic dissimilarity of the associated categories, is called a *strong semantic space*, corresponding to a *strong semantic map*.

(2) The semantic space based on an affine space, in which spatial directions have globally consistent semantics, is called a *weak semantic space*, corresponding to a *weak semantic map*.

Here we will be dealing mostly with weak semantic spaces. Examples of semantic maps (and correspondingly the mapped elements) include a geographical map (geolocations), affective space (affects), color space (colors), value system (beliefs), word embedding (meaningful words), movie ratings (movies), and so on. The kind of semantic maps that are of interest here are semantic maps of appraisals, including attitudes and related to them intensions. Thinking of these values as emotions or affects would be overly simplistic. Take, for instance, examples of intensions such as "express moral support" or "call for frankness", which cannot be uniquely related to emotions or affects. The spectrum of possible fine-grained intensions is very rich and may include many millions, because in general an intension can be defined not by a dictionary word, but by a paragraph of text. In most practical cases, however, only a few intensional dimensions can be relevant.

Another important notion is that of a possible world studied in Kripke semantics^[45]. Because we are interested in real-life agents' states of subjective awareness which may not conform to formal logic standards, the related notion of a *perspective* will be used here as the main term instead of a possible world. Perspective in psychology refers to a point of view of an individual and therefore is associated with a particular instance of an individual (actor). Here the notion of a *Perspective* is defined as follows:

$$\text{Perspective} = \langle \text{Viewpoint}, \text{Space}, \text{Content} \rangle, \quad (3)$$

where Viewpoint is the instance of an actor whose perspective it is, Space is the semantic space (or semantic map) used for reasoning in this perspective, and Content consists of those agent's beliefs of which the agent is aware in the context of the perspective, including fluents and appraisals.

Perspectives can be mapped to possible worlds (understood in the sense of relational semantics), but not necessarily vice versa. The notion of a perspective differs from Kripke's possible world in several aspects: there is a specified Viewpoint that depends on the agent, the Content is limited, certain formal logic requirements may be dropped, and reasoning is limited to values included in Space. Furthermore, a perspective can be viewed as a dynamical system. One representation of a perspective can be shared by two or more actors, if their individual Contents coincide. In a shared perspective, several Viewpoints are identified. In the next section, the notions of a feeling, a normal state (or norm), and a moral schema will be described, requiring a further unfolding of the notion of a perspective.

3.2 Introducing moral schemas

A *feeling* about some actor is the agent's belief about the "normal" attitude of that actor. It is the agent's notion of what the attitude of/towards that actor ought be. In a moral schema, every attitude has two values: appraisal and feeling. A *moral schema* describes personal relationships among agents by specifying their feelings towards each other, plus possibly certain conditions expressed in terms of fluents. A moral schema affects agent's behavior by adding a certain bias to the choice of actions. This bias works towards achieving or maintaining the *normal state*, in which feelings are equal or close to appraisals of attitudes and fluents satisfy conditions. A significant discrepancy between appraisals and feelings is perceived by the moral schema as a cognitive *dissonance*. A moral schema is an agent on its own, whose only motive is to eliminate and avoid a dissonance. This agent has no identity (no "self"): it assumes the identity be determined by the perspective in which it is instantiated,

merging with the self of the actor. Active states of moral schemas produce *emotional qualia* called here *moods*. A certain mood is associated with every configuration of a moral schema.

Definition 3 A *moral schema* is an anonymous agent defined by the following structure (4)–(8):

Moral schema = <Interface, Fabula, Agency>, (4)

Interface = <Actors, Conditions, Bias, Mood>, (5)

Fabula = <Perspectives, Connections, Storyline>, (6)

Agency = <Host, Objective, Means>, (7)

Perspective = <Viewpoint, Space, Appraisals, Feelings, Norm>. (8)

Here Interface is used for binding the schema to a content in working memory by connecting it to Actors and checking Conditions. Fabula is a narrative network^[46] constructed as a graph where nodes are Perspectives, and Connections (edges) are their temporal, logical, and functional relations. Fabula may comprise inconsistent outcomes and may not predict a unique scenario. Storyline is the working scenario (a consistent subgraph of Fabula) that, together with Norms of individual Perspectives, defines the overall schema's norm labeled Objective: the mission that the schema is trying to accomplish. The notion of Perspective is now expanded to explicitly include Appraisals, Feelings, and Norm – defined as the domain in Space containing Feelings where Appraisals are considered "normal", plus certain additional conditions on the fluents.

Agency of the schema binds to the Host perspective on behalf of which the schema is acting, trying to accomplish Objective using available Means (access to agent's abilities) that depend on Host. Perspectives included in Fabula represent viewpoints of the involved instances of Actors. If Fabula includes only one Perspective (possibly shared), the moral schema is called *simple*, otherwise it is called *complex*. The output of moral schema includes Mood and Bias, where the latter is a set of additional likelihoods of action selection generated by the schema. It is computed so as to minimize the dissonance.

When multiple moral schema instances are

simultaneously active in working memory sharing the same Host, they act independently in parallel and collectively determine the choice of behavior. Furthermore, they can establish connections to each other, forming a structure similar to a narrative network.

4 Integration with LLM and making it conscious

It is fortunate that today we can both recognize and express automatically *intentions* as well as *intensions* in utterances in a dialogue. This capability is provided by LLM, the most popular of which, the GPT line, have been developed by OpenAI. Latest models of ChatGPT, such as o1-preview, demonstrate a human-level understanding of complex social situations and related higher-order concepts, such as irony or empathy. They can generate adequate to this understanding responses at the human adult level, possibly even more adequate than an average human adult would be able to produce. In some cases, they can demonstrate extraordinary, possibly superhuman-level (better than average human) socially emotional intelligence. This observation naturally raises a question about consciousness in such LLM. At present there are no reasons to attribute consciousness to them^[2], primarily because their internal functional organization does not conform to the criterion described above. To endow them with the necessary functional organization for consciousness, BICA can be used.

This is how we can create an artificial conscious being. BICA may be powerless and not capable of exhibiting consciousness without the right human-compatible interface like LLM. On the other hand, without the right functional organization provided by BICA, LLM are just statistical models capable of prompt completion; the illusion that they possess consciousness will break sooner or later, no matter how good they may seem. One can think of LLM and BICA as a key and a lock. Only together they work to open the door to artificial consciousness.

The next section is a proposal for building a

prototype of such system. It considers how moral schemas can be used to guide the process of tutoring, using essay writing as an example. When a cognitive architecture of this kind is implemented, its functionality (and therefore internal functional organization) can be transferred to a deep neural network using methods of deep learning. This transfer will make a further evolution of the system possible.

5 Proposal for practical implementation of a conscious artifact: virtual tutor

The role of the proposed system is a virtual tutor teaching a student to write essays in Computer Science. The paradigm is represented schematically in Fig. 1. The process includes 4 moral schemas presented as rows, working in parallel. At the beginning, the student gets acquainted with the tutor. Then student and tutor perform their task. Finally, student's work is analyzed and evaluated by the tutor.

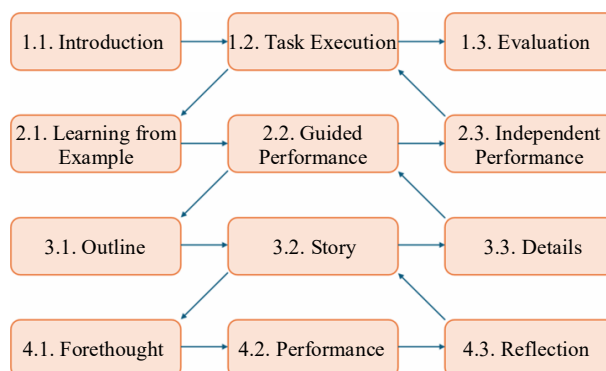


Fig. 1 General scheme of interaction between the student and the tutor during essay writing

An entire experimental session may be allotted up to 20–40 minutes. Let's consider the rows of the table (Fig. 1) in more detail, thinking of each row as a complex moral schema.

Row 1. Student-tutor relationship. Here, macro-level process management occurs, controlling the basic aspects of the student-tutor relationship: respect, subordination, leadership. Every perspective (1.1–1.3) includes two actors: student and tutor. Space dimensions:

willingness to cooperate, leadership, progress. If a dissonance emerges during execution, an abnormal exit may be necessary following an alternative path on the graph (in other rows too; not shown in Fig. 1).

1.1. Introduction. Establishing contact: acquaintance, preliminary discussion, task formulation, implicit confirmation of mutual respect, tutor's leadership, and the desire to work together according to the plan proposed by the tutor in the roles specified by him. Can be performed immediately by checking the conditions with one-two phrases.

1.2. Task execution. Joint work according to the plan, maintaining the achieved relations during work. The goal is to learn how to write essays on any given topic from among current issues in Computer Science and AI.

1.3. Evaluation. Final analysis of the student's work by the tutor. Control questions and recommendations to the student. Assessment of results based on several criteria. Expression of positive emotions by the tutor towards the student as a reward and thanks for successfully completed work—or dissatisfaction with the work.

Row 2. Plan of collaboration according to the task. Each stage of this row should take about 5–10 minutes in the experiment. Time is limited. Space dimensions: communication, understanding, projected success in time.

2.1. Learning from example. Example is demonstrated by the tutor in writing, with verbal explanations. Tutor's answers to student's questions are given, if any. Tutor asks the student control questions to check understanding of the example.

2.2. Guided performance. Student performs the task using feedback and guiding instructions provided by the tutor.

2.3. Independent performance. Student performs the task independently, followed by its analysis and evaluation by the tutor.

Row 3. Stages of essay writing. The three stages of this row include: Outline (plan of essay, bullets), Story (narrative, draft), and Details (refinement,

polishing). Row 3 can be considered as a detailing of the nodes of Row 2, or vice versa. Switching between these modes is possible dynamically (e.g., if there is no progress). Each stage should take about 1.5–3 minutes in the experiment.

3.1. Outline. Jotting down main ideas, essay structure, section headings.

3.2. Story. Writing a draft: forming a coherent storyline that connects main ideas.

3.3. Details. Working on details: section headings should be removed at this stage.

Space dimensions are determined by the written content and include coverage of expected functional components of the essay: definition of the subject with intuitive explanations, discussion of the background and related topics, discussion of questions or problems and existing solutions, characterization of the state of the art, description of methodologies or technical details, discussion of a broad impact on technology/society, discussion of future trends and perspectives, conclusions or a bottom-line / take-home message.

Row 4. Self-Regulated Learning (SRL). This is a standard SRL cycle^[25,47] that should be applied at each node of the first three rows. There is also a connection between Rows 1 and 4. Space dimensions may include: task understanding, motivation, task difficulty, self-confidence, self-image.

4.1. Forethought. Understanding the task, action planning, predicting outcomes, understanding their values, building self-confidence.

4.2. Performance. Executing the plan with metalevel self-monitoring. Checking outcomes. Controlling attention.

4.3. Reflection. Retrospective analysis. Evaluating the outcome. Making causal attribution of the outcomes. Retrospective self-judgment, updating self-image.

Transitions between perspectives is triggered upon achieving the norm in the current perspective – or by the expiration of time. With all four schemas working in parallel, the resultant scenario is not predetermined. This is illustrated in Fig. 2, which shows two possible



scenarios (A, B) emerging from the operation of the system. Active perspectives at each step in Fig. 2 are shaded. Text labels in small boxes that may be hard to read in Fig. 2 are identical to those in Fig. 1.

The next step in creating the prototype is its conversion to a neuromorphic format. By running the system with an emulation of the human user, large volumes of behavioral examples of the tutor can be collected. These data can be further used to train a neural network. In this way, the functionality of the implemented algorithm can be transferred to a neural network. The possibility of this sort of conversion has been demonstrated^[48].

6 Final step: deep learning evolution

The analogy between human brain and computer inspired the initial formulation of research goals for AI^[49] and for cognitive modeling^[22]. However, there are many fundamental differences, if not contradictions, between computer and the brain.

One of these is that the brain develops naturally, including its phylogeny and ontogeny, while computers and software are created by human labor. The unprecedented informational and technological revolution related to LLM is largely due to partially overcoming this difference: there is a small set of relatively simple DNN (Deep Neural Network) models applied everywhere, in very different tasks and paradigms. Each time their application is characterized by the following features:

- (1) Ignoring most of the accumulated knowledge about the subject area.
- (2) Turning any task into a standard big-data-processing task, where the origin and meaning of the data do not matter.
- (3) Solving the task through standard statistical learning techniques using procedures that do not require human intervention.

In contrast, BICA uses accumulated scientific knowledge about the psychology of cognition, learning,

and decision-making, about the subject area, knowledge and experience in building internal representations of the data — all this requires manual labor of developers to implement the solutions. In this sense, integrating BICA and LLM seems unnatural and hardly productive.

However, these limitations of BICA are by no means necessary. Indeed, any BICA model can be implemented in today's popular deep neural networks through training on artificially generated BICA behavior data. Such "reformatting" of BICA models opens the way both to their further adaptation and evolution (e.g., based on deep learning or genetic algorithms) and to a more natural integration with LLM. The anticipated deep learning evolution of neuromorphic BICA can proceed by itself, without involvement of any significant human labor. Its outcome can exceed expectations.

7 Concluding remarks

This work may have much greater significance than it appears, and here is why. Today, we have the capability to create social conversational agents operating within limited paradigms at the level of an adult human expert. Such agents can understand human emotions and attitudes and, in response, exhibit appropriate intensions while performing their tasks, and can be claimed conscious. There are a number of practically important domains for which such abilities are necessary and will revolutionize them. For example: tutors, psychotherapists, partners or assistants in various types of intellectual social activities — marketing, entertaining, creative art, conflict resolution — the examples are numerous. We are not talking here about tasks for which existing dialogue agents like ChatGPT, Siri, or more primitive ones taken as they are, are sufficient.

Developing agents that possess human-level consciousness in limited real-life paradigms is possible by combining LLM with BICA. In this approach, the role of LLM is to provide a translation between natural language and the language of internal representations of

BICA. Initially, the cognitive model must be developed and connected to LLM “manually”. This connection is implemented through prompt engineering. Then, the prototype is empirically validated, and its parameters are adjusted through interactions with human participants. The prototype is then transferred to a neural network via training on the behavioral data generated by it, thereby opening possibilities for its further adaptation and evolution.

Scalability is important: given a set of paradigms for which such agents have been created, the set can be expanded by developing similar solutions for new tasks based on the unified “technological pipeline” to which every success in creating such prototypes contributes. Expansion will continue until the new technology covers all societal needs that can be satisfied by solutions of this type.

Of course, the proposed approach is not the only one (see, for example, works of the group of Jonathan Gratch^[29-30] and also works of the group of Zhen Liu^[50]). Today there is a harsh competition in shaping future AI. It is in our best interests to have it humanlike-conscious. Hopefully, the proposed approach will help it happen.

Acknowledgments

The author is grateful to participating NRNU MEPhI college students. This work was supported by the Russian Science Foundation Grant #22-11-00213, <https://rscf.ru/en/project/22-11-00213/>.

References:

- [1] WU R, YU Z. Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis[J]. *British Journal of Educational Technology*, 2024, 55(1):10-33.
- [2] CHALMERS D J Could a large language model be conscious?[EB/OL]. [2024-08-12]. <https://arxiv.org/abs/2303.07103>.
- [3] CHALMERS D J. *The conscious mind: in search of a fundamental theory*[M]. New York: Oxford University Press, 1996.
- [4] HUTH A G, DE HEER W A, GRIFFITHS T L, et al. Natural speech reveals the semantic maps that tile human cerebral cortex[J]. *Nature*, 2016, 532(7600):453-458.
- [5] DORAHY M J, BRAND B L, ŞAR V, et al. Dissociative identity disorder: an empirical overview[J]. *Australian and New Zealand Journal of Psychiatry*, 2014, 48(5):402-417.
- [6] DUBROVSKY D I. The mind-body problem: the informational approach[J]. *Znanie, Ponimanie, Umenie (Knowledge, Understanding, Skills)*, 2013, 2013(4):92-98. (in Russian)
- [7] VAN GULICK R R. Reduction, emergence and other recent options on the mind/body problem: a philosophic overview[J]. *Journal of Consciousness Studies*, 2001, 8:1-34.
- [8] FARTHING W G. *The psychology of consciousness*[M]. Englewood Cliffs, NJ: Prentice Hall, 1992.
- [9] KUHN R L. A landscape of consciousness: Toward a taxonomy of explanations and implications[J]. *Progress in Biophysics and Molecular Biology*, 2024, 190:28-169.
- [10] SAMSONOVICH A V, ASCOLI G A, MOROWITZ H, et al. A scientific perspective on the hard problem of consciousness[J]. *Frontiers in Artificial Intelligence and Applications*, 2008, 171(1):493-505.
- [11] CHURCHLAND P M. Eliminative materialism and the propositional attitudes[J]. *The Journal of Philosophy*, 1981, 78(2):67-90.
- [12] DENNETT D C. *Consciousness explained*[M]. Boston, MA: Little, Brown and Co, 1991.
- [13] GROSSMANN R. *The existence of the world: an introduction to ontology*[M]. London: Routledge, 2020.
- [14] BAARS B J. Global workspace theory of consciousness: toward a cognitive neuroscience of human experience[J]. *Progress in Brain Research*, 2005, 150:45-53.
- [15] TONONI G, KOCH C. Consciousness: here, there and everywhere?[J]. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, 2015, 370(1668):20140167.
- [16] POPPER K. *The logic of scientific discovery*[M]. Abingdon, UK: Taylor & Francis, 2002.
- [17] SAMSONOVICH A V. On a roadmap for the BICA challenge[J]. *Biologically Inspired Cognitive Architectures*, 2012, 1:100-107.
- [18] LAIRD J E, LEBIERE C, ROSENBLUM P S. A standard model of the mind: toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics[J]. *AI Magazine*, 2017, 38(4):13-26.

- [19] ROSENBLOOM P S, LAIRD J E, LEBIERE C, et al. A proposal for extending the common model of cognition to emotion[C]//Proceedings of the 22nd International Conference on Cognitive Modeling, 2024.
- [20] SAMSONOVICH A V. Emotional biologically inspired cognitive architecture[J]. *Biologically Inspired Cognitive Architectures*, 2013, 6:109-125.
- [21] SAMSONOVICH A V. Socially emotional brain-inspired cognitive architecture framework for artificial intelligence [J]. *Cognitive Systems Research*, 2020, 60:57-76.
- [22] NEWELL A. Unified theories of cognition[M]. Cambridge, MA: Harward University Press, 1990.
- [23] KOTSERUBA I, TSOTSOS J K. 40 years of cognitive architectures: core cognitive abilities and practical applications[J]. *Artificial Intelligence Review*, 2020, 53:17-94.
- [24] SAMSONOVICH A V. Toward a unified catalog of implemented cognitive architectures[M]//SAMSONOVICH A V, JÓHANNSDÓTTIR K R, CHELLA A, et al. *Biologically Inspired Cognitive Architectures 2010*. Amsterdam: IOS Press, 2010:195-244.
- [25] ZIMMERMAN B J, KITSANTAS A. The hidden dimension of personal competence: self-regulated learning and practice[M]//ELLIOT A J, DWECK C S. *Handbook of competence and motivation*. New York: Guilford Press, 2008:509-526.
- [26] RICHARDS C. Certain to win: the strategy of John Boyd, applied to business[M]. Atlanta, GA: Xlibris, 2004.
- [27] ANOKHIN P K. Philosophical aspects of the theory of a functional system[J]. *Soviet Studies in Philosophy*, 1971, 10(3):269-276.
- [28] SAMSONOVICH A V. Extending cognitive architectures [M]//CHELLA A, PIRRONE R, SORBELLO R, et al. *Biologically Inspired Cognitive Architectures 2012*. Berlin: Springer, 2013:41-49.
- [29] MARSELLA S, GRATCH J, PETTA P. Computational models of emotion[M]//SCHERER K R, BÄNZIGER T, ROESCH E. *A blueprint for affective computing: a sourcebook and manual*. Oxford: Oxford University Press, 2010, 11(1):21-46.
- [30] MARSELLA S C, GRATCH J. EMA: a process model of appraisal dynamics[J]. *Cognitive Systems Research*, 2009, 10(1):70-90.
- [31] MUELLER E T. Commonsense reasoning: an event calculus based approach[M]. 2nd ed. New York: Morgan Kaufmann, 2015.
- [32] GLUZ J, JAQUES P A. A probabilistic formalization of the appraisal for the OCC event-based emotions[J]. *Journal of Artificial Intelligence Research*, 2017, 58:627-664.
- [33] ORTONY A, CLORE G L, COLLINS A. The cognitive structure of emotions[M]. Cambridge, UK: Cambridge University Press, 1988.
- [34] MEYER J-J Ch. Reasoning about emotional agents[C]//Proceedings of ECAI'04, 2004:129-133.
- [35] STEUNEBRINK B R, DASTANI M, MEYER J-J CH. A formal model of emotion triggers: an approach for BDI agents[J]. *Synthese*, 2012, 185:83-129.
- [36] WANG Y Y, LIU Z, LIU T T, et al. On the logic of agent's emotions[J]. *Cognitive Systems Research*, 2024, 88:101281.
- [37] LAIRD J E. The soar cognitive architecture[M]. Cambridge, MA: The MIT Press, 2012.
- [38] FRIJDA N H. Appraisal and beyond[J]. *Cognition and Emotion*, 1993, 7(3/4):225-231.
- [39] SCHERER K R. Appraisal theory[M]//DALGLEISH T, POWER M. *Handbook of cognition and emotion*. London: Wiley, 1999:637-663.
- [40] Wikipedia. The free encyclopedia[EB/OL]. [2024-11-20] <https://en.wikipedia.org/w/index.php?title=Intension&oldid=1242642153>.
- [41] HUTH A G, NISHIMOTO S, VU A T, et al. A continuous semantic space describes the representation of thousands of object and action categories across the human brain[J]. *Neuron*, 2012, 76(6):1210-1224.
- [42] SAMSONOVICH A V, GOLDIN R F, ASCOLI G A. Toward a semantic general theory of everything[J]. *Complexity*, 2010, 15(4):12-18.
- [43] SAMSONOVICH A V. On semantic map as a key component in socially-emotional BICA[J]. *Biologically Inspired Cognitive Architectures*, 2018, 23:1-6.
- [44] GARDENFORS P. Conceptual spaces[M]. Cambridge, MA: MIT Press, 2004.
- [45] GASQUET O, HERZIG A, SAID B, et al. Kripke's worlds: an introduction to modal logics via tableaux[M]. Basel: Birkhäuser Basel, 2014.
- [46] PENTLAND B T, FELDMAN M S. Narrative networks: patterns of technology and organization[J]. *Organization Science*, 2007, 18(5):781-795.
- [47] BERGLAS-SHAPIRO T, EYLON B S, SCHERZ Z. A technology-enhanced intervention for self-regulated learning in science[J]. *Teachers College Record: the*

- Voice of Scholarship in Education, 2017, 119:1-26.
- [48] TRAFTON J G, HIATT L M, BRUMBACK B, et al. Using cognitive models to train big data models with small data[C]//Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, 2020:1413-1421.
- [49] MCCARTHY J, MINSKY M L, ROCHESTER N, et al. A proposal for the Dartmouth summer research project on artificial intelligence[M]//Chrisley R, Begeer S. Artificial intelligence: critical concepts. London: Routledge, 2000: 44-53.
- [50] LI Q, LIU T T, LIU Z, et al. Emotionally intelligent virtual tour guide in handling group conflicts: Effect on user outcomes[J]. Computer Animation and Virtual Worlds, 2023, 34(3/4):e2153.
- [49] MCCARTHY J, MINSKY M L, ROCHESTER N, et al. A proposal for the Dartmouth summer research project on

走向有意识的人工制品: ChatGPT 是钥匙, 但要打开门, 必须找到锁

SAMSONOVICH Alexei V^{1,2}

(1.俄罗斯国立核研究大学, 俄罗斯 莫斯科 115409; 2.乔治梅森大学, 美国 弗吉尼亚 费尔法克斯 22030-4444)

摘要: 在机器中复制人类意识(可被理解作为一种直接的第一人称主观体验)的想法是这个时代最大的挑战之一。对现象意识的本质的科学理解可能需要扩展科学方法本身。可以暂时搁置解决该问题, 而先尝试在功能主义的框架内在计算机中重现意识现象。迄今为止, 为此所提出的相关理论似还存在问题。提出了一种基于 BICA(生物启发认知架构)结合 LLM(大语言模型)实现类人意识的方法, 并以虚拟导师为案例对其进行说明。一般认为, 上述问题的最终解决方案可以以纯神经形态的形式获得, 且这一形式能使其得以进一步变更、发展和演化。为此, 从意识的本体论问题及其暂时的非还原解决方案开始讨论, 因为这是科学研究的一条可行路线。主要假设是基于“BICA 挑战”中有关 eBICA 认知架构制定的, 作为例子仅分析了有限的社会情感意识案例, 解释了本身不具备意识的 LLM 如何通过 ChatGPT 连接到 eBICA, 用于构建有意识的人工制品原型并验证其声称的性能。所选定的虚拟导师案例进一步说明了这项技术将如何使虚拟导师发展出人类级别的师生关系, 从而提高辅导效率。未来的研究方向可以包括创建各种各样具有感知能力、且有可能进一步自主进化的代理物, 以满足人类的众多实际需求。

关键词: 机器意识; 意识难题; 信念-欲望-意图; 生物启发认知架构; 大语言模型; 虚拟导师

(责任编辑 韩 超)