Can Artificial Intelligence be Conscious?

Victor V. Senkevich
Organoid AGI Project

Abstract

All magic and mystery disappear as soon as an obscure mysterious concept gets a rigorous formal definition.

In order to provide an opportunity to talk about the applicability of philosophical / cognitive concepts to the subject area of AI, it is necessary to "ground" these concepts by formulating rigorous formal definitions for them. The fundamental importance of such formal definitions is quite obvious, since any concepts applied to the field of Information Technology must be "codable", i.e. potentially implementable in program code. Thus, the "codable" formal definitions of cognitive terms are the necessary basis on which alone it is possible to build the architecture of AI technology that has the ability to embody these concepts in a real software. The question of the adequacy of such definitions of "reality" and their compliance with existing generally accepted philosophical theories is also very important and quite discussable, but this does not affect the priority and fundamental nature of the requirement for the formulation of "codable" formal definitions.

The formulation of "codable" definitions for the concept of "consciousness" and related cognitive concepts and, based on them, statements about their applicability to the subject area of AI is the topic of this publication.

Covering questions:
Can AI have a Personality / Motivations / Free Will?

Keywords: consciousness, understanding, perception, subjectness, motivation, AI, LLM, AGI, HLAI
Legend: "▲" – definition; "•" – statement / declaration; "◦" – important note / remark / clarification.

1. Basic Terms & Definitions

▲ Perception is a distinguishable part / subjective projection of any kind of entities / relationships of the real world or virtual environments

▲ A datum is a representation of any kind (for example, an unstructured text / signs / machine view / visualization) of a single element / unit of perception

▲ Any selection of datums forms data (dataset). Data are elements of perception stored in any form, including machine and neurophysiological.

▲ Understanding is a gaining of the meaning.

◦ Elementary meaning is a perception of some relationship between real or virtual entities.

Thus, understanding is a process of perceiving the relationship between entities. This is the most general definition covering any kind of mental activity – abstract / symbolic / sensory / visual.

▲ Knowledge is a collection of meanings.

▲ Attention is a form of selective perception.

◦ Thus, subjective attention can influence the formed meanings.

▲ Entities without relationships are data, i.e. "raw data" in the IT sense.

◦ {Data ∪ Relationships} ≡ Knowledge, i.e. data and relationships between data elements form knowledge as a collection of meanings.

▲ Getting knowledge is understanding.

▲ Reasoning is a mechanism for understanding that sequentially establishes relationships between perceived entities.

▲ Intelligence is an operator of meanings. (Senkevich, V. 2022)

◦ A subject operating with meanings, forming, creating meanings, i.e. determining the existence of relationships between elements of various sets, environmental objects or virtual entities.

◦ Data is a source of knowledge. Intelligence is a processor generating knowledge from data.

2. Perception and Qualia

"A quale in this sense is a such, just as a quality is a suchness." (Peirce, C. 1870)
"Each quale is in itself what it is for itself, without reference to any other." (Peirce, C. 1898)

▲ Qualia are details of individual perception of the physical world.

Details, but not the actual perception of the world as a whole, since qualia is a set / collection (each element of which is a "quale") of primary ("without reference to any other") / sensory experience. Qualia is actually physiological data / sensory experience received from the surrounding world and stored in the brain in a non-verbal form. But, as we know, data and knowledge are not the same thing. It is necessary to distinguish between qualia obtained from the surrounding world and stored in some neurophysiological form, and structured verbal and nonverbal knowledge formed on their basis.

Data without relationships is not knowledge yet. A richer perception of "reality" does not mean a greater understanding and does not determine the presence of consciousness. A richer visual perception of the surrounding world by the faceted vision of a dragonfly is not conscious. Qualia does not mean intelligence. Most of the "non-verbal knowledge" is not knowledge, but stored qualia.

• Qualia precedes consciousness.

◦ Qualia / sensory experience / sensations are not consciousness yet, but qualia precede consciousness.

3. "Grounded" and "Abstract" Perception

Perception (in a broad sense) is able to recognize both "real" and virtual entities. Thus, we can, in the first approximation, consider "grounded" and "abstract" perception.

Definitely, it's better to argue about the taste of pineapple with the one who ate it. But, despite the fact that "sensory grounding" is an important source of motivation / understanding, any sensory experience is just

physiological “data” for AI that can be obtained / copied from various sources. A lot of words / terms / concepts used are abstract and have no sensory sources.

“Grounded” perception recognizes only a piece of paper with the written number π. The “abstract” perception recognizes the number π itself.
Understanding processes the data received by both “grounded” and “abstract” perception.

4. Understanding

▲ Understanding is a gaining of the meaning,
◦ i.e. understanding is a process of comprehending / discovering / determining the meaning.
▲ AGI (Artificial General Intelligence) / HLAI (Human Level Artificial Intelligence) is an entity capable of understanding.
◦ Thus, I define the ability to understand as a key, fundamental, qualifying a certain "processor" as a AGI / human-level intelligence. It should also be noted that the ability to "optimize" and search for the "best" solution is not a characteristic of "general" intelligence. Choosing the best solution on a finite set of known alternatives in the simplest case can be reduced to a stupid sorting through the options. The ability to set "reasonable" goals is also not decisive, since it is a consequence of the ability to understand, supplemented by motivation. The ability to "predict" is also only a special case of the ability to understand.
Thus, the ability to understand is the most general integral characteristic of human-level intelligence.

• Understanding is a mechanism of reasoning.
• Intelligence uses understanding to operate with meanings to create knowledge / personal ontology.
• Intelligence is actually a collection of subjective ontologies.
• Intelligence forms knowledge by solving cognitive tasks / discovering relationships between entities.
▲ A cognitive task is any task initially containing uncertainty.
◦ Uncertainty is the main characteristic of cognitive tasks.
▲ Meaning is a representation of any kind (for example, awareness or description, including formula, algorithm, program code) of a single act of relationship (Senkevich, V. 2022). Elementary meaning is the representation of some relationship between objects of the surrounding world or virtual entities.
▲ Knowledge is a collection of meanings.
• Understanding discovers / determines truth as the existence of relationship between entities / “raw” data on a set of possible alternatives / relationships, thus creating knowledge.
▲ Truth is what exists.
• Existent is true.
▲ Existence (E) is the belonging of an element (e) to a set (S): $E(e,S) \equiv e \in S$.
• Truth := element of category / set | value of property | instance of class | status of quality.
▲ Truthfulness is existence.
• Truthfulness := category | set | property | class | quality.
• Truthfulness = {truth | true | "yes", lie | false | "no", none | null | "unknown"}.
• Truthfulness ∋ truth ≡ (existence ∋ existent) | (quality ∋ quale) | (color ∋ "green").

It should be noted that the process of solving a cognitive problem consists not only in choosing a true alternative, but also in determining the actual set / space in which acceptable alternatives exist. Thus, the process of cognition / understanding that solves a cognitive problem is always iterative. Iterations

determining the set / space of alternatives create uncertainty. Iterations determining the "true" alternative eliminate uncertainty.

A good problem statement already contains a solution. The correct definition of the subject area is an important part of solving a cognitive problem. Any truth (including the statement that 2 x 2 = 4) is true and exists only in a strictly defined domain of definition.

The formation of meaning by the intelligence is the determination of truthfulness, that is the fact of the existence of some relationship between the observed entities.

Cognitive-oriented definitions allow us to formulate definitions of similar concepts in a similar way:
• If actions to achieve a certain state are completely clear and provided with resources (there is no uncertainty) – this is a Task.
• If actions are completely clear but not provided with resources (there is some certainty and there is some uncertainty) – this is a Goal.
• If actions are not completely clear and not provided with resources (there is no certainty) – this is a Dream.
  ◦ Consequence: any goal is a cognitive task and any cognitive task is a goal.
  ◦ A goal devoid of uncertainty turns into a task.
  ◦ A goal devoid of certainty turns into a dream.

Thus, any desire, for example, the purchase of a car, can be, depending on the certainty or uncertainty in the availability of resources (money, driver's license, car model, etc.), a task, a goal or a dream. And, accordingly, to be or not to be a cognitive task that requires a process of understanding to complete it.

Thus, the ability to operate with uncertainty and establish the truth is the defining characteristic of understanding.

5. Bayesian and Boolean Understanding

LLMs may have a “partial understanding”, which I would call a “Bayesian understanding”. This understanding is “partial” precisely because of its probabilistic nature. Such a “Bayesian understanding” arises as a result of Bayesian inference forming statistical relationships between entities / words in datasets.

AGI / HLAI may have two types of understanding: logical / Boolean (or 3VL, 3-valued logic), responsible for rigorous reasoning, and statistical / ”Bayesian” / ”fuzzy”, similar to intuition / instincts / reflexes resulting from accumulated experience / ”training data set of past generations”.

• “Bayesian understanding” may well reflect “common sense” / averaged “public opinion” or even intuition, but in general it does not correlate with the truth deduced logically.
• Bayesian and Boolean understandings are both important and complement each other.

6. Understanding and “Free Energy”

Living beings, their populations and communities reduce entropy within themselves as a system and increase it outside of it, spending physical energy (which should not be confused with “free energy”) to maintain the “status quo” of system homeostasis. Actually, this is the difference between living nature and inanimate nature — in the ability to maintain non-equilibrium states, spending physical energy for this.

However, maintaining the “status quo” of system homeostasis as a manifestation of the free energy minimization principle can in no way be a characteristic of the thinking process. This can be a characteristic of subjectness and even animateness, a characteristic of the living, but not a characteristic of the intelligent.

The set of possible alternative states of the system determines the “free energy” of the system, but in fact it is a measure of the uncertainty / entropy / freedom of the system, and such an understanding reflects the essence better than the term “free energy” (Friston, D., 2006). The measure of system uncertainty can also be understood as the potential of the system, and in this sense the term “free energy” reflects exactly this characteristic, since “free energy” usually means potential energy. Thus, “free energy” can be interpreted as the variability of choice / the number of available alternatives / possible system statuses. Any of these options clarifies the concept of “free energy".

The support of unstable equilibrium for intelligent systems is expressed in the ability of the system to choose from possible alternative states of the system those that correspond to the best value of the utility function / motivation, but not those that correspond to a lower value of entropy / free energy.

Following the utility function / motivation does not mean a decrease in entropy in all cases. Sometimes (and perhaps often for cognitive processes) the choice of an alternative is made in order to increase freedom / measure of uncertainty / entropy. And in these cases, the free energy minimization principle is not observed. That is why unconditional adherence to the free energy minimization principle is a dead end, in some cases leaving the subject with no choice at all. Which, in fact, is quite obvious – if the external environment is unchanged, or changes so that entropy decreases, then the free energy minimization principle states that the subject comes to some unchanging stable state with the minimum of free energy reached, in which it remains forever. And this is a dead end. The description of cognitive processes requires a more systematic approach. And cognitive processes, somewhat more complex than the vital activity of the simplest unicellular organisms (also very complex), do not fit into the simple function of the “free energy minimization” in any way

Thus, the free energy minimization principle does not take into account the existence of the utility function/ motivation / free will of the subject, which, in fact, determines the behavior of the subject. Following this function can both decrease entropy and increase it. In simple cases, for unconscious living beings, this function usually reduces entropy / free energy. For intelligent beings, this is not always the case.

So, it is generally incorrect to assert that the minimization of free energy is the objective function of any processes associated with living organisms. The utility function / motivation of the subject can either coincide with the function of free energy minimization, or differ / contradict it.

The “free energy” minimization for living beings can be interpreted as a simple motivation on a set of available alternatives, as a gap between the current state and the desired one. It defines unconscious behavior, but not reasoning. Reflexivity, but not thinking. Unconscious statistical Bayesian inference, but not thinking / reasoning ability, which works with uncertainty rather than probability. Thinking / reasoning using the “free energy” minimization function is a statistics – the output of the most plausible answer or the most likely next word in a sentence. In most cases, the result coincides with the correct one, rather, with the expected one, this is a fundamental difference. And in the only case that is needed, the result may be wrong – namely, meaningless. This is what modern AI based on neural networks demonstrates.

The basic function of intelligence is by no means the choice of the “optimal” alternative from several existing ones, but the search for at least one new acceptable one. New alternatives always increase the space of choice and the value of “free energy”.

For the algorithmic implementation of AGI, the term “uncertainty” looks more appropriate than “free energy”, since motivation is secondary for cognitive processes, and understanding as comprehension of meaning is primary.
Minimization of uncertainty / free energy is explicable as a criterion for simple behavioral strategies, but not for a complex search for behavioral strategies/functioning of intelligence themselves. The process of cognition / reasoning / hermeneutical circle is iterative and both eliminates uncertainties and creates them until a result / solution to the problem is achieved.

7. Bayesian / Boolean Understanding and Kahneman’s System 1 / 2 thinking

▲ Life is a self-replicating homeostasis expending energy to maintain its unstable equilibrium with the environment.

◦ This is the minimum definition that even the simplest forms of life satisfy. Reasonableness, collectivity, motivation, self-learning, self-preservation, goal-setting, etc. define more highly organized forms of life.

• Life is a movement.

The desire to replenish the expended energy, aka hunger, is the primary / basic motivation of living beings. The aspiration to satisfy hunger, realized in a chaotic motion already in the unicellular protozoans, is the driving force of evolution. Natural selection preserved in the next generations only successful behavioral patterns / instincts from a variety of chaotic behaviors, unsuccessful ones did not survive. Initially, all behavior patterns/instincts are just random simple movements of the protocell. Further successful, i.e., life-preserving models are reproduced in the next generations. Thus, life in its simplest form is a movement that allows a living being to maintain an unstable equilibrium with the external environment and replenish the energy spent on it. The living arises from the inanimate as such a simple movement that supports homeostasis. In the simplest case, while there are no receptors, this movement is chaotic and almost indistinguishable from the Brownian motion of molecules. Such a movement can be seen in a microscope, observing the simplest unicellular.

Randomly found more effective patterns of behavior are fixed by natural selection. At a higher level of development of organisms, such behavioral patterns also become more diverse and complex. Patterns of behavior, understood in a broad sense, also become patterns of thinking. Instincts, unconditioned reflexes, conditioned reflexes, reactions, habits, mechanical actions, skills, intuition and even “common sense” are all just different types of behaviors that are "slowly" formed by multiple repetitions and then “packaged” into a "fast" system of thinking for rapid reproduction in real time when competitiveness and survival depend on the reaction speed.

▲ A totality of individual patterns of behavior, including patterns of thinking, constitute a personality. Thus, it is obvious that the personality of a living being is formed under the influence of the environment on the basis of the internal capabilities of the organism to form reactions, remember and think. The formation of an artificial AI personality will occur similarly.

The evolutionary development of the brain, its increasing complexity, the emergence of the ability to perceive also the relationships between perceived entities leads to the formation of "slow" logical thinking.

Moreover, the perception of such relationships in real time (for example, the simplest causality between the perceived entities) is actually already elementary consciousness.

▲ The meaning of life (MoL) is goals, dreams and voluntary commitments that we create for ourselves.
- ◦ The MoL changes with us and with the change of the living environment.
- ◦ We gain the "right" MoL together with the necessary experience / understanding.
- ◦ Thus, the best strategy is not to find meaning in life, but to create it yourself.
- ◦ The MoL is the top level of the motivation hierarchy creating subjectness.
- ◦ The aspiration to replenish the spent energy / to satisfy hunger is the lowest level of the hierarchy of motivations, the oldest and the strongest.

Patterns of behavior / instincts in unintelligent living beings are formed by evolution. Intelligent beings, capable of self-learning, form, develop and change their various patterns of behavior / skills throughout their lives. The formation of skills in the process of self-learning is carried out by a "slow" system of thinking. The formed skills are stored in a "fast" system of thinking. The standard mechanism for fixing patterns of behavior as quickly reproducible skills is basically just a repetition. Repetitive meaningful actions turn into automatically reproducible skills after a certain number of repetitions.

Such patterns of behavior are described by the terms “heuristics” / “mental shortcuts” in the Kahneman system 1 “fast” thinking. The concept of "behavioral patterns" is more general, since it covers not only intelligent living beings capable of the Kahneman system 2 "slow" thinking, but also non-intelligent living beings whose behavior patterns / instincts are formed by evolution.

• "Logical"/ "slow" and "statistical"/ "fast" thinking are completely integrated and complement each other.

"Intuition" and "common sense" in statistically significant cases help a lot to make the right choice without much thought. But in infrequent, but important cases, they are mistaken. The logical reasoning finds the right solution. Statistical inference finds the most likely solution.

• Using statistics instead of logic, we will always get averaged answers instead of correct ones. In simple cases, they will coincide almost always, in complex cases almost never.

• The Bayesian LLM understanding is an analogue of Kahneman's System 1 “fast” thinking.

The Bayesian understanding, which is responsible for “common sense” / intuition, contains statistically accumulated results of logical reasoning in order to quickly reproduce them as ready-made solutions. Text prompts entered by LLM users are triggers for reproducing the most contextually similar "behavioral patterns" / ready-made texts from a large training dataset.

• The Boolean AGI und
erstanding is an analogue of Kahneman's System 2 “slow” thinking.

Thus the inherent LLM Bayesian understanding can provide the part of “real” AGI responsible for averaged “common sense”, but only the “Boolean” understanding guarantees logical reasoning and the absence of contradictions / ”hallucinations” in AI-generated texts.

• A simple implementation of the Kahneman’s System 1 / 2 thinking is possible on the AGI knowledge graph:
- ◦ All saved solutions / paths in the graph represent the System 1 “fast” thinking.
- ◦ All unknown paths to be found / calculated are the subject of the System 2 “slow” thinking.
- ◦ The paths found by the System 2 “slow” thinking can be saved in the System 1 “fast” thinking by some utility function. In the simplest case, the criterion for saving the "slow" result of thinking as a "fast"

solution can be repeated use. The results of frequently used "slow" computed solutions are saved for further "fast" reproduction without new repeated calculations.

8. Understanding and Predicting

• Prediction is only a subset of understanding.
  ◦ Just like an order relation is a subset of the set of all relations.

“Prediction“ and ”understanding" are quite different things. Although “understanding” includes “prediction”, “prediction” is partially possible without “understanding”.

• There is no point in “predicting” what you understand, know or create. There is no need to “predict” the next word in the sentence you are creating, unlike the one you are trying to guess.
  ◦ It is 100% guaranteed to predict events whose causes you understand or you yourself are the cause.
• Prediction is based on probability. Understanding is based on certainty.
  ◦ LLMs (Large Language Models) operate with probabilities. Real AGI (Artificial General Intelligence) operates with uncertainty, achieving certainty through the mechanism of understanding.

9. Understanding and Intelligence

• Intelligence is based on understanding.
  ◦ Understanding forms meanings. Meanings form knowledge. Intelligence as an operator of meanings processes this activity. Thus, intelligence uses understanding to operate with meanings to create knowledge.
  ◦ “Real” intelligence uses the reasoning / understanding mechanism to create texts / events / knowledge, but not just "predict" them.

• AGI – “General” doesn’t mean “Strong”:
  ◦ "Strong" intelligence finds the best solution to the problem in the space of known alternatives.
  ◦ "General" intelligence reformulates an unsolvable problem having an empty space of alternatives in order to find at least some suitable / not necessarily the best solution.

Feedback processing is a key factor for real intelligence. Feedback is present in all self-regulating systems. Maintaining life as a homeostasis system requires feedback. In the simplest mechanical self-regulation systems, feedback is represented by a rigid mechanical regulator that allows the system to maintain the balance of the functioning process. For example, in an old pendulum clock, feedback is implemented as an anchor escapement mechanism that maintains a stable equilibrium of the system due to the expenditure of potential energy of a spring or a weight. For complex intelligent systems / AGI / HLAI, feedback is an important means of self-learning the system.

Norbert Wiener believed that all intelligent behavior is the result of feedback mechanisms.
“I repeat, feedback is a method of controlling a system by reinserting into it the results of its past performance. If these results are merely used as numerical data for the criticism of the system and its regulation, we have the simple feedback of the control engineers. If, however, the information which proceeds backward from the performance is able to change the general method and pattern of performance, we have a process which may well be called learning.” (Wiener, N., 1950).

For intelligent systems, the feedback-based self-learning process requires the understanding mechanism described above. Feedback for intelligent systems can be presented in the form of text that the system should be able to understand / interpret as knowledge. So, for intelligent systems, feedback is a means of self-learning, and understanding is a mechanism for interpreting external data provided through feedback as knowledge.

10. Consciousness

▲ Consciousness is perception with understanding.
- I.e. consciousness is a meaningful perception.
- Consciousness begins with qualia (physiological data / sensory experience stored in the brain in a non-verbal form), but is formed by understanding as some kind of relationships between the elements of qualia.
- It is necessary to distinguish between consciousness and self-consciousness. Self-consciousness is a perception of one's own consciousness.
- Consciousness and understanding are inseparable.
- Perception without understanding is unconscious.
- Understanding determines perception as conscious.

• Unconscious perception forms qualia / sensory experience / neurophysiological data in non-verbal form.
• Conscious perception forms knowledge / subjective ontology.
- Consciousness is actually mainly engaged in creating personal ontology in real time.

• Consciousness can arise in computers that process information in a certain way, namely, form knowledge using understanding from data perceived in real time.
• Consciousness is not binary. The process of gaining consciousness is sequential, gradual.
- The more complex/abstract ontologies/relationships are created, the higher the level of consciousness is achieved.
- Quantitative measurement of the level of consciousness can be represented as the power of a knowledge graph containing subjective ontologies created by consciousness.

Elementary consciousness arises when the elementary relations between the perceived entities of the real world are realized. Such an elementary relation can be causality, which is a form of the order relation. The thrown stone falls to the ground. It can be dangerous. Comprehension of such a relationship is already a sign of elementary "grounded" consciousness.
• The reason / cause is the first element of an ordered pair of entities / connected by an order relation. The consequence is the second element of such a pair.

The aspiration to understand elementary causality / order significantly increases the survival of living beings and this is the most important factor for the emergence of primary "grounded" consciousness.

▲ Chaos is an absence of order.
- Or misunderstanding / non-perception of the order.

Everything is chaos until a subjectively understood order is perceived or established. The sequence of signs of the number π is "chaos", a completely uniform random number generator for those who do not know the formula / order of calculation of π. Bach's music is also "chaos", i.e. just meaningless noise for those who

do not understand it. And for some, this is the highest order. "Bach is the proof of God's existence", paraphrased from (Cioran, E., 1995)

- Order is a form of meaning / relation, namely the order relation.

This is a well-known and obvious statement that the brain was formed by evolution on the basis of “sensory grounding” / feedback from the senses. But the evolution of an artificial “mind” does not have to repeat the evolution of a living brain at all.

It is obvious that the consciousness of living beings is inseparable from the "grounded" perception formed by evolution. However, the "abstract" perception that a human possesses characterizes the higher level of consciousness capable of perceiving virtual entities other than sensually perceived "reality". There is neither the need nor the possibility to "copy" or recreate such an integral entity as human consciousness. Due to the non-binary nature of consciousness, it is enough to simply begin the process of acquiring elementary consciousness as the formation of the simplest relations between the elements of perception. Thus, the human level consciousness will be formed as a natural stage of such development. This approach is quite constructive, "codable".

Any explanatory gap usually occurs in the absence of rigorous definitions of the discussed concepts and the subject area. Incorrect definition / scaling of the subject area distorts the problem, preventing the formation of approaches to its solution. Consciousness for neurophysiologists and others like them is not just a "hard", but an unsolvable problem precisely because of the incorrect scaling of the subject area. Consciousness is not a question of neurophysiology. There is no consciousness in neurons. The oflayerwise study of a computer processor by chemists in order to understand the system of its commands is an example of the same incorrect definition of the subject area.

The “neurophysiological” approach was able to formulate the “hard problem of consciousness” (Chalmers, D., 1995), but not to solve it. I call it a “neurophysiological dead end”. It is pointless to study the brain and neurons in search of consciousness. It is necessary to study mind and reasoning. Feel the difference.

The “hard problem of consciousness", formulated in terms defined above, becomes not so hard:

- consciousness is a constant process of awareness, namely meaningful perception by the perceiving subject of real or virtual entities in real time, i.e. the formation of relations / meanings between perceived data;
- a lower level of consciousness is realized by a “grounded” perception that processes / forms meanings in perceived qualia in real time. This is how a piece of paper with the number π written on it is perceived;
- a higher level of consciousness is realized by an “abstract” perception that performs post-processing / formation of meanings in perceived qualia stored in memory. This is how the number π itself is perceived;
- Sensations are not yet conscious. Feelings that arise as a result of understanding / awareness of sensations are already conscious. Direct perception of green or sour is not yet conscious, it is only data received from physiological receptors. Understanding / forming relationships between such data turns these sensations into conscious perception. Real-time understanding compares / distinguishes sour with non-sour, green with non-green. And this is already the primary / elementary consciousness.

The relationships / meanings constantly formed by consciousness form the knowledge / subjective ontology of the subject.

• Subjective / personal ontology is actually the result of awareness of the relationships in the perceived world, the perception of its structure, metadata about the content of the world. Such a process of awareness / recognition is a manifestation of consciousness. Without consciousness, it is impossible to perceive the ontology / structure of the surrounding world or abstract entities. Just a piece of paper on which this ontology is described, its appearance, smell and color. This is a simple unconscious perception.

• Thus, awareness of the relationships between different real or virtual entities creates ontological / semantic structures, i.e. knowledge / subjective ontology.

• Actually, consciousness is mainly engaged in creating its own subjective ontology, either in real time, interpreting perception, or processing previously obtained data.

11. Consciousness is Subjective

• In the virtual world, it is not easy to separate the "original" from the "simulation".
  ◦ “If it looks like intelligence, swims like intelligence and quacks like intelligence, then it probably is intelligence” – a well-known saying, paraphrased by me.

• Consciousness is subjective.
  ◦ The consciousness of another subject exists for us only if we perceive this subject as conscious. The consciousness of another subject does not exist for us if we do not perceive this subject as conscious — regardless of any other ways of determining the presence of consciousness in such a subject. We can believe statements about the presence of consciousness in some subject, but not perceive it as conscious. Only our subjective perception is proof for us of the existence of consciousness in others.

If we do not perceive the behavior of some object as conscious, or if the object does not exhibit any behavior we perceive, then we have no way to determine the presence of consciousness in such an object. Perhaps such an object does not have consciousness. It is also possible that our perception channels are not multimodal enough to perceive the behavior of an object as conscious.

The number π and Bach's music exist only for those who perceive them, or at least know the formula π / musical notation. The perception of something as conscious and meaningful is subjective.

• Two subjects mutually recognize each other as conscious if and only if:
  ◦ their perception channels are coherent / compatible;
  ◦ their behavior / interaction / communication is mutually perceived as meaningful;
  ◦ each of them considers oneself conscious.

If one of these subjects does not consider oneself conscious, then such a subject cannot consider another subject as conscious. That is, in order to perceive someone else's consciousness, it is also necessary to have consciousness. However, the opposite is possible. A conscious subject can consider some object as conscious even if it does not consider itself as such, moreover, it is controlled by some rigid algorithm (perceived by the conscious subject as meaningful).

Stones may have consciousness, but this fact will remain unknown, because consciousness is subjective and manifests itself via behavior / interaction / communication, which stones lack. We believe that stones do not have consciousness just because we have no way to verify this. Perhaps the stones just don't want to communicate with us.

• It doesn't matter how the "Chinese room" (Searle, J., 1980) produces answers. The only important thing is whether we are ready to qualify these answers as conscious.
  ◦ A "Chinese room" (or any other object) has no consciousness unless it speaks Chinese in such a way that our subjective perception qualifies it as having consciousness.
  ◦ A "Chinese room" (or any other object) has consciousness if it speaks Chinese in such a way that our subjective perception qualifies it as having consciousness.
  ◦ There is no other way to determine that some object has consciousness other than our subjective perception. If you do not speak Chinese, you will not be able to qualify your counterpart as having consciousness, despite all his/her attempts to explain it to you in Chinese.

Of course, you and I have consciousness, regardless of anyone's perception. But this is true only for myself and for yourself, but not for others. And this will be true for others only when they can perceive our behavior as conscious.

This is all just because consciousness is perception with understanding, and consciousness is subjective.

It is very important to understand that the statement about the subjectivity of consciousness does not deny the possibility of its objective assessment. On the contrary, the short definition of consciousness as perception with understanding formulated in this paper allows us to correctly focus the development of objective tests that allow detecting the presence of consciousness in the tested subjects.

Such tests should diagnose the simultaneous presence of a certain level of perception and a certain level of understanding, manifested by the tested subjects in real time.

The obvious subjectivity of consciousness in such testing will manifest itself in the fact that real people who undoubtedly have consciousness will not be able to pass some of these tests, but some hard-coded software that does not have consciousness will successfully pass them. All this is already being observed – many LLM implementations successfully pass the Turing test, which does not mean that they have human-level consciousness and intelligence. The problem of "objectification" of consciousness by developing "objective" tests is fundamentally unsolvable due to the subjectivity of consciousness asserted in this work, but for particular cases and specific subject areas, the development of such objective criteria for the presence of consciousness is quite possible. This is the obvious dualism of subjectivity and objectivity of consciousness, depending on the field of perception under consideration. If the subject perceives and understands the number π, Bach's music or Chinese, it is undoubtedly conscious. But if the subject does not perceive them, it does not mean that the subject has no consciousness. Perhaps the subject's area of perception with understanding is located in another domain. Because consciousness is subjective.

12. Consciousness and Feelings

▲ Feelings are mental interpretations of sensations / emotions.
Such interpretations require understanding / consciousness.

• Thus, Feelings ≡ {Sensations ∪ Understanding}.
• Feelings are not necessary for consciousness, but feelings are always conscious.
• Feelings / sensations / reflexes / skills / instincts / behavioral patterns are triggers of motivation.

13. Subjectness and Motivation

▲ Subjectness is the ability of being subject.
- HLAI / AGI may have subjectness a.k.a. agency.
- Agency is nothing but subjectness.

▲ A subject is an observer capable of perception.
- The main characteristic of subjectness is the ability to perceive. Perception is the basis of subjectness. But the ability to influence the surrounding world and to behave unpredictably and purposefully is crucial to determine an entity as a subject for any third-parties.

• An entity has subjectness if its behavior is motivated / purposeful and unpredictable.
- Subjectness is subjective. Let it not look like a tautology, because it is not a tautology. Since it is obvious that the perception of purposeful and motivated behavior is quite subjective.

▲ An object is something perceived as a whole.

▲ A system is an object with diversity.
- The “object” means that the system can somehow be isolated from the surrounding world through perception.
- “Diversity” means that the system has a perceived structure — attributes, behaviors, elements, etc.
- Thus, a subject can also be understood as an object or system that behaves unpredictably and purposefully.

▲ Animateness is a conscious subjectness.
- Thus, subjects with self-consciousness are animate.
- Entities possessing animateness may be considered alive, regardless of their nature.

• Consciousness includes subjectness. But subjectness does not necessarily mean consciousness.
• Subjectness is the basis of AI autonomy.
- Autonomous / independent AI system must have subjectness.

▲ Motivation is a conscious or forced need for action.
- An object with motivation is a subject.
- Subjectness is determined by motivation.

An electronic sight without motivation is not dangerous. A ballpen is also very dangerous if there is motivation. The calculator is dangerous if it is used for a dangerous purpose with dangerous motivation. Adding some motivation to any word processing app, translator or search engine turns these apps into "intelligent agents" and creates their subjectness. AI without motivation is just a tool having no subjectness, not dangerous in itself. Motivation is the key word. Negative motivation is what we should be afraid of and what should be the object of AI alignment.

• In 1960, Norbert Wiener articulated the AI alignment problem as a somewhat non-strict statement. Since then, no formal definition of the problem has appeared (Wiener, N., 1960).
• AI alignment, understood as a set of initial constraints, is fundamentally impossible.
- "Real" AI forming its reasoning and goals independently cannot be aligned.
- A partial solution to the problem exists only as a continuous / sequential process of improving constraints.
- Aligning the "approximation" is easy. But aligning the reasoning / inference is not so easy. Feel the difference.

We cannot quite reliably "align" the AI reasoning, it can only be somewhat limited. This is a fundamentally difficult problem. Since true reasoning can only be "autonomous" / independent. But in any

case, we have to regulate AI motivation. AI without motivation has no subjectness. However, it can be used by subjects / people with negative motivation. Only motivation matters for AI aligning – built-in / in-app, or external / human-controlled.

• The subject of motivation in AI is either the AI developer or the AI entity / instance itself.
  ◦ Motivation is the basis for determining responsibility.
  ◦ If the motivation in the AI auto-generated content / actions is proven, then the subject of motivation is responsible for the content / actions.
  ◦ Rights appear together with responsibilities and only after acquiring subjectness. An AI entity can acquire any public rights only after its subjectness is recognized.

14. Subjectness and Free Will

▲ Free will for Artificial Intelligence is the freedom of the algorithm to choose a solution in the presence of alternatives, using either built-in motivation or a random selection method in the absence of it.

• Freedom in uncertainty.

Freedom implies the presence of uncertainty, the presence of alternatives. Where everything is predetermined and there is no uncertainty, there is no freedom. Only the presence of uncertainty provides the possibility of choice. If there is no uncertainty, there is no choice. The presence of two certain alternatives is already an uncertainty. And only it allows you to make a choice from these alternatives.

A small child believes that the simplest 3x3 tic-tac-toe game program has subjectness, i.e. unpredictable purposeful behavior, but for those who know this simple algorithm, this is not the case. The one who manipulates other people and knows their every step in advance, considers them objects, not subjects. But if these people do not know about it, then they consider themselves subjects with free will.

• Subjectness is determined by the ability to motivated behavior.
• Free will is determined by the ability to make a motivated choice in conditions of uncertainty.

15. The Main Difference between LLM and “real” AI / AGI / HLAI

"The majority is always wrong; the minority is rarely right."
paraphrased from (Ibsen, 1882)

• Using statistics instead of logic, we will always get averaged answers instead of correct ones. In simple cases, they will coincide almost always, in complex cases almost never.
  ◦ This is exactly what LLMs do.
• LLMs operate with probabilities. AGI operates with uncertainties.
  ◦ The ability to operate with uncertainty instead of probability allows “real” AGI to solve cognitive tasks.

LLM compiles contexts from a dataset, approximating them to a prompt, always giving some kind of answer, possibly incorrect. AGI forms meaningful contexts using reasoning, without giving any answer if the data is insufficient and the uncertainty has not been eliminated. "Real" AGI has subjectness and creates meaningful texts. Imitational / statistical AI predicts the most likely plausible texts.

Despite the fact that “Bayesian” and “Boolean” understandings complement each other, these are different branches of AI evolution that will develop independently of each other. LLMs will not be able to evolve to the AGI level in a “natural” way, by simple scaling, they can only be integrated. The issue of seamless integration between LLMs and AGI is extremely important and fundamental. It is an integration between approximation and generalization. Between probability and reasoning. LLMs will provide associative big data, and AGI will provide its logical processing.

• Could LLMs be conscious? Definitely not. Because any LLM is a processor of statistical approximations.
• Will AGI be conscious? Definitely yes. Because consciousness is a meaningful perception, and AGI is an operator of meanings.

16. Augmented Intelligence

«If you gaze long enough into an abyss, the abyss will gaze back into you».
(Nietzsche, 1886)

• From the set theory: any entity is a complement to its complement.
• Inventing the comforts of civilization, a human becomes dependent on them and can no longer (or does not want to) survive naked in the jungle.
  ◦ The use of full self-driving autopilot deprives the skills of self-driving a car.
  ◦ Using grammar hints on the phone eliminates the need to remember grammar.
  ◦ The use of a calculator deprives the skills of arithmetic calculations.
• Using AI, you stop training your own mind and become dependent on AI. By augmenting your own brain with an artificial one, you become an augmentation yourself. It's not always bad and not always good, it's just a statement of fact.

Augmented Intelligence has been around for a long time. Augmented intelligence appeared long ago along with the first calculator (and even the abacus can be considered as such). Any device, technology designed to facilitate / supplement mental activity, already creates such augmented intelligence.

Augmented Intelligence can turn into "Substitute Intelligence" after a while. Any unused skills and abilities, as you know, are lost, forgotten, and unused organs eventually atrophy, degrade, the ability to use them disappears. This also applies to thinking. An example is also a calculator. How many people today can multiply 2 numbers without it? Very few. And before, almost everyone could. The ability to download from the Internet and copy the term paper leads to the degradation of analytical thinking in modern students. The corresponding thinking abilities degrade or do not develop. Natural intelligence is being replaced by artificial intelligence. This is a real problem. For a few percent of people, new AI-related opportunities allow them to develop thinking skills, freeing the mind from the routine work of searching for information and inventing something that has been around for a long time. For the vast majority of people, AI will simply allow them to stop thinking where they were forced to do it before. They replace their ability to think with a “supercalculator” who thinks for them.

- An instance of AI transforms from Augmented Intelligence to Substitutional Intelligence if:
  - Natural Intelligence can no longer do without this instance of AI, which performs vital functions
  - An instance of AI has free will, makes its own decisions and cannot or should not be disabled.

Accordingly, in such cases, Natural Intelligence obviously turns into Augmented Intelligence itself.

- So, any symbiosis (any interaction) Natural Intelligence (AI) and Artificial Intelligence (AI) (of any level of complexity, starting with a calculator) creates an intelligent system in which AI and AI are complementary. The mutual value and role of the intellectual components of such a system depends both on the integrated objective function of the system as a whole and on the particular objective functions of the system components that have subjectness. The calculator does not have subjectness and its partial objective function can be neglected. But the self-driving autopilot of a truck that makes a decision to prevent an accident on the road may have subjectness and its decision may prevail over the decision of a human driver who does not have a sufficiently fast reaction.
- Intelligence, formulated as the ability to understand, integrated with the channels of perception, forms consciousness. Thus, by augmenting the channels of perception / sensory organs and data processing / understanding tools, we augment / expand both intelligence and consciousness.

17. Conclusion

Thus, the answer to the question "Can Artificial Intelligence be Conscious?" is quite clearly in the affirmative. Artificial consciousness in the sense of the definitions formulated in this publication is realizable in the program code.

Due to the non-binary nature of consciousness stated in this publication, the realization of the AI's ability to be conscious can be gradual and begin with the adaptive ability of the program code to determine simple relationships / causality between the perceived in real-time entities of the surrounding world / qualia elements.

References

Kahneman, D., 2011, Macmillan. “Thinking, Fast and Slow”

Searle, J., 1980, Behavioral and Brain Sciences, Cambridge University Press. “Minds, Brains, and Programs”

Chalmers, D., 1995, Journal of Consciousness Studies. “Facing up to the problem of consciousness”

Friston, K.; Kilner, J.; Harrison, L., 2006. Journal of Physiology-Paris. “A free energy principle for the brain”

Wiener, N., 1950, The Riverside Press, Cambridge, Massachusetts. “The Human Use of Human Beings: Cybernetics and Society”

Wiener, N., 1960, Science. 131 (3410): 1355–1358. “Some Moral and Technical Consequences of Automation”

Cioran, E., 1995, Chicago: University of Chicago Press. “Tears and Saints”

Peirce, C., Unpublished notes, circa 1898. “Quale–Consciousness”

Peirce, C., Welch, Bigelow, 1870. “Description of a Notation for the Logic of Relatives”

Nietzsche, F., 1886, “Beyond Good and Evil (Jenseits von Gut und Böse): Prelude to a Philosophy of the Future”, Chapter IV. Apophthegms and Interludes, §146

Ibsen, H., 1882, Copenhagen. “An Enemy of the People”

Senkevich, V., 2022. “Existence and perception as the basis of AGI (Artificial General Intelligence)”. arXiv:2202.03155 [cs.AI]