Consciousness in Machines: Faulty Declarations and Basic Principles

Imants Vilks
Artificial Intelligence Foundation Latvia researcher
Ikskile, Ozolu str. 40, LV-5052
e-mail: imants.vilks@gmail.com

Abstract

The article is a short review about one particular but important and essential question in the field of AI: the consciousness. The main idea is to cast away non-scientific pronouncements about mystics and impossibility in the field and to define basic notions (emergence of new properties) and the main theme - consciousness. The article proposes temporary definition of consciousness, consciousness emergence conditions and how to recognize and test the consciousness.

Keywords: *Emergent properties, consciousness, emergence conditions.*

1. Emergence

Emergence is the origin of new properties or processes in a physical system consisting of some parts and connections between them. The phenomenon of emergence in a scientific community is an issue broadly discussed and there is not an unanimous attitude toward it. Although some say that "*it is a perennial philosophical puzzle*" and "*it is uncomfortably like magic*" [1], the current paper shows that there is nothing mystical about the emergence.

All molecules and atoms of physical matter, all biological cells, millions of human inventions and patents, all mathematical and physics formulae are emergent property devices. For the simple system we can predict and calculate the new properties of the system, for example, for the ordinary dice. New properties-the probability of coming up a chosen number - are predictable and calculable using the knowledge about mass distribution inside the cube, the knowledge about outer surfaces and edges.

A bit more complicated systems like the loops for catching animals, different traps and holes are still predictable. Their properties are predictable and explainable, and the emerged properties are reducible to constituent parts.

For more complicated systems, e.g., for the atoms and molecules our theories, our models are incomplete. In many cases the new features are hardly predictable. Why? We don't have enough information about the constituent parts, about their connections and properties. Simply, we don't have good enough theories, system models. Some information is missing. This is why we measure the properties of new substances experimentally.

For example, even today in chemistry we encounter new elements with unknown properties. Philip Ball writes:

"Chemistry is a messy business. The elements are so diverse that their interactions can be unpredictable and sometimes bizarre. Often, chemists rely on nothing more than intuition to tell them what may or may not be possible. Sometimes that leads them astray. History is littered with ideas that were derided or dismissed at first, but eventually changed the rules of the game" [2].

Philip Ball tells us five stories of chemistry they said could never happen. Add the missing information and you have an invention or a new substance.

For some authors the notion of emergence is incomprehensible:

"I believe that this notion of emergence is incomprehensible—rather like a naive conception of the big bang. The idea that everything (matter, space-time, their antecedent causes, and the very laws that govern their emergence) simply sprang into being out of nothing seems worse than a paradox" [3].

Received July 11, 2012; Revised August 25, 2012; Accepted September 17, 2012

2. The solution is based on main mathematics and physics principles:

- always when we have sufficiently good, complete theory or the model of the process, we can predict and calculate the emerged properties. That means, the complexity is reducible to elements, e.g., we can calculate the orbit of a satellite, the mass center of some physical system or the masses of products in a chemical reaction;
- due to the lack of complete physical theory or due to the complexity of mathematical calculations we can't always predict the properties of newly created devices, e.g., the properties of new organic compounds;
- all properties of newly created objects are explicable, reducible to component properties-in accordance with our knowledge about constituent parts.
- incomplete knowledge about constituent parts and their collaboration creates an illusion of mystical experience [3].

3. Example

In an electronic generator sinusoidal oscillations emerge when necessary physical conditions and connections are realized. The oscillation **emergence conditions** are:

- power amplification, i.e., power of oscillations at the output is about known amount bigger than the power at the input;
- frequency selective element, e.g., appropriate resistor-capacitor (RC) or inductance-capacitor (LC) circuit resonating at some frequency;
- positive feedback, i.e., some determined part of the output signal is correctly (in a right phase, i.e., in such a way that output signal, added to the input increases the output signal amplitude) fed back to the input;
- non-linear element (or circuit, i.e., the connection of some elements) providing constant output amplitude.

How do the oscillations emerge? All electrical circuit elements have thermal (and other specific-vacuum tube or transistor) noise at their terminals. At the moment when all emergence conditions are fulfilled noise at the input of the system is amplified, some part from it is fed back to input, summed together with the input noise, and amplified again. Exponential growth of the noise level sets in.

At the same time frequency selective circuit passes the noise signals corresponding to its resonance frequency and attenuates the other signals. At that time sinusoidal oscillations gradually grow out (are filtered out) from the noise and increase exponentially. We can say - emerge. Amplitude, frequency, rise time and all other parameters (output power, noise level, non-linearity) are calculable and measurable-they are determined by circuit's elements.

Is this complexity irreducible? Often, if we don't have sufficiently complete model of the system, it seems so. In this case we have a sufficiently complete mathematical model and we can reduce emergence of oscillations to constituent parts.

If we take a Furrier transform from the amplifier's output signal we find that the output noise consists of infinite number of sinusoids with different amplitudes and phases. After the amplifier is switched on (everything properly connected and power supply is on) the sinusoids near the resonance frequency are amplified but the sinusoids with other frequencies are attenuated. In this way the pure sinusoidal signal at the output emerges.

All properties of the output signal are reducible to elements properties, but they are not equal to elements properties. Like the sum 5 = 2+3. It is reducible to elements 2 and 3 but is not equal.

The elements determine the output signal parameters, the parameters can be calculated. Even the very fact of emergence of oscillations is **reducible to elements properties**. In radio engineering the set of these properties is called '**generation conditions'** (the 4 conditions above) and they are exactly describable.

4. Consciousness

Human knowledge about consciousness has developed synchronous with our knowledge in biology and computer science. The British psychologist Stuart Sutherland (1927-1998) wrote:

"*Consciousness is a fascinating but elusive phenomenon: it is impossible to specify what it is, what it does, or why it has evolved. Nothing worth reading has been written on it*." [4].

In recent years in thinking about consciousness we can observe two big and partially overlapping directions:

- knowledge gained by classifying known facts and features, defining and combining the new ones, and discussing their features;
- knowledge based on accomplished work. In the first direction vagueness, ambiguity and some mystics is easy to find.

In Wikipedia consciousness is defined: "***Consciousness** is a term that refers to the relationship between the <u>mind</u> and the world with which it interacts*".

Some other definitions are simply tautological: Consciousness is "*the state or fact of being <u>conscious</u> of an external object, state, or fact"* [5].

Obscure pronouncements about consciousness are common:

"Arranging atoms in a certain way appears to bring consciousness into being. And this fact is among the deepest mysteries given to us to contemplate…

Perhaps the emergence of consciousness is simply incomprehensible in human terms…

But couldn't a mature neuroscience nevertheless offer a proper explanation of human consciousness in terms of its underlying brain processes? We have reasons to believe that reductions of this sort are neither possible nor conceptually coherent. Nothing about a brain, studied at any scale (spatial or temporal), even suggests that it might harbor consciousness." [3].

For some the consciousness is a miracle and mystery:

"At some point in the development of certain complex organisms, however, consciousness emerges. This miracle does not depend on a change of materials — for you and I are built of the same atoms as a fern or a ham sandwich. Rather, it must be a matter of organization. Arranging atoms in a certain way appears to bring consciousness into being. And this fact is among the deepest mysteries given to us to contemplate." [3].

Jeff Hawkins characterizes the situation:

"…*most people think consciousness is some kind of magical sauce that is added on top of the physical brain. You've got a brain, made of cells, and you pour consciousness, this magical sauce, on it, and that's the human condition. In this view, consciousness is a mysterious entity separate from brains. That's why zombies have brains but they don't have consciousness. They have all the mechanical stuff, neurons and synapses, but they don't have the special sauce*." [6], p.195.

This situation will last until AI researchers will manage to get some system conscious (at the beginning-in some special fields) and we will recognize the consciousness in it.

What is consciousness?

Jeff Hawkins explains:

"*We can be certain that however consciousness is defined, memory and prediction play crucial roles in creating it... Your cortex creates a model of the world in its hierarchical memory. Thoughts are what occur when this model runs on its own; memory recall leads to predictions, which act like sensory inputs, which lead to new memory recall, and so on.* ***Our most contemplative thoughts are not driven by or even connected to the real world; they are purely a creation of our model****. We close our eyes and seek quiet so that our thinking will not be interrupted by sensory input. Of course* ***our model was originally created by exposure to the real world through our senses, but when we plan and think about the world****, we do so via the cortical model, not the world itself.* [6], p.199. (bold –I.V.).

To the cortex, our bodies are just part of the external world. Remember, the brain is in a quiet and dark box. It knows about the world only via the patterns on the sensory nerve fibers. From the brain's perspective as a pattern device, it doesn't know about your body any differently than it knows about the rest of the world. There isn't a special distinction between where the body ends and the rest of the world begins. But the cortex has no ability to model the brain itself because there are no senses in the brain. Thus we can see why our thoughts appear independent of our bodies, why it feels like we have an independent mind or soul. The cortex builds a model of your body but it can't build a model of the brain itself. Your thoughts, which are *located in the brain, are physically separate from the body and the rest of the world. Mind is independent of body, but not of brain.* [6], p.200.

Your understanding of the world and your responses to it are based on predictions coming from your internal model. At any moment in time, you can directly sense only a tiny part of your world. That tiny part dictates what memories will be invoked, but it isn't sufficient on its own to build the whole of your current perception". [6], p.202.

5. Definition

Using existing knowledge we can create a temporary definition of consciousness. **Consciousness is a stream of previously accumulated experience in which the physical entity called oneself cooperates with environment in different situations of life. This is achieved by arousing groups of neurons in which previous experience is stored.**

Previous experience is a sum of many signal streams: external audio, visual, touch, thermal, scent and inner body (emotional and somato-sensory) signals. They all are remembered **in sum**, i.e., they create rich and colorful pictures about past events.

In order to create the memories about past situations, those situations must be remembered. Remembering is formed in higher cortical regions by filtering out the most essential features of current events.

Just this is what is being stopped by anesthesia: back propagation, i.e., creating the material for storing in memory. Experiments described by Alkire et al. 2008 in "Consciousness and Anesthesia" suggest that:

"Consciousness vanishes when anesthetics produce functional disconnection in this posterior complex, interrupting cortical communication and causing a loss of integration; or when they lead to bistable, stereotypic responses, causing a loss of information capacity. Thus, anesthetics seem to cause unconsciousness when they block the brain's ability to integrate information." [7].

Linda Geddes writes:

"EEG studies on anesthetized animals suggest it is the backwards signal between these areas that is lost when they are knocked out."

"*Both propofol and the inhaled anesthetic sevoflurane inhibited the transmission of feedback signals from the frontal cortex in anesthetized surgical patients. The backwards signals recovered at the same time as consciousness returned* ."

"In 'global workspace' theory, incoming sensory information is first processed locally in separate brain regions without us being aware of it. We only become conscious of the experience if these signals are broadcast to a network of neurons spread through the brain, which then start firing in synchrony.

The idea has recently gained support from recordings of the brain's electrical activity using electroencephalograph (EEG) sensors on the scalp, as people are given anesthesia. This has shown that as consciousness fades there is a loss of synchrony between different areas of the cortex - the outermost layer of the brain important in attention, awareness, thought and memory"

At the deepest levels of anesthesia *…"Long-distance communication seems to be blocked, so the brain cannot build the global workspace. It's like the message is reaching the mailbox, but no one is picking it up*." [8].

6. Emergence conditions

The necessary physical conditions for emergence of consciousness are:

1. Memories from previous experience, in which a physical entity we call 'individual' plays the main role: receives the signals from external world and reacts on environmental stimuli. Most of the experience must be gained by learning, not to be pre-programmed. Why? Only by learning the acquired experience will be connected to machine's sensor and actuator neurons.
2. Neural program which builds the models of external world and predicts future events and actions of itself [6], pp. 87-97.

3. Neural program which compares different previous actions and chooses the most appropriate action for the situation at a moment. In order to choose the best action the target function or the action prize-list for different external situations must be introduced or created by individual experience.
4. External stimuli initiating the stream of memories.

All neural reinforcement learning programs create external world ‘maps’, accumulate external world reactions, predict future events and choose the best action for the current situation. In simplest cases they miss two essential things: they don’t create the notion of acting subject - as a part of external world (see quote [6] p. 200 and 202 above) and, second, they create very simple models (e.g., action-reaction pairs) of external world. These two things can be more or less elaborated, outspoken and, accordingly, the consciousness is more or less outspoken and noticeable for the outside observer.

Humans have developed special symbols and languages (physics, mathematics, ordinary languages) and created special world models (theories) for describing and predicting the external world processes. This is why we say that humans are the most intelligent beings in our world.

Hawkins writes:

“When a ball is thrown, three things happen. Thirst, the appropriate memory is automatically recalled by the sight of the ball. Second, the memory actually recalls a temporal sequence of muscle commands. And third, the retrieved memory is adjusted as it is recalled to accommodate the particulars of the moment, such as the ball’s actual path and the position of your body. The memory of how to catch the ball was not programmed into your brain; it was learned over the years of repetitive practice, and it is stored, not calculated, in your neurons”. [6], p.69.

Aleksander and Dunmall have described similar 5 conditions:

“Let A be an agent in a sensorily-accessible world S. For A to be conscious of S it is necessary that:

***Axiom 1** (Depiction): A has perceptual states that depict parts of S.*

***Axiom 2** (Imagination): A has internal imaginational states that recall parts of S or fabricate S-like sensations.*

***Axiom 3** (Attention): A is capable of selecting which parts of S to depict or what to imagine.*

***Axiom 4** (Planning): A has means of control over imaginational state sequences to plan actions.*

***Axiom 5** (Emotion): A has additional affective states that evaluate planned actions and determine the ensuing action.”* [9].

Consciousness is a gradual quality, which can be more or less outspoken and, therefore, more or less recognizable. From information theory axiom “any information processing system’s output is restricted by the information previously gathered and by the information generated inside” follows: **machine or robot will show more or less consciousness in accordance with its sensory experience**.

For example, chess-playing programs and other cognitive systems with self-awareness mechanisms, where “a *system is continuously generating meanings from continuously updated self-models*” [10] are only partly self-conscious. If not humanlike, we will not recognize the consciousness in the machine.

7. How we will recognize the consciousness?

By Turing test.

Jeff Hawkins challenges the Turing test suitability for consciousness tests and offers the machine’s ability to make predictions instead: *“We can now see where Alan Turing went wrong. Prediction, not behavior, is the proof of intelligence.”* [6], p.105.

This argument is right-finding out the most essential feature of consciousness is obligatory step. But in any case in order to test the machine, some questions and behavioral test will be necessary.

Temporarily we have to stay at the Turing test looking for answers indicating that the machine has built-in models ([6], pp. 6 and 225) of external world and itself, makes predictions

([6], pp. 2 and 88) and is able to choose appropriate actions. This means, 'understands' the external world [6], p.89.

For example, for internet to achieve human-like (which we will recognize) consciousness is necessary:

- human-like (at least-partially) input sensors and actuators;
- local body-like physical system isolated from external space;
- programs and memories listed in emergence conditions linked with machine's sensors and actuators.

The only limit for the machine to pass Turing test will be the machine's '**experience**' (touch, temperature, visual, audio, smell, inner body signals, social contacts, and emotions) **library**. Of course, the machine like every human will not give reasonable answers to the questions about the senses and experiences it does not have.

8. Conclusion

In this paper much discussed notion of emergence is shown to be a natural and widely known (but not always recognized) process of our physical world. The very fact of emergence of new properties in physical systems is understandable and reducible to elements properties, in a broader view-reducible to the known laws of physics and mathematics.

The phenomenon of consciousness is shown to be nothing special but just a separate case of emergence. The temporary definition of consciousness is given-in accordance with the existing knowledge. When researchers will create (and recognize) consciousness in some neural network we will be empowered to say that we have complete theory of consciousness. Now, for the current moment we have temporary knowledge and definition.

Author proposes to investigate physical processes like consciousness in neural networks to be grounded on known laws of physics and mathematics. The article shows that the consciousness is a gradual quality. The conditions of emergence and the criteria for recognizing and testing the consciousness in any physical system are given.

References

[1] Bedau, M. Weak Emergence. *Philosophical Perspectives: Mind, Causation, and World*. 1997; 11: 375-399.
[2] Ball, P. Impossible reactions: Five chemistry rules broken. *New Scientist*. 2012; (2848): 31.
[3] Harris, S. The Mystery of Consciousness*. Sam Harris Blog*, 2011.
[4] http://www.thebigview.com/mind/.
[5] Merriam-webster dictionary. http://www.merriam-webster.com/dictionary/consciousness
[6] Hawkins, J. On Intelligence. New York: Times Books. 2004.
[7] Alkire et al. Consciousness and Anesthesia. *Science. 2008; 322* (*5903*): 876-880*.*
[8] Linda Geddes. Banishing consciousness: the mystery of anesthesia. *New Scientist.* 2011; *29*.11.
[9] Aleksander, I., B. Dunmall. Axioms and Tests for the Presence of Minimal Consciousness in Agents. *Journal of Consciousness Studies*. 2003; 10 (4–5): 7–18.
[10] Carlos Hernandez, Ricardo Sanz, Ignacio L´opez. Consciosusness in Cognitive Architectures. 2008-02-27, http://core.kmi.open.ac.uk/display/87106