Journal of Artificial Intelligence and Consciousness
Vol. 7, No. 1 (2020) 73–82

DOI: 10.1142/S2705078520500046

On Artificial Intelligence and Consciousness

Pentti O. A. Haikonen

Department of Philosophy
University of Illinois at Springfield
One University Plaza
Springfield, IL 62703, USA
pentti.haikonen@pp.inet.fi

Published 29 April 2020

The popular expectation is that Artificial Intelligence (AI) will soon surpass the capacities of the human mind and Strong Artificial General Intelligence (AGI) will replace the contemporary Weak AI. However, there are certain fundamental issues that have to be addressed before this can happen. There can be no intelligence without understanding, and there can be no understanding without getting meanings. Contemporary computers manipulate symbols without meanings, which are not incorporated in the computations. This leads to the Symbol Grounding Problem; how could meanings be incorporated? The use of self-explanatory sensory information has been proposed as a possible solution. However, self-explanatory information can only be used in neural network machines that are different from existing digital computers and traditional multilayer neural networks. In humans, self-explanatory information has the form of qualia. To have reportable qualia is to be phenomenally conscious. This leads to the hypothesis about an unavoidable connection between the solution of the Symbol Grounding Problem and consciousness. If, in general, self-explanatory information equals to qualia, then machines that utilize self-explanatory information would be conscious.

Keywords: Artificial Intelligence; Machine Consciousness; Symbol Grounding Problem; Qualia.

1. State of the Art

According to recent publicity, Artificial Intelligence (AI) would seem to be doing well. Advances have been reported in various areas such as game playing, natural language translation, information search, robots, smart phones, smart assistants, and self-driving cars, to name a few. This publicity and recent movies about AI and conscious robots have created the popular impression that AI is an entity that will soon, if not already, be able to think, understand and be intelligent in superior ways and even be conscious. Unfortunately, we are not there yet, and contrary to the popular perception, the progress towards superior general AI has been very slow.

It is obvious that major advances in AI cannot be achieved without the consideration of the very idea of AI. What is AI, what can be achieved with it, what are its limitations? Could or should AI be conscious? These questions must be considered from both theoretical and practical viewpoints. Remarkable theoretical and philosophical groundwork has been done in recent years by well-known researchers like Baars, Boltuc, Block, Chalmers, Chrisley, Dennett, Harnad, Reggia, Searle, Sloman, Tononi and others.

On the practical side, the impactive works of Aleksander, Franklin, Hesslow, Kinouchi, Manzotti, Shanahan, to name a few, are well known. Biologically inspired robots have been studied by Chella, Brooks, Goertzel, Haikonen, Holland, Kawamura, Sanz, Takeno and others.

A building is only as strong as its foundations. The same goes for AI. Therefore, it is useful to inspect the early foundations of AI.

2. Early Foundations of AI

The theoretical foundations of AI go back to the 1950s and were explicitly presented by Newell and Simon [1976]. In the early days of computers and AI, Newell and Simon proposed that intelligence is based on symbol manipulation, and a physical symbol system (computing machinery) has all the necessary and sufficient means for general intelligence. Thus, according to Newell and Simon, a physical symbol system is required whenever intelligence is to be produced. Consequently, also the human brain has to be a physical symbol system. On the other hand, a digital computer is known to be a physical symbol system; it manipulates binary word symbols by given rules, the program. It is understood that a computer can compute everything that can be expressed as proper algorithms in the form of computer programs. (This statement is true, because it is a tautology. Any proper algorithm is computable by definition.)

Newell and Simon concluded that both the computer and the brain are physical symbol systems and are therefore computationally equivalent; a suitably programmed computer can execute every algorithm that is executed by a brain. These conclusions are also known as the Physical Symbol System Hypothesis (PSSH). However, Newell and Simon were not able to prove this hypothesis directly. Instead of this, they presented indirect evidence: There are no other hypotheses that could explain how intelligence could be produced in the brain or a machine. Fodor seconded this by stating that physical symbol manipulation is "the only game in town" [Fodor, 1975].

However, the PSSH leaves open some fundamental questions. Obviously, the brain can operate as a physical symbol system, but could it also be something else, negating the computer−brain equivalence? Also, if that were the case, would a physical symbol system really have all the necessary and sufficient means for the production of general intelligence? So far, no artificial general intelligence (AGI) has

been produced by computer programs. Would that be due to lack of trying or would the PSSH be invalid?

It is a fact that the existing AI programs are effective only in narrow area applications; they do not exhibit any general intelligence. Therefore, they are called Weak Artificial Intelligence (Weak AI), as opposed to the currently hypothetical Strong AI (also AGI), which would be comparable to human mind and intelligence.

3. Weak Artificial Intelligence

Computer programs that can generate apparently intelligent results in well-defined narrow areas are called Weak AI. With the exception of few experiments, the existing AI applications represent Weak AI.

Remarkable results have been recently reported with Weak AI. Considering this, it can be asked, what would be the point of Strong AI, if, already in the near future, stacked Weak AI programs were able to produce all that is practically required. Unfortunately, this will not happen. Weak AI will remain weak for a simple reason; it does not understand anything. This is a direct consequence of its fundamental way of operation.

Weak AI is based on algorithmic symbol processing. An algorithm is a sequence of rules to be followed exactly for the production of the desired outcome. Computer programs are algorithms. Algorithms have a benefit: The executor of an algorithm does not have to understand what the rules are ultimately about. Consequently, algorithms can be executed by persons, agents and machines that do not understand the meanings of the used symbols nor the actions taken. For example, a calculator operates with symbols (numbers), but does not know what is being computed. The external meanings of the numbers are not entered into the calculator.

This benefit of algorithm is also its greatest weakness. An algorithm is able to do only what it is designed to do; therefore, its application area is inherently narrow. A chess-playing algorithm is not able to do anything else, no matter how trivial. Weak AI is not readily able to do any outside-the-box reasoning. A computer does only what the program commands it to do, not necessarily what would be needed. This, of course, could be remedied by providing a rule or a program for each and every possible situation.

If one rule is not enough, then how many rules would be needed? It turns out that a very large number of rules, perhaps even conflicting ones, would be required for the production of AGI by brute force. Real human intelligence does not work that way; it is not rule following. Instead, it is something that is used when rules don't work. It is about finding and learning a new response that fits the situation. But this would call for the understanding the situation; meanings would have to be manipulated instead of blind manipulation of symbols. This requirement leads to the Symbol Grounding Problem: How to attach meanings to the used symbols.

4. The Symbol Grounding Problem

There is no understanding without meanings. In order to understand something, a computer would have to operate with meanings, but this does not happen, because physical symbol systems manipulate symbols, not their meanings.

However, how come then that certain already existing AI-based assistants, like Alexa and others, are able to converse in a natural language? After all, they seem to understand what has been asked, and are able to answer more or less properly. The inconvenient truth is these systems do not really understand anything. The words refer to words, not to their meanings, and tricks are used to create the illusion of meaningful conversation.

In a digital computer, symbols can be made to refer to other symbols, but not to their meanings because the meanings are not readily attached to the symbols. The meanings of the used symbols cannot be defined ultimately by other symbols within the system; they would have to be imported from the outside. But the imported meanings cannot be imported as additional symbols, as also these would call for interpretation. In pure symbol processing systems this would appear to be an unsolvable problem as illuminated by the Chinese Room thought experiment of Searle [1980]. How to give meanings to the symbols? This is the well-known "Symbol Grounding Problem".

The Symbol Grounding Problem is also present in the process of thinking. In humans, thinking manifests itself as the flow of inner imagery and inner speech. Inner speech utilizes a vocabulary of words and also syntax, the apparent formal rules for the structure of sentences. With the help of syntax, words are able to refer to other words, but this is not enough. Thinking is manipulation of meanings and therefore the used mental patterns, symbols, must ultimately refer to their intended meanings.

Technically, the Symbol Grounding Problem involves three issues; first, how to generate a symbol, second, how to attach a meaning to it, and third, how to manipulate the meanings, not only the symbols.

5. Strong AI and Consciousness

Strong AI, also known as AGI, is supposed to produce human-like cognition; free thinking and reasoning without area limitations. Humans understand what they are thinking about because thinking operates with meanings. AGI is also supposed to understand what it is doing, but without meanings that will not happen. Therefore, in working AGI systems, the Symbol Grounding Problem has to be solved.

The fact that meanings cannot be imported into physical symbol systems in the form of additional symbols leads to the conclusion that the imported meanings must be in the form of non-symbols. Harnad [1990] and others proposed that the Symbol Grounding Problem could be solved by grounding the meaning of symbols in non-symbolic representations of the external world, produced by sensory processes. This principle would appear to be a straightforward one, but problems remain, see for

instance Taddeo and Floridi [2005]. How could a digital computer accommodate both non-symbols and symbols? It may not.

It has been argued that the Symbol Grounding Problem cannot be solved in digital computers because they can only accept information in symbolic (usually numerical) form [Haikonen, 2019]. The sensory information acquired by cameras, microphones, etc. must be digitized into streams of binary numbers. The acquired information is now in symbolic form, but the intrinsic meanings are lost. Naked numbers do not convey external meanings.

Recently, it has been proposed that in *other than digital computers*, the Symbol Grounding Problem can be solved by importing external meanings in the form of self-explanatory sensory information [Haikonen, 2019]. In humans, self-explanatory forms of sensory information would appear as qualia. Qualia are qualitative appearances of direct sensory percepts and the virtual percepts of mental content like inner speech, imaginations, pain and pleasure. Qualia are the meaning (red is red, pain is pain), no interpretation is necessary.

The percept of red color is red, the percept of a circle is a round pattern. The percept of a heard word is a sound pattern. As such, the mind takes them as actual properties of the outside world, even though they actually are sensory responses to these.

Next, the transition to symbolic processing is required. This can be done by associating percepts with percepts. For instance, a sound pattern may be associated with a sensory percept of an entity or action. Thereafter, the sound percept would also stand for the associated entity, while appearing as the percept of its original stimulus. This process also detaches the associated percept from its temporality; the associated meaning can now be evoked by its symbol also when it is not sensorily present. The Symbol Grounding Problem is solved [Haikonen, 2003, 2007, 2019].

What has consciousness to do with this? Introspection shows that the presence of reportable qualia constitutes the content of consciousness. This situation is demonstrated by the fact that when all qualia vanish, like in deep dreamless sleep, also consciousness vanishes, because there will be nothing to be conscious of; the content of consciousness will be void.

Self-explanatory forms of sensory information that are used to solve the Symbol Grounding Problem in cognitive machines would also appear to have the form of qualia. If, indeed, this were the case, then the cognitive machines that utilize self-explanatory information would be conscious in the aforestated sense, even though their qualia would not necessarily be similar to human qualia. This leads to the hypothesis about an unavoidable connection between true AGI and consciousness.

Would it be possible to create conscious AI by computer programs? Can a computer have qualia? Qualia are subjective sensory experiences, not numbers, and therefore they are neither computable nor importable to computers. Therefore, it seems that true human-like AGI cannot be achieved solely by symbolic computations and digital computers. Different kinds of systems and machinery would be needed.

6. From Networks of Neurons to Networks of Meanings

The brain is a neural network that consists of a very large number of brain cells, neurons, and their connecting points, synapses. The neurons communicate with each other via synapses and form large interconnected networks. These can be modeled and realized artificially, either as computer simulations or as actual electronic hardware. The earliest neuron models were devised by McCulloch and Pitts Jr. [1943] and Rosenblatt [1958]. The principles of Rosenblatt's Perceptron model are still used in many applications.

A simple artificial neuron is a threshold device, which receives a number of input signals via attenuators, artificial synapses, and produces an output signal if the sum of the attenuated (weighted) input signal intensities is higher than a set output threshold. The attenuation (the synaptic weight) of each artificial synapse is adjustable, and can be controlled by different means.

Properly adjusted synaptic weights allow neurons and neuron networks to respond only to certain input signal patterns, and in this way, they can be used as pattern classifiers and recognizers. The adjustment of the individual synaptic weights usually involves the tweaking of the weights against each other, by supervision or self-learning. For multilayer neural networks, various synaptic weight adjustment algorithms exist, such as Back Propagation and Deep Learning.

Neural networks can be made to detect and label low-level feature patterns and high-level combination patterns, and this information may be used by computer programs for the production of useful results. Is the Symbol Grounding Problem thus solved and does the neural network−computer program combination operate now with meanings? For instance, neural networks may be used for the detection of words in the stream of heard speech, but the crucial question is; does this lead to the capture of symbols or the capture of meanings? Will the computer program now manipulate meanings instead of symbols?

Natural language uses words and their syntactic combinations, sentences, to describe situations and actions. In order to understand what a sentence means, the words and their syntactic combination must evoke a mental qualia-based "image" or "experience" of the described situation; a multimodal mental model [Zwaan and Radvansky, 1998; Haikonen, 2003]. Ideally, this "image" would be rather similar to the one that would be evoked by the sensory act of actually experiencing the situation. The presence of these mental models and "imagery" should be obvious to anyone reading a novel.

Thus, it is not enough simply to detect the words and the syntactic structure of the sentence; the meanings must be resolved. This can be done by the associative evocation of relevant already learned information. Suitable mental "images" or "experiences" cannot be evoked, if they have not been acquired earlier. It should be noted that here "experiences" may involve dynamic conditions and reactions of the experiencer.

Pattern detection alone does not constitute understanding. In addition to pattern detection another function is required, namely the associative linking of real and virtual sensory percepts, qualia. This calls for neurons that can both detect patterns and link them associatively. Examples of these neurons exist, e.g. Haikonen [2019].

Cognition does not arise from the linking of symbols but from the linking of meanings. Artificial cognitive neural networks should not only be associative networks of neurons, but they should also amount to associative networks of meanings.

7. Emotional Robots

Autonomous robots should be able to operate without supervision in various, perhaps unpredictable environments. Therefore, the robot should be self-motivated and be able to determine the significance and urgency of environmental situations and conditions. These requirements are related to the Frame Problem, attention control and motivation.

Associative networks of meanings are able to accumulate a large number of possible associative connections between entities. Consequently, a cue like a sensory percept would be able to evoke a large number of associations. This leads to combinatorial explosion and to the Problem of Choice; which associations would be relevant to the framework of the current situation? This problem is also known as the Frame Problem.

Dennett [1987] has provided an example of the Frame Problem: A robot enters a room to retrieve a given object. In the room, there are a large number of various objects, including a bomb with a burning fuse. Which object should be attended to?

Generally, context would frame and limit the scope of choice and would solve the Problem of Choice, but in Dennett's example, something out of the context would be more important. Obviously, context is not enough. A general mechanism that is able to evaluate instantly the significance of each perceived object and situation is needed. This evaluation should be able to override the context-related attention, and focus attention on the more significant percepts. In associative neural networks, this action would call for another network for attention control in the form of neural threshold control lines.

Percepts have qualia-based self-explanatory meanings and also learned associatively connected meanings. However, for the solution of the Problem of Choice, another additional meaning is required, namely, the emotional significance. All real and virtual percepts should have an attached good−neutral−bad significance, an emotional value that guides attention and initiates proper reactions. A robot with good−bad significance evaluation would approach good objects and would try to execute good tasks, while bad objects and actions would be avoided. The good−bad significance could also be used to motivate the robot to execute desired actions also in unpredictable conditions.

How would the meanings of good and bad be established in the first place? In humans, this takes place basically via the perception of pain and pleasure as well as the perception pleasant and unpleasant tastes and odors. Also match and mismatch conditions seem to generate pleasure and displeasure; pleasure follows, when an expectation is met, and when it is not met, disappointment and displeasure follows. Similar mechanisms could be used in robots, both with and without supervision. The inclusion of these mechanisms would amount to some kind of machine emotions, not too much different from the human ones.

8. Cognitive Architectures for Artificial Minds

The human brain, the senses and the various response systems including muscles form an integrated system that is able to think, control the body and produce actions and behavior in planned interactive and communicative ways. Autonomous robots should have a similar cognitive system. A number of architectures for artificial neural and symbolic cognitive systems have been devised, see e.g. Samsonovich [2010].

Systems that understand must operate with grounded meanings. Therefore, a proper cognitive architecture has to be a perceptive system with multimodal sensors that produce self-explanatory percepts. Symbols are not self-explanatory, repeating neural activity patterns may be, if they are sensorily derived from the real world. Therefore, a proper cognitive architecture should be a neural one, one that is able to support the networking of meanings. One example of these kinds of architectures is the author's neural associative Haikonen Cognitive Architecture (HCA) [Haikonen, 1999, 2003, 2007, 2019].

The HCA architecture consists of parallel perception/response feedback loops with associative memories for each sensory modality, including internal sensors for gaze direction, body positions and muscle tensions. The perception/response feedback loops produce non-symbolic sensory percepts of their kind and also virtual percepts of the mental content via the feedback loops. Real and virtual percepts are of the same kind, and can be matched against each other resulting in reportable match, mismatch and novelty conditions [Haikonen, 2014]. These can be used to guide attention. Somewhat similar perception feedback loops have been proposed also by Chella [2008], Hesslow [2002] and Steels [2003].

The feedback loops of each sensory modality are associatively cross-connected with the outputs of the feedback loops of other modalities. These cross-connections allow sensorimotor integration, the association of additional meanings with percepts, the seamless transition from sub-symbolic to symbolic processing and the emergence of a grounded natural language for communication and inner speech.

The HCA utilizes distributed representations [Hinton *et al.*, 1986], where an entity is represented as a combination of feature signals. This method facilitates the recognition of an entity also when its percept only reminds of the actual entity. Distributed representations allow mental modification of imagined objects by varying their features.

The HCA utilizes neurons that can both detect patterns and link them associatively. Parallel and serial auto-association and hetero-association (cross-association) of neural patterns and sequences are used for learning, memorization and linking.

Fast associative learning is used in short-term memories and in emotional learning. The System Reactions Theory of Emotions [Haikonen, 2003, 2007, 2019] is used for the solving of the Problem of Choice in attention and responses.

As a feasibility test, the HCA has been implemented in a minimalist way in the author's XCR-1 robot[a] [Haikonen, 2011, 2019].

9. Conclusions

AI has not yet reached its full potential. Existing Weak AI has a shortcoming that limits its power; it operates algorithmically without the utilization of meanings. Without meanings, AI cannot understand anything, and without understanding, there cannot be any real intelligence. This is a real challenge especially for autonomous robots, which should be able to operate in rather unpredictable everyday environments, as easily as humans do.

General Artificial Intelligence would be the ultimate state of AI. It would have to operate with meanings, but there is a problem. Existing computers operate only with symbols; symbols refer to symbols, not to their meanings. Moreover, the meaning of symbols cannot be ultimately defined by additional symbols. Therefore, meanings must be imported in a self-explanatory way. For this purpose, novel technical approaches and system architectures are needed.

Qualia are self-explanatory information. If this holds also the other way around, then machines with self-explanatory information have qualia. This leads to an interesting philosophical question. To have reportable qualia is to be phenomenally conscious. Therefore, would machines that use self-explanatory information for symbol grounding be phenomenally conscious?

The history of technology shows that more of the same is usually good. But it also shows that real breakthroughs can only come from major paradigm changes. A wrong way may look promising, but it will not lead to the desired destination, no matter how far you go. More research (and money) and outside-the-box thinking are still needed for the perfection of AI, and many remarkable inventions and great opportunities await the brave.

References

Chella, A. [2008] Perception loop and machine consciousness, APA *Newsl. Philos. Comput.* **8**(1), 7–9.

Dennett, D. [1987] Cognitive wheels: The frame problem of AI, in C. Hookway (ed.), *Minds, Machines and Evolution* (Baen Books), pp. 41–42.

Fodor, J. [1975] *The Language of Thought* (Crowell).

[a]Updated demo videos of the XCR-1 robot can be viewed at https://www.youtube.com/user/PenHaiko.

Haikonen, P. O. [1999] *An Artificial Cognitive Neural System Based on a Novel Neuron Structure and a Reentrant Modular Architecture with implications to Machine Consciousness*, Doctoral Thesis. Series B: Research Reports B4. Helsinki University of Technology, Applied Electronics Laboratory.

Haikonen, P. O. [2003] *The Cognitive Approach to Conscious Machines* (Imprint Academic).

Haikonen, P. O. [2007] *Robot Brains: Circuits and Systems for Conscious Machines* (John Wiley & Sons).

Haikonen, P. O. [2011] XCR-1: An experimental cognitive robot based on an associative neural architecture, *Cognitive Comput.* **3**(2), 360–366.

Haikonen, P. O. [2014] Yes and No: Match/mismatch function in cognitive robots, *Cognitive Comput.* **6**(2), 158–163, doi: 10.1007/s12559-013-9234-z.

Haikonen, P. O. [2019] *Consciousness and Robot Sentience*, 2nd ed. (World Scientific).

Harnad, S. [1990] The symbol grounding problem, *Physica D* **42**, 335–346.

Hesslow, G. [2002] Conscious thought as simulation of behaviour and perception, *Trends Cognitive Sci.* **6**(6), 242–247.

Hinton, G. E., McClelland, J. L. and Rumelhart, D. E. [1986] Distributed representations, in D. E. Rumelhart & J. L. McClelland (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*: *Foundations* (MIT Press), pp. 77–109.

McCulloch, W. and Pitts, W. [1943] A logical calculus of the ideas immanent in nervous activity, *Bull. Math. Biophys.* **5**, 115–133.

Newell, A. and Simon, H. [1976] Computer science as empirical inquiry: Symbols and search, *Commun. ACM* **19**(3), 902–915.

Rosenblatt, F. [1958] The Perceptron: A probabilistic model for information storage and organization in the brain, *Psychol. Rev.* **65**(6), 386–408.

Samsonovich, A. V. [2010] Toward a unified catalog of implemented cognitive architectures, in A. V. Samsonovich, K. R. Johannsdottir, A. Chella & B. Goertzel (eds.), *Biologically Inspired Cognitive Architectures 2010* (IOS Press), pp. 195–244.

Searle, J. R. [1980] Minds, brains, programs, *Behav. Brain Sci.* **3**(3), 417–457.

Steels, L. [2003] Language re-Entrance and the "Inner Voice", in O. Holland (ed.), *Machine Consciousness* (Imprint Academic), pp. 173–185.

Taddeo, M. and Floridi, L. [2005] Solving the symbol grounding problem: A critical review of fifteen years of research, *J. Exp. Theor. Artif. Intell.* **17**(4), 419–445.

Zwaan, R. A. and Radvansky, G. A. [1998] Situation models in language comprehension and memory, *Psychol. Bull.* **123**(2), 162–185.