State of the art in achieving machine consciousness by using cognitive frameworks

Marius Marten Kästingschäfer
Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands

This review paper examines the developments in the field of artificial and machine consciousness, and focuses on state of the art research and the application of cognitive frameworks. This paper examines to what extent computer science has made use of existing theories regarding consciousness. Different projects using computational models of consciousness are compared and analyzed. It investigates if current artificial networks should be called conscious. The paper concludes that we are a long way from artificial consciousness comparable to the human brain.

Keywords: Artificial Consciousness; Consciousness; Computational model; Review

"*Would the world without consciousness have remained a play before empty benches, not existing for anybody, thus quite properly not existing?*"

[Erwin Schrödinger, Cambridge, 1958]

1. Introduction

The use of cognitive and neural models in information technology has become increasingly popular during the last decade. Cognitive models or frameworks are explaining human behavior as an algorithm. This paper deals with how these algorithms could be applied and used to develop a new system. Especially in the field of engineering it was concluded that nature, through natural selection, already found solutions for many problems. An example for such a problem would be how humans should orientate in a complex environment. Part of nature's strategies for this problem is the integration of multiple senses and to make sense out of social cues. Solutions like this only need to be understood and copied in order to achieve comparable intelligent or efficient solutions. Applications such as computer vision, face recognition, natural language processing and autonomous driving have gained inspiration from bioinspired frameworks. Computer science can rely on frames used in neuroscience, psychology, cognitive science and philosophy. Conversely, cognitive scientists can build computational models to test and advance their models. Replicating systems pushed the boundaries of state of the art technology further. Developments in machine learning (ML) such as reinforcement, supervised and semi-supervised learning made these applications possible (Kaelbling, Littman, & Moore, 1996). These learning models were able to crunch and make sense of the large amount of data necessary for tasks such as computer vision. Comparable groundbreaking inventions were made in developing (deep) artificial neural nets (ANNs) inspired by biological neural networks (Dayhoff & DeLeo, 2001). ANNs are distributed knowledge representations, based on artificial neurons called nodes. The nodes are connected and able to transmit signals from one to another. ANNs is the umbrella term and different specifications exist. Convolutional neural networks (CNNs) for example were inspired by the organization of the visual cortex. Regulatory feedback networks algorithms and recurrent neural networks (RNNs) were influenced by animal sensory recognition systems and the human brain. The potential benefit of using neurobiological influenced frameworks is no longer academia-based anymore but is also encountered in software company's such as Google (Hassabis, Kumaran, Summerfield, & Botvinick, 2017). Machine learning for example is today widely used in natural language processing, image detection and fraud detection. Speech based assistants like Siri, Cortona or Alexa in recent years made these developments available for everyone.

These speech assistants are not yet able to pass the Turing test and so still distinguishable from humans. Nevertheless, machines act more human-like than ever before. Rising investments and interest in artificial intelligence have leveraged these developments and have led to questions regarding machine consciousness. Since the human brain is the best working conscious system known, its operating principle could be used to build conscious machines. During the last century, science and especially engineering tried to rebuild every algorithm they understood. The widespread use of ANNs shows that understanding a concept is often only the first step, followed by rebuilding, and improving the concept. There is no reason to believe that the understanding of the conscious mind as algorithm will differ. This development of understanding consciousness in mathematical terms has already begun. Efforts to build computational models of consciousness have increased during the last years, creating a field known as artificial consciousness. Models used in this field largely fall into five categories (Reggia, 2013) and are based on what the theories select as most fundamental to consciousness: a global workspace, information integration, internal self-model, higher-level representation, and attention. Particularly from theories such as the global workspace theory from Baars, (1997, 2005) dealing with the *theatre* of consciousness, the multiple drafts model by Dennett, (1991, 2001), Cohen, Cavanagh, Chun, and Nakayama, (2012) dealing with information processing and especially the integrated information theory by Tononi, (2008, 2012) dealing with the question which processes become conscious, inspiration is gained. These theories set benchmarks from which frameworks and models could be derived. They further give a general guideline how consciousness should be approached, examined and assessed. It can be concluded that engineers will extract the essence of these models and use them to build models of conscious machines. Scientific efforts focus on distill the essence of consciousness.

The relation between the brain and consciousness is heavily investigated in neuroscience today (Atkinson, Thomas, & Cleeremans, 2000; Block, 2005; Crick & Koch, 2003; Koch, Rees, & Kreiman, 2002). Empirical evidence gained from electroencephalography (EEG), positron-emissions-tomography (PET) or functional magnetic resonance imaging (fMRI) shed lights on the neural correlates of consciousness. Machine consciousness could have a significant positive impact on real-world challenges. Rising amounts of data, higher complexity and faster technological progress can only be processed by better algorithms, that’s why they are useful and heavily needed. The likelihood of negative consequences is depending on how society will make use of the new technology. China for example is using artificial intelligence to cut its citizens privacy, whereas the company Deepmind uses similar technology to advance mammography screening for breast cancer. The impact on society is a double-edged sword, however the decision where to put the boarder is beyond the scope of this paper. Nevertheless, within the next decade the so called technological singularity (invention of artificial super-intelligence that leads to unpredictable consequences (Kurzweil, 2005; Chalmers, 2010) followed by a *terminator scenario* is unlikely. Within the next decades hardware constraints will still limit technologies ability.

Until now it was explained how cognitive ideas and biological systems influenced computer science, the important consciousness theories are named, and the social impact of the topic is discussed. The next section, outlines past work on artificial consciousness and provides more detail about frameworks mainly used to approach consciousness. Some neurobiological correlates that have inspired computational models are discussed. Different computational models are compared. Intelligent Dispatching System (IDA), the advanced model learning IDA (LIDA), the neurocontroller, the virtual machine architecture, the Adaptive Resource-Allocating Vector Quantizer (ARAVQ), and CRONOS. In addition, other models are introduced, and advantages and limitations of current models are discussed. In the end, an overview about the findings of this review are given and the implications for further developments of the field are discussed.

2. Measurement of consciousness

As already mentioned the field of artificial consciousness unifies different scientific disciplines. That makes it necessary to predefine some terms. First, the distinction between stimulated and instatantiation consciousness is important. Stimulated consciousness is functional and referring to the easy problem of consciousness (Chalmers, 1996). Instatantiation consciousness is phenomenal and referring to the hard problem (Chalmers, 1996, 2007). The easy problem is concerned with the question of how the brain functions, for example, how humans discriminate stimuli. The hard problem on the other side is asking the question of why these physical processes are accompanied by subjective experience. This distinction is comparable to the distinction made between strong and weak AI (Seth, 2009). Artificial consciousness is a heterogeneous research area including numerous fields without unified goals. To measure machine consciousness this paper will in general follow the classifications by Gamez (2008). The classifications are used to compare different systems. Different classifications set different thresholds, some easier to achieve than others. The Measurements of consciousness (MC) are: MC1 external behavior associated with human consciousness, MC2 cognitive characteristics associated with human consciousness, MC3 architecture with cause or correlates of human consciousness, and MC4 phenomenal conscious machines. This categorization could be seen as ordinal levels, each of which qualitatively closer to real artificial consciousness.

MC1 External behavior associated with human consciousness. A limited number of behavior is unconsciously processed. More complex activities such as introspection and expressing of complex feelings are an example of conscious behavior. These behaviors are only seen in systems that appear to be conscious. Criteria MC1 is met when a computer or another system replicate aspects of human behavior associated with consciousness. This criterion is, unlike the other three criteria, often also passed by models that aim for general intelligence. Intelligence in AI research is based on specific behavioral criteria which can be assessed by the Turing Test (Turing, 1950). The Turing Test is best equipped to measure *shown behavior* but unable to measure underlying processes. A machine could pass the Turing Test when a human is unable to distinguish between the machine and another human. Current systems do not pass the Turing test, but up-coming generations of Siri or Alexa might fulfill MC1.

MC2 Cognitive characteristics associated with human consciousness. Unlike MC1 Siri or Alexa are not build to fulfill the requirements for MC2. Criterion MC2 includes computers and systems with the cognitive characteristics associated with consciousness. Examples would be, internal models of the system’s body, emotions, and imagination. MC2 is often accompanied by criteria MC1 and MC3 but could also be created without them in a model with only internal behavior. MC2 does not imply MC4 (stimulated water is not wet and stimulated fear is not phenomenological fear).

MC3 Architecture with cause or correlates of human consciousness. Criterion MC3 is the simulation of architectures linked to human consciousness. These simulations often arise from the desire to model or test theories of consciousness. Notably models computing the neural correlates of consciousness fulfill this requirement. Data involving MC3 regularly overlaps with MC1 and MC2 data.

MC4 Phenomenal conscious machines. Criterion MC4 captures an artificial conscious system with phenomenal awareness, the ability to be aware of own feelings. This approach is the most controversial on since it is philosophically problematic. It is unclear whether MC4 is achievable independently of MC1-MC3. Only this approach is related to instantiation consciousness.

3. Past work on artificial consciousness

Measurements of consciousness show how systems can be compared (i.e., what is a system able to do?). The next part shows which broader categories of systems exist (i.e., where do they belong?). To structure the overall review of past work on artificial consciousness, the fundamental categories most central to consciousness proposed by Reggia (2013) will be explained. These categories are based on cognitive theories. Engineers made use of these theories and build their models within the framework provided by these theories. Categories are named after the models they contain, namely:

1. a global workspace
2. information integration
3. an internal selfmodel
4. higher-level representation
5. attention mechanisms

In this review higher-level representation and attention mechanisms are only mentioned briefly. Some additional frameworks and models worth to mention will be named under the category ’Other models’. For each of the broader categories a theoretical overview of the theory will be given, followed by a review of the computational models using these frameworks and an assessment of the level of consciousness on the classification (MC1-MC4).

3.1 Global workspace model

3.1.1 Theory In the global workspace model, different parallel processes compete to place information in the global workspace (Baars, 1988, 2002). Baars referred to this global workspace as the *theatre* of consciousness. The model views modular specialized and automatic processors responsible for information processing, motor control, language, and so forth. The processing is located in different human brain regions and is mainly unconsciousness. The specialized modules compete to make their information available in the interconnected global workspace. Information reaches consciousness in this model when a certain threshold of activity is crossed. The model is often extended and connected with the thalamo-cortical system in the human brain (Newman, Baars, & Sung-Bae, 1997). The thalamo-cortical system (Koch, Rees, & Kreiman, 1998, 2011) and the ascending arousal system are closely associated with human consciousness (Posner, Saper, Schiffer, & Plum, 2007).

3.1.2 Application The global workspace account is particularly well suited for the idea of artificial consciousness since there is nothing intrinsically biological about the global workspace. The first system inspired by the global workspace theory is the Intelligent Distributed Agent naval Dispatching System (IDA) (Baars & Franklin, 2007; Franklin, 2003; Franklin & Graesser, 1999). IDA is a multi-agent system consisting of the processing modules defined by the global workspace hypothesis. The system was used for interactions with databases, natural language conversation and checks on job requirements. A typical task is to check on job requirements and the in-take of personal data via listening to conversations. The IDA system is not actively used anymore. Agents in IDA communicate via a global workspace, also called coalition manager or spotlight controller. This makes the system functionally conscious (Franklin, Strain, Snaider, McCall, & Faghihi, 2012). The system produces behavior requiring consciousness in humans (MC1). Further the system has some cognitive characteristics associated with consciousness (MC2) and an architecture linked to human consciousness (MC3). It remains unclear if IDA is phenomenally conscious (MC4), but it is not ruled out entirely, because it could not be tested. The second system builds up on the advanced version of Baars’ main idea, called learning IDA (LIDA). It is the successor of IDA but it differs in fundamental matters; this is why it is examined separately. The LIDA framework is a neural network model and incorporates more biologically-realistic features. This enables the model to predict human correlates of behavior better. LIDA became advanced and merged into a system that is called Neural Simulations of Global Workspace also known as the neural global Workspace Model (ngWM) (Dahaene, Kerszberg, & Changeux, 1998; Dahaene & Naccache, 2001). NgWM derived from an earlier model that was developed for solving a Stroop task. The Stroop task is used to demonstrate the effect of interference on reaction time. Even more recent neural models also incorporated spiking neurons, which make them more biological plausible. The ngWM consists of response units, input, global workspace neurons, reward and vigilance system to modulate activity in the global workspace. Reinforced learning based on Hebbian weight change is added to train the network, this enables machine learning and reduces the amount of hand-crafted knowledge the system needs. The model is able to make predictions about brain patterns measured during conscious effortful tasks (Dahaene et al., 1998), the attentional blink (Dahaene & Naccache, 2001) and the spontaneous activity during the inattentional blink (Dahaene & Changeux, 2005). The power to predict certain brain activities correct is taken as evidence for the proper method of its operating. Especially since the predicted brain patterns are meaningful for the understanding of consciousness. This work is an example of MC3, it is consistent with evidence that widespread cortical activation correlates with conscious brain activity and provides a clear statement about on how to differentiate conscious from unconscious brain states (global distribution and localized distribution). This is in line with for example Lamme (2003), found recurrent processing and interaction between brain parts as essential for conscious processing.

3.2 Integrated information theory

3.2.1 Theory Integrated information theories view information processing and integration as the integral element in explaining consciousness (Agrawala, 2012; John, 2002). Tononi's Integrated Information Theory (IIT) takes this approach, starting from essential phenomenal properties of experiences. IIT proposes that every physical system that integrates information is conscious (Tononi, 2008, 2004). Overall, consciousness for IIT is not an all-or-none property, but graded, measured with φ. A network's φ value is assessing interactions between the network's components, not statistical dependency only (Tononi & Sporns, 2003). Distributed connectivity in a network was found to maximize information integration, an idea also consistent with the global workspace theory. Computational work within this frame- work is greatly influenced by the IIT. The quantitative measure φ seems to make the concept applicable and sets a valuable threshold. However, the practical calculation remained an unsolved problem. Gamez (2008) calculated that applying Tononi's measurement of φ in a network of 18.000 neurons would take up to 10^{9000} years. This makes it unlikely to be able to compute φ for the entire human brain in the foreseeable future (Gamez, 2010), even if projects like the Blue Brain project might come up with rough estimations (Markram, 2006).

3.2.2 Application Gamez (2010) built a neurocontroller using a spiking neural network and a robotic vision system. The network was modelled using the SpikeStream stimulation, which directs the eyes of SIMNOS, a software based virtual robot. The trained network's distribution of integrated information was examined to assess the extent to which the system, or subsystems, could be considered as conscious. Activity in the system was gated by activity through emotion and inhibition regions. This made the system more psychological plausible. The region with the highest φ value (103) involved 91 neurons and included all inhibition regions and most of the emotion region. Here subsystems were serving as gating circuits. Gating in neural architectures might play an important role in conscious processing (Regan, 2012). Gating in this context refers to a systems ability to subjectively experience only certain stimuli by focusing on them. Is not possible to address the question regarding how much consciousness was present with the value of 103 (e.g. 10% of an average waking human brain) (Gamez, 2010). This makes it difficult to interpret the obtained results. MC1 is limited since the model focused on internal behavior. The architecture used by Gamez is roughly inspired by the brain (MC2 and MC3). In Tononi's terms the model would be conscious (MC4), if this consciousness is comparable to human consciousness, still remains questionable. Furthermore, Tononi himself asked the question how machines could be conscious (Koch & Tononi, 2008). He emphasizes the role of the hardware; the models were conscious only on neuromorphic devices and not on commonly used Von Neuman architectures. Von Neuman architecture is based on silicon-based semiconductor transistors, whereas neuromorphic devices are more organic. Due to progress in developing and designing silicon neurons and synapses (Indiveri, Chicca, & Douglas, 2006; Wijekoon & Dudek, 2008) and advancements in multichip systems to communicate using neurone-spike-like signals (Chicca et al., 2007) it might be possible to run models on biological plausible neuromorphic devices in the future.

3.3 Internal self-models

3.3.1 Theory For internal self-models a key component of consciousness is a self-model based on an internal representation of the body (Metzinger, 2003), called body image, which often is related to the sensorimotor cortex (Goldenberg, 2003). The body image is a virtual model, an adaptable inner representation of the physical body. Metzinger (2000a, 2000b) argues that the conscious experience of a self arises due to the omnipresence somatic and proprioceptive input that is constantly fed into the internal self-model.

3.3.2 Application Embodied agents with an internal model encompassing the external environment, including a self-model, were proposed to have conscious experience. Models using the internal self-model framework further often apply a dynamic system approach. Robots within this approach need to discover the properties of their body and environment themselves, starting with as little rules as possible (e.g. repeated hypothesize-and-test cycles). The first type of model using an internal self-model is the virtual machine architecture (Sloman & Chrisley, 2003). The conscious 'human' mind in this model is seen as a virtual machine executed by the brain, comparable to software and hardware. Phenomenal consciousness in this model is seen as a side-effect of the virtual machine executing different components. The second model is the Adaptive Resource-Allocating Vector Quantizer (ARAVQ) used to build models of sensorimotor data from a Khepera robot. Distinct combinations of sensory inputs and motor outputs called concepts were used to store sequences of experiences economically. Economically in this case refers to make use of the lowest amount of resources, an example would be to filter incoming stimuli and to store only the relevant once. Holland and Goodman (2003), were able to show that an internal model could control the robot precisely, process novel data and detect anomalies. The experiment gave evidence for the importance of internal models and proved that they could be studied in simple systems with- out the need for complex supercomputers. This work is an example of MC2 since some internal models are integrated into conscious cognitive states. If the assumption of internal self-models is correct, complex systems could contain phenomenal states (MC4). The third model within this framework is called CRONOS, a large project explicitly funded to work on machine consciousness. CRONOS is a robot based on the human musculoskeletal system, running with a software called SIMNOS. SIMNOS is a physics-based simulation of the robots environment inspired by the visual system. Further CRONOS uses SpikeStream for simulating spiking neural activity (Holland, 2007). Simulating spiking neural activity is used to achieve artificial activity comparable to neuronal brain activity, in this sense it makes the model more realistic. CRONOS uses a self-model intended to process sensory information. Holland (2007) claimed that this process reflects the cognitive contents of human consciousness. Overall the project focuses on MC2-4. The network contains cognitive characteristics of consciousness (MC2) and the basic architecture is based on neural correlates of consciousness (MC3). It remains unclear if the model is really phenomenally conscious (MC4), which is the same challenge faced by other models in the field. However, the developer of CRONOS stated that the model is not complex enough to reach a level of consciousness comparable to the human brain.

3.4 Other models and applications

The first category of the other models uses the higher-level representation (also sometimes referred to as Higher-Order Thought (HOT) theories) framework. Conscious mental activity is rather seen to involve a higher-level representation than as an unconscious mental activity. Important models are; the CLARION system using a representational difference approach (Sun, 1997, 1999; Sun & Franklin, 2000b) and CERA-CRANIUM using simulated and physical robots (Arrabales, Ledezman, & Sanchis, 2009). The second category of other models uses the attention mechanisms model as framework proposing that a person is only consciously aware about a fraction of the ongoing stream of incoming information. The attention mechanisms responsible for selecting are closely linked to consciousness, in humans and the model (Koch & Tsuchiya, 2006). A Model working with this framework is Corollary Discharge of Attention Movement (CODOM) based on an engineering control theory of attention (Taylor, 2007). An example of engineering control of attention is to exclusively focus on stimuli currently necessary for the task. Comparable is Haikonen's (2012) model based on different properties, for example inner speech and the ability to report on inner states. The model is consistent with neurobiological evidence. Thirdly, the world's largest brain simulations should also be taken into account when looking at systems that might be conscious. The Blue Brain Project is one out of many projects that should be more deliberately examined regarding their possible artificial consciousness. Markram's (2006) Blue Brain Project uses an IBM "Blue Gene" supercomputer and simulates neural signaling in a cortical column of the rat brain at ion-channel level of detail. The long-term goal for the project is a simulation of the human cortex. It is unclear whether rebuilding the lowest level necessarily lead to the emergence of more complex behaviors or even behavioral correlates of consciousness.

4. Discussion and conclusion

The cognitive frameworks building on the current understanding of the human brain were applied in a wide range of systems. The systems made good use of the fundamental ideas of the models. Today not only academia makes use of cognitive models but also industry. Company's like Google and Microsoft fast adopted machine learning techniques based on biological systems. Consequently, they are also gone be the first once who will integrate conscious frameworks into their devices to further expand Alexa or Cortana. Concerning the review of past work on artificial consciousness, three conclusions seem appropriate. First, currently no approach on machine consciousness is convincing. None of the existing systems is fulfilling the broad array of applied criteria. Systems mainly fail to model the neural complexity of the brain due to the small scale of simulations. Even the most optimistic estimates assume that we are currently not able to simulate more than 5% of the human brain. Even a simulation of a rat's brain is only within reach at the end of the decade. This is due to the problem of stimulating the interconnectivity and plasticity of the brain, especially under the power consumption restraints set by a real brain. Models were still able to emulate, capture and replicate some of the main predictions, findings and neuronal correlates. Making computational modeling of consciousness is a feasible method to some insides into stimulation underlying biological process. Second, the main problem with using different frameworks to achieve machine consciousness is that every criterion brings the applied measures itself. For example, a project build on Tonoi's integrated information theory, might be conscious within the framework itself but only because the models set the threshold that needs to be passed. Results are predetermined in a way that models were measured on thresholds within the model itself. This issue could be solved by finding a more widely accepted definition of consciousness and general threshold. A solution for this would be to combine current approaches of consciousness into a unified theory of consciousness. The main theories stated in this paper are not exclusive, but complementary to each other. Converging to the same question from different angles, only in sum the sum of theories might be able to grasp a phenomena as complex as consciousness. The next generation of theories might fulfill this goal. Third, acting or access consciousness is possible to simulate, being consciousness is questionable as long as the explanatory gap remains unsolved. A solution for the easy problem of consciousness is necessary, but even more relevant, since it is more fundamental, is to find a solution for the hard problem. The integrated information theory might be a first step into solving this problem. Overall, human kind is a long way from artificial consciousness comparable to the human brain. Also the philosophical question remains in how far systems can be called artificial intelligent or consciousness if they are coded by humans.

5. Future work on artificial consciousness

Computational modeling and machine learning are fast growing fields and artificial consciousness will benefit from developments in both. Furthermore, increasing computational power and especially the invention of more efficient algorithms will spur the developments. Also, a better understanding of the brain in neuro- and cognitive science might lead to better empirical data, more specific models and detailed frameworks will yield some benefits for computational modeling in general and for artificial consciousness in particular. Machine consciousness would lead to difficult ethical questions and far reaching consequences for society as a whole. The development of conscious non-human agents might change the world beyond expectations. Currently there is no reason to belief that consciousness is limited to biological entities. Best expressed by Christof Koch (2001) "*...we know of no fundamental law or principle operating in this universe that forbids the existence of subjective feelings in artefacts designed or evolved by humans.*" In the future machine consciousness will be achieved, since one algorithm exists that is conscious already, the human brain.

References

Agrawala, A. (2012). *A framework for consciousness and its interactions with the physical world*. Toward a science of consciousness, Arizona: Tucson.

Arrabales, R., Ledezman, A., & Sanchis, A. (2009). CERA-CRANIUM: a test bed for machine consciousness research. *Proceedings of the international workshop on machine conscious- ness*, 105-124.

Atkinson, A., Thomas, M., & Cleeremans, A. (2000). Consciousness: mapping the theoretical landscape. *Trends in Cognitive Sciences*, *4*, 372-382.

Baars, B. (1988). A Cognitive Theory of Consciousness. *Cambridge: Cambridge University Press*.

Baars, B. (1997). In the theatre of consciousness. Global Workspace Theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies*, *4*(4), 292-309.

Baars, B. (2002). The conscious access hypothesis. *Trends in Cognitive Sciences*, *6*, 47-52. doi: https://doi.org/10.1016/S1364-6613(00)01819-2

Baars, B. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Progress in Brain Research*, *150*, 45-53. doi: https://doi.org/10.1016/S0079-6123(05)50004-9

Baars, B., & Franklin, S. (2007). An architectural model of conscious and unconscious brain function. *Neural Networks*, *20*, 955-961. doi: https://doi.org/10.1016/j.neunet.2007.09.013

Block, N. (2005). Two neural correlates of consciousness. *Trends in Cognitive Sciences*, *9*, 46-52.

Chalmers, D. (1996). *The conscious mind*. Oxford: Oxford University Press.

Chalmers, D. (2007). The hard problem of consciousness. *M. Velmans, S.Schneider (Eds.), The blackwell companion to consciousness*, 225-235.

Chalmers, D. (2010). The singularity. *Journal of Consciousness Studies*, *17*, 7-65.

Chicca, E., Whatley, A. M., Lichtsteiner, P., Dante, V., Delbruck, T., Giudice, P., Indiveri, G. (2007). A multi-chip pulse-based neuromorphic infrastructure and its application to a model of orientation selectivity. *IEEE Trans. Circuit. Syst. I*, *5*, 981-993. doi: https://doi.org/10.1109/tcsi.2007.893509

Cohen, M. A., Cavanagh, P., Chun, M. M., & Nakayama, K. (2012). The attentional requirements of consciousness. *Trends in Cognitive Sciences*, *16*(8), 411-417. doi: https://doi.org/10.1016/j.tics.2012.06.013

Crick, F., & Koch, C. (2003). A Framework for consciousness. *Nature Neuroscience*, *6*, 119-126. doi: https://doi.org/10.1038/nn0203-119

Dahaene, S., & Changeux, J. (2005). Ongoing Spontaneous Activity Controls Access to Consciousness: A Neuronal Model for Inattentional Blindness. *Public Library of Science Biology*, *3(5)*, e141. doi: https://doi.org/10.1371/journal.pbio.0030141

Dahaene, S., Kerszberg, M., & Changeux, J. (1998). A neuronal model of a global workspace in efforful cognitive tasks. *Proceedings National Academy of Sciences*, *95*, 14529-14534. doi: https://doi.org/10.1073/pnas.95.24.14529

Dahaene, S., & Naccache, L. (2001). Towards a cogntive neuroscience of consciousness. *Cognition*, *79*, 1-37. doi: https://doi.org/10.1016/S0010-0277(00)00123-2

Dayhoff, J. E., & DeLeo, J. M. (2001). Artificial neural net- works. *Cancer*, *91*(S8), 1615–1635. doi: 10.1002/1097-0142(20010415)91:8+<1615::AID-CNCR1175>3.0.CO;2-L

Dennett, D. (1991). *Consciousness explained*. Little, Brown, Boston, MA.

Dennett, D. (2001). Are we explaining consciousness yet? *Cognition*, *79*(1-2), 221-237. doi: https://doi.org/10.1016/S0010- 0277(00)00130-X

Franklin, S. (2003). IDA: A conscious artifact? *Journal of Consciousness Studies*, *10*, 47-66.

Franklin, S., & Graesser, A. (1999). A software agent model of consciousness. *Consciousness and Cognition*, *8*, 285-305. doi: https://doi.org/10.1006/ccog.1999.0391

Franklin, S., Strain, S., Snaider, J., McCall, R., & Faghihi, U. (2012). Global workspace theory, its LIDA model, and the underlying neuroscience. *Biologically Inspired Cognitive Architectures*, *1*, 32-43. doi: https://doi.org/10.1016/j.bica.2012.04.001

Gamez, D. (2008). Progress in machine consciousness. *Consciousness and Cognition*, *17*, *Issue 3*, 887-910. doi: https://doi.org/10.1016/j.concog.2007.04.005

Gamez, D. (2010). Information integration based predictions about the conscious states of a spiking neural network. *Conscious. Cogn.*, *19*, 209-310. doi: https://doi.org/10.1142/S1793843009000086

Goldenberg, G. (2003). Disorders of body perception and representation. *In T. Feinberg, M. Farah (Eds.), Behavioral neurology and neuropsychology*, 285-294.

Haikonen, P. (2012). *Consciousness and robot sentience*. CWorld Scientific.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, *95*, 245-258. doi: http://dx.doi.org/10.1016/j.neuron.2017.06.011

Holland, O. (2007). A Strongly Embodied Approach to Machine Consciousness. *Journal of Consciousness Studies*, *14, Number 7*, 97-110.

Indiveri, G., Chicca, E., & Douglas, R. (2006). A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity. *IEEE Trans. Neural Netw.*, *17*, 211- 221. doi: https://doi.org/10.1109/TNN.2005.860850

John, E. (2002). The neurophysics of consciousness. *Brain Research Reviews*, *39*, 1-28. doi: https://doi.org/10.1016/S0165-0173(02)00142-X

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement Learning: A Survey. *Journal of Artifical Intelligence Research*, *4*, 237-285. doi: http://dx.doi:10.1613/jair.301

Koch, C. (2001). Final Report of the Workshop "Can a Machine be Conscious". *The Banbury Center, Cold Spring Harbor Laboratory*. doi: Available from: http://www.theswartzfoundation.org/abstracts/2001$_s$*ummary.asp*

Koch, C., Rees, G., & Kreiman, G. (1998). The neuronal basis for consciousness. *The Royal Society*, *353*, 1841-1849. doi: https://doi.org/10.1098/rstb.1998.0336

Koch, C., Rees, G., & Kreiman, G. (2002). Neural correlates of Consciousness in Humans. *Nature Review Neuroscience*, *3*, 261-270. doi: https://doi.org/10.1038/nrn783

Koch, C., Rees, G., & Kreiman, G. (2011). The thalamic dynamic core theory of conscious experience. *Consciousness and Cognition*, *20*, 464-486. doi: https://doi.org/10.1016/j.concog.2011.01.007

Koch, C., & Tononi, G. (2008). Can machines be conscious? *IEEE Spectrum*, *June*, 55-59. doi: https://doi.org/10.1109/mspec.2008.4531463

Koch, C., & Tsuchiya, N. (2006). Attention and consciousness: two distinct brain processes. *Trends in Cognitive Sciences*, *11*, 16-22. doi: https://doi.org/10.1016/j.tics.2006.10.012

Kurzweil, R. (2005). *The singularity is near*. Viking.

Lamme, V. A. (2003). Why visual attention and awareness are different. *TRENDS in Cognitive Sciences*, *7 No.1*, 12-18. doi: https://doi.org/10.1016/S1364-6613(02)00013-X

Markram, H. (2006). The Blue Brain Project. *Nature Reviews Neuroscience*, *7*, 153-160. doi: http://dx.doi.org/10.1038/nrn1848

Metzinger, T. (2000a). *The subjectivity of subjective experience. In T. Metzinger (Ed.), Neural correlates of consciousness*. MIT Press.

Metzinger, T. (2000b). *Neural correlates of consciousness*. MIT Press.

Metzinger, T. (2003). *Being no one*. Cambridge Massachusetts

Newman, J., Baars, B. J., & Sung-Bae, C. (1997). A Neural Global Workspace Model for Conscious Attention. *Neural Networks*, *10, Issue 7*, 1195-1206. doi: https://doi.org/10.1016/S0893-6080(97)00060-9

Posner, J., Saper, C., Schiffer, N., & Plum, F. (2007). *Plum and Posner's diagnosis of stupor and coma*. Oxford University Press.

Regan, J. (2012). How to build a robot that is conscious and feels. *Minds and Machines*, *22*, 117-136. doi: https://doi.org/10.1007/s11023-012-9279-x

Reggia, J. A. (2013). The rise of machine consciousness: Studying consciousness with computational models. *Neural Networks*, *44*, 112-131. doi: http://dx.doi.org/10.1016/j.neunet.2013.03.011

Seth, A. (2009). The strength of weak artificial consciousness. *International Journal of Machine Consciousness*, *1*, 71-82. doi: https://doi.org/10.1142/S1793843009000086

Sloman, A., & Chrisley, R. (n.d.). Virtual machines and consciousness. *Journal of ConsciousnessStudies*.

Sun, R. (1997). Learning, action and consciousness. *Neural Networks*, *10*, 1317-1331. doi: https://doi.org/10.1016/S0893-6080(97)00050-6

Sun, R. (1999). Accounting for the computational basis of consciousness. *Consciousness and Cognition*, *10*, 529-565. doi: https://doi.org/10.1006/ccog.1999.0405

Sun, R., & Franklin, S. (2000b). *Computational models of consciousness. In P. Zelazo, M. Moscovitch (Eds.)*. Cambridge University Press.

Taylor, J. (2007). CODAM: a neural network model of consciousness. *Neural Networks*, *20*, 983-992.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, *5*, 42. doi: https://doi.org/10.1186/1471-2202-5-42

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *Biol Bull*, *215*(3), 216-242. doi: https://doi.org/10.2307/25470707

Tononi, G. (2012). Integrated information theory of consciousness: an updated account. *Archives Italiennes de Biologie*, *150*, 290-326. doi: https://doi.org/10.4449/aib.v149i5.1388

Tononi, G., & Sporns, O. (2003). Measuring information integration. *BMC Neuroscience*, *4*, 31. doi: https://doi.org/10.1186/1471-2202-4-31

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *59*, 433-460.

Wijekoon, J., & Dudek, P. (2008). Compact silicon neuron circuit with spiking and bursting behaviour. *Neural Netw.*, *21*, 524-534. doi: https://doi.org/10.1016/j.neunet.2007.12.03