Journal of Artificial Intelligence and Consciousness
Vol. 7, No. 1 (2020) 95–107

DOI: 10.1142/S270507852050006X

Artificial Conscious Intelligence

James A. Reggia

Department of Computer Science and UMIACS
University of Maryland
College Park, MD 20742 USA
reggia@cs.umd.edu

Garrett E. Katz

Department of Electrical Engineering and Computer Science
Syracuse University
Syracuse, NY 13244 USA
gkatz01@syr.edu

Gregory P. Davis

Department of Computer Science
University of Maryland
College Park, MD 20742 USA
grpdavis@cs.umd.edu

Published 29 April 2020

The field of artificial consciousness (AC) has largely developed outside of mainstream artificial intelligence (AI), with separate goals and criteria for success and with only a minimal exchange of ideas. This is unfortunate as the two fields appear to be synergistic. For example, here we consider the question of how concepts developed in AC research might contribute to more effective future AI systems. We first briefly discuss several past hypotheses about the function(s) of human consciousness, and present our own hypothesis that short-term working memory and very rapid learning should be a central concern in such matters. In this context, we then present ideas about how integrating concepts from AC into AI systems to develop an artificial conscious intelligence (ACI) could both produce more effective AI technology and contribute to a deeper scientific understanding of the fundamental nature of consciousness and intelligence.

Keywords: Artificial Consciousness; Artificial Intelligence; Computational Correlates of Consciousness; Computational Explanatory Gap; Neural Virtual Machine; Working Memory.

1. Machine Intelligence versus Machine Consciousness

In his foundational paper on machine intelligence, Turing [1950] did the emerging field of artificial intelligence (AI) a great service. At the time, just a few years after

the construction of the first electronic digital computers, there was a lot of discussion about whether or not a machine could think or have a mind. Turing found questions like these to be "too meaningless to deserve discussion", and instead proposed what we today call the Turing Test as the criterion for machine intelligence. While this specific criterion has faced well-deserved criticism in contemporary AI, Turing's basic notion that we should judge whether a machine is intelligent based on its *behavior*, rather than on vaguely defined concepts such as the existence of an underlying mind, was liberating and persists as the dominant paradigm in AI to this day. In a very real sense, Turing's approach made research into AI respectable. This is because the idea of machine intelligence per Turing only refers to whether a machine can exhibit intelligent behavior, and does not represent a claim to having created a machine that has subjective mental experiences, can think, or have a mind. Avoiding these latter difficult issues has to a great extent enabled the pursuit of and substantial successes of AI as a technology.

On the other hand, an additional consequence of this dominant viewpoint is that, with very few exceptions, it has largely side-lined work in AI on challenging issues surrounding the possibility of an artificial mind or a conscious machine. Many AI researchers find such issues to be uninteresting or insufficiently well defined to be of any relevance to AI [Bringsjord, 2007; McDermott, 2007]. As a result, the field of artificial consciousness (AC) has largely developed outside of mainstream AI [Reggia, 2013]. In our opinion, this is regrettable because there is substantial room for synergistic work in these two fields.

We have previously considered the issue of how work in AI might contribute to advancing AC [Reggia *et al.*, 2014, 2017]. Our central point in this regard is that a *computational explanatory gap* currently limits our ability to advance work in AC. The computational explanatory gap is our lack of understanding of how consciously accessible high-level cognitive information processing can be mapped onto low-level neural computations. The computational explanatory gap is a purely computational issue and not a mind–brain problem — it is a gap in our understanding of how cognitive algorithms (executive control, goal-directed problem solving, planning, etc.) can be mapped into the sub-symbolic computations supported by neural networks that use a distributed representation of information. This issue is clearly relevant to AI in general, and encouragingly increasing attention is being paid to it in studying "programmable neural networks", for example [Devlin *et al.*, 2017; Graves *et al.*, 2016]. The computational explanatory gap also makes cognitively oriented models in AI much more relevant to AC than is often recognized, especially given the philosophical concept of cognitive phenomenology.[a]

Having previously considered how work in AI may contribute to AC, here we address the converse question: How might concepts developed via work in AC and consciousness studies in general enhance the functionality of AI systems? To answer

[a]Cognitive phenomenology asserts that parts of our cognitive processes are consciously accessible above and beyond their sensory representations [Bayne and Montague, 2011]. Our other points about the significance of the computational explanatory gap hold regardless of the validity of cognitive phenomenology.

this question, we first summarize some past ideas about what the function of consciousness is in people (Sec. 2). We next present our own hypothesis that working memory and very fast learning/unlearning of its contents should be a central concern in such matters (Sec. 3), summarizing the results of some recent work we have done studying this issue (Sec. 4). With these considerations about the function of biological consciousness in hand, we then return to the question of how work in AC is relevant to the development of future AI systems (Sec. 5) and provide a summary of our conclusions on these matters (Sec. 6).

2. What is the Function of Human Consciousness?

We approach the question of what consciousness might contribute to AI systems by first asking what its function is in human cognition, and then exploring whether such functionality might provide/improve similar, currently absent/limited functionality in machine intelligence. This of course presumes that consciousness does have a biological function, an assumption that we make here. While this assumption is controversial, with some arguing that consciousness is just an epiphenomenon, we note that the evolution of consciousness in at least humans and some animal species supports the idea that it contributes to survivability and reproductive fitness, and we explore the consequences.

There is no shortage of past hypotheses concerning the function(s) of human consciousness. Here, we take asserting that something is a *function* of consciousness to implicitly indicate that a causal relationship is involved: that consciousness causes and is in large part necessary for that function. Many AC investigators have hesitated to make such causal claims and have instead proposed neural or computational *correlates* of consciousness. A neural correlate of consciousness is a minimal neurobiological state whose presence is sufficient for the occurrence of a corresponding state of consciousness [Metzinger, 2000]. A computational correlate of consciousness is a minimal computational mechanism that is specifically associated with conscious aspects of cognition but not with unconscious aspects [Cleeremans, 2005; Reggia *et al.*, 2014]. Being a function of consciousness generally implies being a correlate of consciousness, but not vice versa.[b] With this understanding, we now give a non-exhaustive listing of functions of consciousness previously proposed in the literature, ordered arbitrarily:

- global access to and integration of information [Baars, 1997; Tononi, 2008]
- symbol grounding [Chella *et al.*, 2008; Haikonen, 2019; Kuipers, 2008]
- high-level symbolic cognition [Pasquali *et al.*, 2010; Sun and Franklin, 2007]
- supports executive functions [Rosenthal, 2008; Shanon, 1998]
- error detection and correction [Baars, 1997; Taylor, 2007]
- novelty detection and generation [Baars, 1997; Mudrik *et al.*, 2012]
- self-awareness/modeling [Chella *et al.*, 2008; Holland, 2007; Perlis, 1997; Takeno, 2013]

[b]A neural/computational correlate of consciousness could be something caused by consciousness, something that causes consciousness, or neither (for example, there might be a separate underlying cause of both consciousness and the correlate).

- source of intrinsic motivation [DeLancey, 1996; Sanz *et al.*, 2012]
- evoking/informing volitional actions [Earl, 2014; Pierson and Trout, 2017]
- attention mechanisms, control [Haikonen, 2019; Taylor, 2007]

The large number of these past hypotheses is remarkable, but is consistent with the sizable number of theories about the nature of consciousness [Katz, 2013]. This is ameliorated somewhat by the fact that these hypotheses are generally not mutually exclusive or independent (e.g., symbol grounding and inference [Brody *et al.*, 2016]), and it could be that consciousness has multiple functions. Our point here is that, at the present time, there is no clear consensus on an identifiable function of consciousness that provides an adaptive advantage. For example, several of the potential functions of consciousness listed above have been criticized on various grounds due to inconsistency with empirical data or theoretical considerations [Manzotti, 2012; Mudrik *et al.*, 2012; Rosenthal, 2008; Seth, 2009].

3. Memory, Learning and Consciousness

Contemporary AI recognizes that intelligent agents can be composed of multiple functional components, some of which deal with the *processing* of information (reflexive condition-action rules, symbolic reasoning, executive decision making, taking actions, etc.) and some of which deal with the *memory and learning* of information [Russell and Norvig, 2010]. From this AI perspective, it is striking that the diverse list of previously proposed functions of consciousness in the preceding section generally have one thing in common: They largely deal with some facet of the *processing* of information (its integration, manipulation, use for inference or decision making, etc.). In contrast, here we propose an alternative, complementary possibility that the adaptive function of human consciousness is to be found in its contribution to *memory and learning* rather than to the subsequent processing of that information. Specifically, *we hypothesize that the fundamental function of consciousness and its contribution to intelligence will most likely be found in its role in supporting short-term working memory and its associated learning and control mechanisms.*

Psychologists distinguish different memory systems in explaining various neuroscientific and behavioral data [Squire and Dede, 2015]. Human memory at the top level is typically characterized in terms of long-term memory versus short-term memory. Long-term memory is often sub-divided into distinguishable types, such as semantic, episodic, and procedural memory, and we do not consider these further here. Short-term memory is also sub-divided into types, one of which is *working memory* and that serves as our focus here. Working memory stores recently experienced information, typically for a period of seconds to minutes, that is being used in problem solving or other cognitive activities. In contrast to long-term memory with its enormous storage capacity, short-term memory is characterized by a very limited capacity, and is able to retain just a few independent items at any one time [Cowan *et al.*, 2005].

Why focus on working memory as a function of consciousness? One reason is that working memory is widely recognized in philosophy and psychology to involve conscious, reportable cognitive activity [Baars and Franklin, 2003; Baddeley, 2012; Carruthers, 2015; Persuh *et al.*, 2018]. Our view of this relationship is that what psychologists refer to as "working memory" is mostly the same as what some philosophers would characterize as the state of a conscious mind. For information to be consciously accessible and reportable essentially requires that information to be actively represented in working memory. Whether there are also things that are in working memory that are not conscious, or there are things that are conscious but not in working memory, are open questions at present.

Another reason for focusing on working memory is that it is a fundamental underlying element of cognition that provides a unifying perspective for the multiple possible functions of consciousness that have been proposed in the past (listed in Sec. 2). For example, the neurobiological mechanisms that underlie working memory appear to be fairly widespread throughout cerebral cortex [Lara and Wallis, 2015], consistent with the hypothesis that consciousness supports global access to and integration of information. The representation of symbolic information in working memory supports the importance of symbol processing and grounding in human consciousness. The top-down, goal-directed control of working memory that distinguishes it from low-level sensorimotor processes is consistent with past proposals that high-level cognition, executive functions, and attention mechanisms are all key aspects of conscious mind. In other words, what makes working memory "working" is that its contents are actively manipulated by cognitive processes: it is at the intersection of algorithms and data structures. Working memory may turn out to be a common underlying factor in all of these previously proposed functions of consciousness since it is such a foundational aspect of cognition.

4. Working Memory and Computational Correlates of Consciousness

Can computational models of working memory suggest any specific computational correlates of consciousness that might ultimately be used to enhance AI systems? We have recently been examining this issue. Our initial work focused on application-specific models based on standard psychological tests of working memory such as the *n*-back task [Sylvester *et al.*, 2013] and on solving problems involving card matching tasks [Sylvester and Reggia, 2016]. Recently, we greatly generalized our computational models of working memory in the context of developing a *neural virtual machine* (NVM) that is capable of universal computation [Katz *et al.*, 2019].[c]

[c]The NVM is only capable of universal computation in the limit as the number of neurons goes to infinity. While none of our neurocomputational working memory models described here are intended to capture biologically-realistic neural circuitry, they all incorporate separate modules for working memory proper (stores ongoing problem solving information; believed to be widely distributed across human cerebral cortex) and other modules for representing executive-level control of working memory functionality (store action sequences; most closely associated with human prefrontal cortex) [Lara and Wallis, 2015].

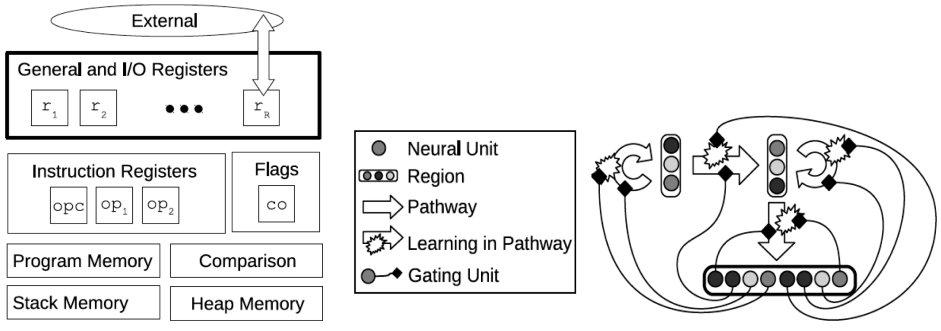

Fig. 1. (a) The virtual machine supported by the NVM. (b) A toy example of an underlying recurrent neural network supported by the NVM that uses gating connections to modulate activation and learning in other pathways. Each block arrow represents many individual connections (not shown) that are being gated.

The NVM is a purely neurocomputational, application-independent software environment that allows one to instantiate cognitive-level algorithms in neural networks. Such algorithms are currently readily implemented via mainstream symbolic AI methods, but much less so via existing programmable neural networks. Importantly, the NVM's modeled knowledge and cognitive processes are acquired through a learning process and represented by distributed patterns of activity over an underlying neural substrate. From a user's perspective, to model a cognitive process using the NVM, one writes an assembly language level program for a virtual machine that is emulated by the NVM (see Fig. 1(a)). However, in actuality, the NVM converts that given program into a region-and-pathway system of recurrently connected neural networks that perform the indicated computations on distributed activity patterns representing symbols, based on the network's dynamics and synaptic weight changes. In short, the NVM can be viewed as a step toward bridging the computational explanatory gap: unlike hybrid systems it is purely neurocomputational. A detailed description of the NVM with a link to an open-source implementation is available [Katz *et al.*, 2019].

How can we use these models of working memory to develop a better understanding of consciousness? Much past work in AC considering computational correlates has started from the premise that some underlying mechanism is a key aspect of consciousness (global processing, attention, self-modeling, etc.) and then explored the implications of that premise via computational modeling. In contrast, our recent work has taken the opposite approach: Start with a model of working memory and ask what core, distinguishing neurocomputational mechanisms were needed to implement that model. The idea is that such distinguishing mechanisms suggest new candidates for computational correlates of consciousness, based on the fact that working memory is tightly associated with the conscious mind. Such correlates might be useful in AI systems. This approach has led us to suggest three new correlates,

all of which are incorporated into the NVM's implementation of working memory, as follows.

The first potential computational correlate of consciousness that we identified is learned *itinerant attractor sequences*, i.e., sequences of learned attractor states of the underlying recurrent neural network's activity, where each sequence element represents a cognitive state of working memory. Each learned attractor state corresponds to an action or "instruction" that is currently active in working memory as a task is being performed. In contrast to previous proposals that individual attractor states or activity trajectories in general may be computational correlates of consciousness, we specifically mean that (i) the trajectory is composed of a sequence of attractors, (ii) this sequence contributes to control of agent behavior and working memory processes, (iii) it involves learned states rather than pre-wired genetically determined circuitry, and (iv) it involves cognitive states used in high-level problem solving and reasoning. Such sequences can represent not only arbitrary procedures, but also arbitrary sequences of items in general such as list data structures. This computational correlate supports past suggestions that the functions of consciousness include symbol processing (each attractor can be viewed as representing a symbol in working memory) and error detection/correction (the transitions between working memory states need only be approximate as the system's dynamics will correct for errors by converging on the nearest attractor state).

The second possible computational correlate of consciousness suggested by our modeling work is the *top-down gating* of working memory by which high-level cognitive processes control what is stored, manipulated and learned by working memory. For example, as illustrated in Fig. 1(b), the underlying neural networks that the NVM uses to implement given algorithms make heavy use of multiplicative gating to turn on/off the flow of information through a network's pathways and to enable/disable learning on network connections. Each action/instruction in a procedural attractor sequence is performed by using multiple coordinated gating operations. We postulate that these gating operations, driven by the sequences of attractor states in the executive component of our working memory models, represent consciously reportable cognitive activities in working memory, and for that reason we take them to be possible computational correlates of consciousness that may contribute to a sense of agency and mental causation. This computational correlate supports past suggestions that the functions of consciousness include top-down executive processes, attention mechanisms, and the evoking/controlling of actions.

Finally, the third computational correlate is *very fast weight changes* that provide for immediate, simultaneous one-step learning and unlearning in working memory. Human short-term working memory is remarkable in its ability to reliably learn new information immediately from just a single presentation of that information. For example, if one is verbally told "add 16 to 17", the numbers involved are immediately retained in working memory as the computation is done — there is generally no need for a person to hear the problem stated several times to learn what the problem is.

Such learning is very different from what is done in many neural network learning systems, including those based on gradient descent methods that require numerous iterative presentations of material to be learned. Our recent modeling work with the NVM implements very fast additions/deletions to working memory contents via synaptic weight changes that involve simultaneously using (1) one-step Hebbian learning to retain new information in working memory, and (2) one-step anti-Hebbian unlearning that actively removes old information that is no longer needed in working memory [Katz *et al.*, 2019]. This fast store-erase Hebbian learning mechanism introduced in the NVM is responsible for its ability to control dynamically what is retained and what is removed from working memory during problem solving. Further, it has proven effective for both representing information about the state of a problem being solved (temporally symmetric weight changes) and information about the behavioral action sequences that control problem-solving (temporally asymmetric weight changes). This third computational correlate suggests, for the first time to our knowledge, that the very fast learning/unlearning of information in working memory may be an important function of consciousness.

5. ACI=AI+AC

We will refer to AI systems that incorporate concepts from AC as *artificial conscious intelligence* (ACI).[d] Put simply, ACI = AI + AC. Having considered above what the function(s) of consciousness might be, we now return to our central question: How might work done in AC enhance the functionality of future AI systems? There are at least two distinct answers to this question about ACI depending on whether one considers simulated or instantiated machine consciousness.[e] By *simulated consciousness*, we mean simulations that attempt to capture some aspect of consciousness or its neural or behavioral correlates in a computational model. Most work in AC falls in this category and thus involves nothing truly mysterious. Just as a computational model of any real-world phenomenon does not imply that the model is actually that phenomenon (e.g., simulating a rain storm does not make a computer wet [Searle, 1980]), simulating aspects of consciousness does not imply that the machine involved actually becomes conscious. In contrast, by *instantiated consciousness* we mean efforts to produce an artificial system that actually is phenomenally conscious, i.e., that experiences qualia and has subjective experiences and thus represents "synthetic phenomenology" [Chrisley, 2009]. Currently no existing work in AC has produced a generally accepted demonstration of instantiated machine consciousness, or even compelling evidence that instantiated machine consciousness is possible. Conversely, there is currently no compelling theoretical or experimental proof that this will not be possible in the future. With this distinction between

[d]This is analogous to distinguishing artificial general intelligence (AGI) systems that study general purpose AI from the more common practice of creating application-specific AI systems.

[e]Our simulated consciousness corresponds to MC1-MC3 and our instantiated consciousness to MC4 in the taxonomy of machine consciousness given in [Gamez, 2018].

simulated and instantiated machine consciousness in hand, we can now return to the question about how work in AC may contribute to improving future AI systems.

A first answer to this question is that work on *simulated* consciousness is directly and immediately relevant to enhancing existing practical AI technology. For example, many existing AI systems are very brittle in the context of novel situations, including both AI systems based on traditional symbol processing methods and those based on contemporary deep learning methods (e.g., adversarial images for deep convolution networks). This is especially a problem with autonomous physical systems where a lack of trustworthiness, both in general but especially in the face of unanticipated novel situations, can be dangerous, and it has significantly limited the practical use of AI in such systems. There is substantial evidence that people, when confronted with novel situations, evoke conscious reasoning and learning to deal with these situations [Mudrik *et al.*, 2012]. This is true regardless of whether the situation is unexpected (e.g., a person driving a familiar highway route suddenly sees two cars collide up ahead) or simply a pre-planned novel experience (e.g., learning to ride a bike or play a game). Current AI systems also generally do not reflect on their internal models to reason about the causes of failure or difficulties. All of this suggests that AC studies relating consciousness to functions such as executive decision making, novelty detection, attention mechanisms, working memory, metacognition, motivations, and informing volitional activities appear promising avenues to explore in creating more effective ACI systems.

Another example of how AC work on simulated consciousness may contribute to practical AI systems relates to the latter's interactions with people. Current human−computer interactions involving AI systems are quite limited. For example, there is no existing AI system that can consistently pass the Turing Test. Conscious self-monitoring would be expected to improve human−robot interactions because of the intimate relationship between self-awareness and the awareness of roles and perspectives [Trafton *et al.*, 2005]. In other words, understanding of roles in various situations is valuable in anticipating the behavior of others, and arguably this understanding relates to self-consciousness. These considerations suggest that simulated AC studies relating consciousness to functions such as working memory with its rapid one-step learning, self-awareness, self-modeling, source of motivations, and symbol grounding would be promising avenues to explore in developing ACI.

A second answer to the question about how work in AC may contribute to creating future AI systems relates to *instantiated* consciousness. While there is no generally accepted proof that instantiated machine consciousness can or cannot be created, we speculate here about what it would mean for AI if a phenomenally conscious machine is someday possible. From a technological perspective, an ACI system based on instantiated consciousness would be anticipated to provide many of the same benefits of robustness, improved human−computer interactions, etc. as would an ACI system involving simulated consciousness. Further, Haikonen [2019] compellingly argued that for an AI system to truly understand the outside world, its symbols must

be grounded in qualia because qualia are self-explanatory forms of sensory information. In addition to these technological implications, perhaps even more significant would be how an instantiated machine consciousness would relate to the scientific study of consciousness. If we can successfully create and confirm an instantiated ACI, something that would effectively be the first artificial mind, we will have made a fundamental advance in consciousness studies in general. Such an ACI would permit the study of consciousness at a much deeper level than is currently possible. For example, it would be expected to shed light not only on the core underlying mechanisms of consciousness, but also on improved criteria for rationally determining the presence/absence of consciousness in machines and animals, and the possibility of mind uploading. It might also lead to major advances in our understanding of psychiatric and neurocognitive disorders, such as schizophrenia, amnesia, and dementia.

6. Discussion

In this paper, we have asked the question of how concepts developed via computational models in AC might contribute to advancing the creation of more effective intelligent agents, or ACI, than can currently be supported by contemporary AI technology. We approached this question by reviewing past hypotheses about what the biological functions of consciousness are, most of which focus on the processing of information. In contrast, we hypothesized that short-term working memory and the associated very fast learning/unlearning of working memory's contents may also be considered as a function of consciousness, and one that complements/unifies previously suggested functions. In this context, we reached three conclusions from our analysis.

First, work on simulated consciousness is immediately relevant to advancing the technology of AI, most prominently in terms of improving robustness and human−computer interactions. Particularly promising avenues for future ACI research can literally be read off from the list of previously hypothesized functions of consciousness in Sec. 2: detecting and managing novelty, symbol grounding and its impact on AI effectiveness, top-down executive control of behavior, the role of a self-model and motivations in machine intelligence, etc.

Second, the development of instantiated machine consciousness (or MC4 [Gamez, 2018]), when combined with AI methodologies, would be a major technological and scientific advance that enables communicating with and studying in depth a conscious mind in ways that are currently not possible. Particularly exciting is the possibility of gaining insights into neurocognitive disorders. The critical direction for future ACI research in this case is how to create an artifact that experiences qualia. At the current time there is no consensus on how this might be done, or even if it is possible.

Third, short-term working memory and especially the rapid learning/unlearning of its contents have been a largely overlooked possible function of consciousness.

For the future, our own research is examining whether or not the three computational correlates of consciousness that have been suggested by AC models of working memory will be sufficient to support compositional working memory *in general*.[f] To examine this issue, we are currently applying the NVM to challenging imitation learning tasks involving cause-effect reasoning where symbolic AI methods, but not neurocomputational methods, have previously been shown to work effectively [Katz *et al.*, 2018].

Acknowledgments

Our work on modeling working memory, one-step learning, and the NVM described in Sec. 4 were supported by ONR award N00014-19-1-2044.

References

Baars, B. [1997] In the theatre of consciousness, *J. Conscious. Stud.* **4**, 292−309.
Baars, B. and Franklin, S. [2003] How conscious experience and working memory interact, *Trends Cogn. Sci.* **7**, 166−172.
Baddeley, A. [2012] Working memory, *Ann. Rev. Psychol.* **63**, 1−29.
Bayne, T. and Montague, M. (eds.) [2011] *Cognitive Phenomenology* (Oxford University Press).
Bringsjord, N. [2007] One billion dollars for a conscious robot, *J. Conscious. Stud.* **14**, 28−43.
Brody, J., Barham, S.Dai, Y, Maxey, C., Perlis, D, Sekora, D. and Shamwel, J. [2016] Reasoning with grounded self-symbols, in *Artificial Intelligence for Human-Robot Interactions* (AAAI), pp. 16−19.
Carruthers, P. [2015] *The Centered Mind* (Oxford University Press).
Chella, A., Frixione, M. and Gaglio, S. [2008] A cognitive architecture for robot self-consciousness, *Artif. Intell. Med.* **44**, 147−154.
Chrisley, R. [2009] Synthetic phenomenology, *Int. J. Mach. Conscious.* **1**, 53−70.
Cleeremans, A. [2005] Computational correlates of consciousness, in S.Laureys (ed.), *Progress in Brain Research*, Vol. 150 (Elsevier), pp. 81−98.
Cowan, N., Elliott, E., Saults, J., Morey, C., Mattox, S., Hismjatullina, A. and Conway, A. [2005] On the capacity of attention, *Cogn. Psychol.* **51**, 42−100.
DeLancey, C. [1996] Emotion and the function of consciousness, *J. Conscious. Stud.* **3**, 492−499.
Devlin, J., Bunel, R., Singh, R., Hausknecht, M. and Kohli, P. [2017] Neural program meta-induction, in *NIPS Proc. Advances in Neural Information Processing Systems*, pp. 2077−2085.
Earl, B. [2014] The biological function of consciousness, *Front. Psychol.* **5**, 1−18.
Gamez, D. [2018] *Human and Machine Consciousness* (Open Book).
Graves, A. *et al.* [2016] Hybrid computing using a neural network with dynamic external memory, *Nature* **538**(7626), 471.
Haikonen, P. [2019] *Consciousness and Robot Sentience* (World Scientific).
Holland, O. [2007] A strongly embodied approach to machine consciousness, *J. Conscious. Stud.* **14**, 97−110.

[f]By "compositional" working memory, we mean that working memory supports computational mechanisms that are sufficiently powerful to dynamically build high-level structured solutions (tree-like data structures, cause-effect networks, etc.) during problem solving, decision making, planning, and learning.

Katz, B. [2013] An embarrassment of theories, *J. Conscious. Stud.* **20**, 43−69.

Katz, G., Davis, G., Gentili, R. and Reggia, J. [2019] A programmable neural virtual machine based on a fast store-erase learning rule, *Neural Netw.* **119**, 10−30.

Katz, G., Huang, D., Hauge, T., Gentili, R. and Reggia, J. [2018] A novel parsimonious cause-effect reasoning algorithm for robot imitation and plan recognition, *IEEE Trans. Cogn. Dev. Syst.* **10**, 177−193.

Kuipers, B. [2008] Drinking from the firehose of experience, *Artif. Intell. Med.* **44**, 155−170.

Lara, A. and Wallis, J. [2015] The role of prefrontal cortex in working memory, *Front. Syst. Neurosci.* **9**, 173.

Manzotti, R. [2012] The computational stance is unfit for consciousness, *Int. J. Mach. Conscious.* **4**, 401−420.

McDermott, D. [2007] Artificial intelligence and consciousness, in M. Moscovitch and E. Thompson (eds.), *Cambridge Handbook of Consciousness* (Cambridge University Press), pp. 117−150.

Metzinger, T. [2000] *Neural Correlates of Consciousness* (MIT Press).

Mudrik, L., Deouell, L. and Lamy, D. [2012] Novelty, not integration: Finding the function of conscious awareness, in S. Kreitler and O. Maimon, (eds.), *Consciousness: Its Nature and Functions* (Nova Science), pp. 265−276.

Pasquali, A., Timmermans, B. and Cleeremans, A. [2010] Know thyself: Meta-cognitive networks and measures of consciousness, *Cognition* **117**, 182−190.

Perlis, D. [1997] Consciousness as self-function, *J. Conscious. Stud.* **4**, 509−525.

Persuh, M., LaRock, E. and Berger, J. [2018] Working memory and consciousness, *Front. Hum. Neurosci.* **12**, 78.

Pierson, L. and Trout, M. [2017] What is consciousness for? *New Ideas Psychol.* **47**, 62−17.

Reggia, J. [2013] The rise of machine consciousness, *Neural Netw.* **44**, 112−131.

Reggia, J., Huang, D. and Katz, G. [2017] Exploring the computational explanatory gap, *Philosophies* **2**, 5, doi: 10.3390/philosophies201005.

Reggia, J., Katz, G. and Davis, G. [2018] Humanoid cognitive robots that learn by imitation, *Front. Robot. AI* **5**, 1.

Reggia, J., Monner, D. and Sylvester, J. [2014] The computational explanatory gap, *J. Conscious. Stud.* **21**, 153−178.

Rosenthal, D. [2008] Consciousness and its function, *Neuropsychologia* **46**, 829−840.

Russell, S. and Norvig, P. [2010] Intelligent agents, in *Artificial Intelligence: A Modern Approach*, Chap. 2 (Prentice Hall), pp. 34−63.

Sanz, R., Hernandez, C. and Sanchez-Escribano, M. [2012] Consciousness, action selection, meaning and phenomenic anticipation, *Int. J. Mach. Conscious.* **4**, 383−393.

Searle, J. [1980] Minds, brains, and programs, *Behav. Brain Sci.* **3**, 417−424.

Seth, A. [2009] The strength of weak artificial consciousness, *Int. J. Mach. Conscious* **1**, 1−82.

Shanon, B. [1998] What is the function of consciousness, *J. Conscious. Stud.* **5**, 295−308.

Squire, L. and Dede, A. [2015] Conscious and unconscious memory systems, *Cold Spring Harbor Perspect. Biol.* **7**, a021667.

Sun, R. and Franklin, S. [2007] Computational models of consciousness, in P. Zelazo and M. Moscovitch (eds.), *Cambridge Handbook of Consciousness* (Cambridge University Press), pp. 151−174.

Sylvester, J. and Reggia, J. [2016] Engineering neural systems for high-level problem solving, *Neural Netw.* **79**, 37−52.

Sylvester, J., Reggia, J., Weems, S. and Bunting, M. [2013] Controlling working memory with learned instructions, *Neural Netw.* **41**, 23−38.

Takeno, J. [2013] *Creation of a Conscious Robot* (Pan Stanford).

Taylor, J. [2007] CODAM: A neural network model of consciousness, *Neural Netw.* **20**, 983−992.

Tononi, G. [2008] Consciousness as integrated information, *Biol. Bull.* **215**, 216−242.

Trafton, J., Cassimatis, N., Bugajska, M., Brock, D., Mintz, F. and Schultz, A. [2005] Enabling effective human-robot interaction using perspective-taking in robots, *IEEE Trans. Syst. Man Cybern. A* **35**, 460−470.

Turing, A. [1950] Computing machinery and intelligence, *Mind* **59**, 433−460.