What are the Computational Correlates of Consciousness?

James A. Reggia^{1,2}, Garrett Katz¹, and Di-Wei Huang¹

¹*Department of Computer Science & ²UMLACS, University of Maryland*
reggia@cs.umd.edu, garrett.katz@gmail.com, dwh@cs.umd.edu

Abstract

Cognitive phenomenology refers to the idea that our subjective experiences include deliberative thought processes and high-level cognition. The recent ascendance of cognitive phenomenology in philosophy has important implications for biologically-inspired cognitive architectures and the role that these models can play in understanding the fundamental nature of consciousness. To the extent that cognitive phenomenology occurs, it provides a new route to a deeper understanding of consciousness via neurocomputational studies of cognition. This route involves identifying computational correlates of consciousness in neurocomputational models of high-level cognitive functions that are associated with subjective mental states. Here we develop this idea and compile a summary of potential neurocomputational correlates of consciousness that have been proposed/recognized during the last several years based on biologically-inspired cognitive architectures. We conclude that the identification and study of computational correlates of consciousness will lead to a better understanding of phenomenal consciousness, a framework for creating a conscious machine, and a better understanding of the mind-brain problem in general.

Keywords: cognitive phenomenology, computational explanatory gap, phenomenal consciousness, machine consciousness, artificial consciousness

1 Introduction

The term *phenomenal consciousness* refers to the subjective qualities (qualia) of sensory phenomena, such as the redness of an object or the pain from a skinned knee, that we experience when awake (Block, 1995). Phenomenal consciousness is very poorly understood at present. Here we describe a framework for biologically-inspired computational investigations that we believe could ultimately shed substantial light on the fundamental nature of phenomenal consciousness, and perhaps even lead to a conscious machine experiencing qualia and subjective experiences. Such a claim may initially seem improbable given current beliefs about the nature of consciousness and the many

controversies that surround the field of consciousness studies. For example, there appears to be fairly widespread agreement that existing computers are *not* phenomenally conscious. If one asks university students whether or not they believe that contemporary electronic computers are conscious, 82% say no, 3% say yes, and 15% are undecided (Reggia, Huang & Katz, 2015). This result is particularly striking in that people in general are usually quite generous in attributing consciousness to others, frequently identifying animals, including ants and worms, and in some cases even plants*, as being conscious (see Figure 1). Some philosophers would go even further, arguing that phenomenal machine consciousness will *never* be possible for a variety of reasons: the non-organic nature of machines (Schlagel, 1999), it would imply panpsychism (Bishop, 2009), the absence of a formal definition of consciousness (Bringsjord, 2007), or the insufficiency of computation to underpin consciousness (Manzotti, 2012; Piper, 2012). More generally, it has been argued that the objective methods of science cannot shed light on phenomenal consciousness due to its subjective nature (McGinn, 2004), making computational investigations irrelevant.

Figure 1: Number of responses at each location when 228 university students were asked to indicate the dividing line between conscious and non-conscious entities on a horizontal scale (Reggia, Huang and Katz, 2015). More than half of the responses indicated that all animal life is conscious. There was no statistically significant difference in how subjects responded based on whether they identified themselves as dualists, idealists, materialists, or otherwise. Such results are consistent with those found in other related studies (Arico et al., 2011; McDermott, 2007). ■

Fortunately, such negative sentiments have not completely deterred the study of artificial consciousness over the last two decades, and this work has led to substantial progress and informative results (Reggia, 2013). For example, neurocomputational models have been studied that support global workspace theories of consciousness (Dehaene et al., 1998), that allow expectation-driven robots to effectively pass the well-known mirror test used to identify self-recognition in animals (Takeno, 2013), and that match behavioral data from human blindsight subjects during post-decision wagering tasks (Pasquali et al., 2010). However, this progress in artificial consciousness has mainly involved computationally simulating neural/behavioral/cognitive aspects of consciousness, much as is done in using computers to simulate other natural processes (e.g., models of weather/climate). There is

*While the notion that plant life is conscious may strike some as surprising, there is growing scientific evidence that plants do ecologically-relevant information processing and communicate with one another via chemical signals, e.g., some plant species can “hear”, discriminate and respond appropriately to sounds that indicate a threat (Appel and Cocroft, 2014). This has contributed to a substantial literature, online discussions, and media coverage about plant consciousness and intelligence that may have influenced such views (Cvrckova et al., 2009; Marder, 2012; Nagel, 1997; Trewavas, 2003).

nothing particularly mysterious about such work: Just as one would not expect that a computer used to simulate a nuclear reactor would become radioactive, one would not expect that a computer used to model some aspect of conscious information processing would become phenomenally conscious. There is no claim that phenomenal consciousness is actually present in a machine in this type of work. While computational modeling has become a widely accepted tool for studying consciousness, there is currently no existing computational approach to artificial consciousness that has yet presented a compelling demonstration or design of phenomenal consciousness in a machine, or even clear evidence that it will eventually be possible.

In the following we describe our framework for computational investigations that we believe will ultimately provide insight into the nature of phenomenal consciousness, and perhaps even lead to a conscious machine experiencing qualia and subjective experiences. We begin with brief background information about *cognitive phenomenology*, which asserts that our subjective experiences are not restricted to just traditional qualia but also encompass deliberative thought processes and high-level cognition (Bayne and Montague, 2011b). We then ask, from a purely computational/engineering viewpoint: What is the main practical barrier to further progress on creating phenomenal machine consciousness? The answer suggested here is that it is a *computational explanatory gap*, our current lack of understanding of how high-level cognitive computations can be captured in low-level neural computations (Reggia, Monner and Sylvester, 2014). The significance of this gap derives from the claims of cognitive phenomenology, in that bridging it may be a critical step in gaining a better understanding of phenomenal consciousness. In other words, we are suggesting that a key implication of cognitive phenomenology is that it provides a new route to a deeper understanding of consciousness via computational studies. This route would focus on identifying possible *computational correlates of consciousness*, computational processes that are exclusively associated with phenomenally conscious information processing (Cleeremans, 2005). Here our focus will be solely on *neurocomputational correlates of high-level cognition* that can be related to the computational explanatory gap. Identifying these correlates depends on neurocomputational implementation of high-level cognitive functions that are associated with subjective mental states. Our main results in this context are to compile in one place a summary of potential computational correlates of consciousness that have been proposed/recognized during the last several years based on biologically-inspired cognitive architectures. We conclude that the identification and study of computational correlates of consciousness will lead to a better understanding of phenomenal consciousness, a framework for creating a conscious machine, and a better understanding of the mind-brain problem in general.

2 Cognitive Phenomenology

Since much of what follows is based on concepts introduced over the last few years concerning cognitive phenomenology, we briefly summarize these concepts here. The central assertion of cognitive phenomenology is that our subjective, first-person experience is not restricted to just traditional qualia involving sensory perception, emotions and visual imagery, but also encompasses deliberative thought and high-level cognition. Such an assertion might seem straightforward to the non-philosopher, but it has proven to be controversial in philosophy. Introductory overviews that explain this controversy in a historical context can be found in (Bayne and Montague, 2011a; Siewert, 2011), and a balanced collection of essays arguing for and against the basic ideas of cognitive

phenomenology is given in (Bayne and Montague, 2011b). A very recent detailed philosophical analysis of cognitive phenomenology is also available (Chudnoff, 2015).

Most of this past discussion in the philosophy literature has focused on reasons for accepting or rejecting cognitive phenomenology. Philosophical theoreticians generally appear to accept that some aspects of cognition are accessible to consciousness, so the arguments in the literature have focused on the specific issue of whether there are phenomenal aspects of cognition that cannot be accounted for by traditional sensory qualia and mental imagery. For example, an advocate of cognitive phenomenology might argue that there are different non-sensory subjective qualities associated with hearing the sentence “I’m tired” spoken in a foreign language if one understands that language than if one does not, given that the sensory experience is effectively the same in both cases.

Here we take a perspective that differs from most of this past discussion that has focused on the validity of cognitive phenomenology. We start with the assumption that cognitive phenomenology is correct and ask what that might imply. In other words, we start by *assuming* that there exist distinct non-sensory subjective mental experiences associated with some aspects of cognition. What would follow from this assumption? Past discussions concerning the implications of cognitive phenomenology have been relatively limited, although it has been speculated that the occurrence of cognitive subjective states would inform our understanding of the trustworthiness of introspection, of the functional role of consciousness, and of the relationship between the phenomenal mind and the intentional mind (Bayne and Montague, 2011a). Here we consider another somewhat more radical hypothesis about such implications: *Cognitive phenomenology provides a route to both a deeper understanding of human consciousness and also to a phenomenally conscious machine.* This hypothesis is plausible because cognitive phenomenology makes a large body of research in cognitive science, work that is developing neurocomputational implementations of high-level cognition, relevant to the issue of phenomenal consciousness. This work, while recognized as being relevant to intentional aspects of consciousness, has not generally been considered of direct relevance to phenomenology. This implication is an aspect of cognitive phenomenology that has apparently not been recognized previously.

3 The Computational Explanatory Gap

What are the main practical barriers at present to creating phenomenal machine consciousness? There is no shortage of well-known candidates. These include the absence of a generally agreed-upon definition of consciousness, our limited understanding of its neurobiological correlates, and the “other minds problem” applied to artifacts (how could one possibly know whether or not a machine is conscious?). While each of these difficulties is substantial, from a computer science/engineering viewpoint none appear to be sufficient to account for the lack of progress that has occurred. Significant progress has been made in artificial intelligence and artificial life without having generally accepted definitions of either intelligence or life. Our current inadequate understanding of the neurobiological basis of consciousness does not prevent one from conducting non-biological experimentation via computer modeling of existing theories, and the other minds problem has not prevented studies of phenomenal consciousness and self-awareness in humans and animals.

Our central point here is that there is another less recognized barrier, the *computational explanatory gap*, which is also critically important. We define the computational explanatory gap to be

a lack of understanding of how high-level cognitive information processing can be mapped onto low-level neural computations (Reggia, Monner and Sylvester, 2014). By “high-level cognitive information processing” we mean aspects of cognition such as goal-directed problem solving, executive decision making, cognitive control, planning, language understanding, and metacognition – cognitive processes that are widely accepted to at least in part be consciously accessible. By “low-level neural computations” we mean the kinds of computations that can be achieved by networks of artificial neurons like those that are widely studied in contemporary computer science, engineering, psychology, and neuroscience.

The computational explanatory gap can be contrasted with the well-known *philosophical explanatory gap* between a successful functional/computational account of consciousness and accounting for the subjective experiences that accompany it (Levine, 1983). While the computational explanatory gap is ultimately relevant to the philosophical explanatory gap, the former is *not* a mind-brain issue per se. Rather, it is a gap in our general understanding of how computations (algorithms and dynamical states) supporting goal-directed reasoning and problem solving at a high level of cognitive information processing can be mapped into the kinds of computations/algorithms/states that can be supported at the low level of neural networks. In other words, as illustrated in Figure 2, *it is a purely computational issue*.

Figure 2: The well-known philosophical explanatory gap (on the left) and less recognized computational explanatory gap (on the right). The argument presented here is that the latter may ultimately prove to be the more fundamental problem, and that focusing on solving it may be the key to advancing future work on phenomenal machine consciousness. ■

The computational explanatory gap is also not an issue specific to computers or even computer science. It is a generic issue concerning how one type of computations at a high level (serial goal-directed deliberative reasoning algorithms associated with conscious aspects of cognition) can be mapped into a fundamentally different type of computations at a low level (parallel distributed neural computations and representational states). It is an *abstraction* independent of the hardware involved, be it the electronic circuits of a computer or the neural circuits of the brain.

Comparing the computational explanatory gap and the philosophical explanatory gap further is informative. In philosophy, the computational explanatory gap relates to the long-standing discussion concerning the “easy problem” of accounting for cognitive information processing versus the “hard problem” of explaining subjective experience, the latter being associated with the philosophical explanatory gap (Chalmers, 1996). This characterization of the easy problem is often qualified by comments that the easy problem is not really viewed as being truly easy to solve, but that it is easy to imagine that a solution can be obtained via functional/computational approaches, while this is not the case for the hard problem (e.g., Block, 1995). Nonetheless, the easy/hard contemporary philosophical distinction tends to largely dismiss solving the computational explanatory gap, as it is just part of the easy problem. Such a dismissal fails to explain why the computational explanatory gap has proven to be largely intractable during more than half a century of efforts (since McCulloch and Pitts (1943) first captured propositional-level logical inferences using neural networks). This intractability is somewhat

mysterious to us, given that the brain somehow readily bridges this gap. While the brain's structure is of course quite complex, it appears unlikely that this complexity alone could be the whole explanation, given that mainstream top-down AI has been qualitatively more successful in modeling high-level cognition when compared to neurocomputational methods. We believe that something more fundamental is being missed here. More specifically, it may turn out that the unfortunate terminology of the “easy” and “hard” problems that dominates contemporary philosophical thought will ultimately prove to be precisely backwards. In other words, from our perspective the computational explanatory gap is actually the more fundamental and difficult-to-resolve issue, and that once this gap is bridged, the philosophical explanatory gap may be found to be much more tractable and fade away.

The computational explanatory gap is clarified by the efforts of cognitive psychologists to characterize the properties that distinguish human conscious versus unconscious cognitive information processing. Human conscious information processing has been characterized as being serial, relatively slow, and largely restricted to one task at a time (Baars, 1988, 2002; Dehaene and Naccache, 2001). In contrast, unconscious information processing is parallel, relatively fast, and can readily involve more than one task simultaneously with limited interference between the tasks. Conscious information processing appears to involve widespread “global” brain activity and is internally consistent, while unconscious information processing appears to involve more localized activation of brain regions and varies in its internal consistency. Conscious information processing is often associated with inner speech and is operationally taken to be cognition that is reportable,[†] while unconscious information processing has neither of these properties. The key point here is that psychologists explicitly trying to characterize the differences between conscious and unconscious information processing have implicitly, and perhaps unintentionally, identified the computational explanatory gap. The properties they have identified as characterizing unconscious information processing – parallel processing, efficient, modular, non-reportable – often match reasonably well with those of neural computation. For example, being “non-reportable” matches up well with the nature of neurocomputational models where, even once a neural network has learned to perform a task very successfully, what that network has learned remains largely opaque to outside observers and often requires a major effort to determine or express symbolically (Huynh and Reggia, 2012). In contrast, the properties associated with conscious information processing – serial processing, relatively slow, holistic, reportable – are a much better match to symbolic top-down algorithms, and a poor match to the characteristics of neural computations. In this context, the unexplained gap between conscious cognition and the underlying neural computations that support it is strikingly evident.

The computational explanatory gap also relates to substantial efforts in neuroscience that have focused on identifying *neural correlates of consciousness* (Crick and Koch, 1990). Roughly, a neural correlate of consciousness is some minimal neurobiological state whose presence is sufficient for the occurrence of a corresponding state of consciousness (Chalmers, 2000). Neurobiologists have identified several candidate neurobiological correlates, such as specific activity patterns in the brain's electrical activity (e.g., synchronized cortical oscillations), activation of specific neural structures (e.g., thalamic intralaminar nuclei), widespread brain activation, etc. (Metzinger, 2000b). However, in spite of an enormous neuroscientific endeavor in general over more than a century, there remains a large

[†] Although such a criterion has substantial limitations, being verbally reportable has long been widely used in experimental psychology as a major objective criterion for accepting that a person is conscious of (subjectively aware of) an event (Baars, 1988; Dehaene and Naccache, 2001).

difference between our understanding of unconscious versus conscious information processing in the brain, particularly in our limited understanding of the neural basis of the high-level cognitive functions that are relevant to cognitive phenomenology. For example, a great deal is currently known at the macroscopic level about associating high-level cognitive functions with brain regions (pre-frontal cortex “executive” regions, language cortex areas, etc.), and a lot is known about the microscopic functionality of neural circuitry in these regions, all the way down to the molecular and genetic levels. What remains largely unclear is how to put those two types of information together, or in other words, how the brain maps high-level cognitive functions into computations over the low-level neural circuits that are present, i.e., the computational explanatory gap. It remains unclear why this mapping from cognitive computations to neural computations is so opaque today given the enormous resources that have been poured into understanding these issues. The key point here is that this large gap in our neuroscientific knowledge about how to relate the macroscopic and microscopic levels of information processing in the brain to one another is, at least in part, also a manifestation of the underlying computational explanatory gap.

Why is bridging the computational explanatory gap of critical importance in developing a deeper understanding of phenomenal consciousness, and in addressing the possibility of phenomenal machine consciousness? The reason is that bridging this gap would allow us to do something that is currently beyond our reach: It would allow us to directly and cleanly compare (i) neurocomputational mechanisms associated with conscious/reportable high-level cognitive activities, and (ii) neurocomputational mechanisms associated with lower-level unconscious information processing. In effect, this comparison would allow us to determine whether or not there are computational correlates of consciousness in the same sense that there are neurobiological correlates of consciousness.

A *computational correlate of consciousness* can, in general, be defined to be any principle of information processing that characterizes the differences between conscious and unconscious information processing (Cleeremans, 2005). Such a definition is quite broad: It could include, for example, symbolic AI logical reasoning and natural language processing algorithms. Consistent with this possibility, there have been past suggestions that intelligent information processing can be divided into symbolic processes that represent conscious aspects of the mind, and neurocomputational processes that represent unconscious aspects of cognition (Chella, 2007; Kitamura et al., 2000; Sun, 1999, 2000). However, while such models/theories implicitly recognize the computational explanatory gap as we explained it above, they do not attempt to provide a solution to it. The critical issue that we are concerned with here is how to *replace* the symbolic modules of such models with neurocomputational implementations. How such replacement is to be done remains stubbornly mysterious today.

Accordingly, in the following we use the term “computational correlate of consciousness” to mean minimal neurocomputational processing mechanisms that are specifically associated with conscious aspects of cognition but *not* with unconscious aspects. In other words, in the context of the computational explanatory gap we are specifically interested in *neurocomputational* correlates of consciousness, i.e., computational correlates related to the representation, storage, processing, and modification of information that occurs in neural networks. Computational correlates of consciousness are a priori distinct from neural correlates (Cleeremans, 2005). As noted above, proposed neural correlates have in practice included, for example, electrical/metabolic activity patterns, neuroanatomical regions of the brain, and biochemical phenomena (Chalmers, 2000) - correlates within the realm of biology that are not computational. Computational correlates are abstractions that

of may well find implementation in the brain, but as abstractions they are intended to be independent of the physical substrate that implements them (brain, silicon, etc.).

Our suggestion here is that, should computational correlates of consciousness as defined above be discovered, they will provide a direct route to investigating the possibility of phenomenal machine consciousness, to identifying candidate properties that could serve as objective criteria for the presence of phenomenal consciousness in machines, animals and people, and to a better understanding of the fundamental nature of consciousness. In other words, what we are suggesting is that a complete characterization of high-level cognition in neurocomputational terms may, along the way, show us how subjective experience arises mechanistically, based on the concepts of cognitive phenomenology discussed above. Cognitive phenomenology emphasizes that our subjective first-person experience encompasses deliberative thought and high-level cognition, precisely the subject at issue in the computational explanatory gap as we have defined it. While the cognitive phenomenology viewpoint is controversial (Bayne and Montague, 2011a), to the extent that this viewpoint holds it indicates that bridging the computational explanatory gap will provide evidence directly relevant to a deeper understanding of phenomenal consciousness. Even if bridging the computational explanatory gap does not fully account for subjective experience, it will at least greatly demystify it.

4 Identifying Computational Correlates of Consciousness

From a computational point of view, the recognition of cognitive phenomenology that high-level cognitive processes may have distinct subjective qualities associated with them has striking implications for computational research related to consciousness. Cognitive phenomenology implies that ongoing research focused on developing biologically-inspired computational models of higher cognitive functions is directly relevant to understanding the nature of consciousness. Much of this work on neurocognitive architectures attempts to map higher cognitive functions into neurocomputational mechanisms independently of any explicit relationship between these functions and consciousness. Such work may, intentionally or inadvertently, identify neurocomputational mechanisms associated with phenomenally conscious information processing that are not observed with unconscious information processing. This would potentially provide examples of computational correlates of consciousness, and ideally would provide a deeper understanding of the nature of consciousness in general, just as molecular biology contributed to demystifying the nature of life and deprecating the theory of vitalism.

If one allows the possibility that the computational explanatory gap is a substantial barrier both to a deeper understanding of phenomenal consciousness and to developing machine consciousness, then the immediate research program becomes determining how one can bridge this gap and identify computational correlates of consciousness. Encouragingly, cognitive phenomenology suggests that the substantial efforts by researchers working in artificial consciousness and by those developing neurocognitive architectures involving high-level cognitive functionality are beginning to identify computational correlates. Much of this past work can be interpreted as being founded upon specific candidates for computational correlates of consciousness, and this observation suggests that such models are directly related to addressing the computational explanatory gap. In the following, we compile a summary of representative examples of such computational correlates that are identifiable from the existing literature, dividing them into the five classes indicated in Table 1. All of the examples we include are based on neurocomputational models.

Table 1: Five Classes of Potential Computational Correlates of Consciousness

<u>Type</u>	<u>Basis</u>	<u>Examples</u>
representation properties	how information is encoded	attractor stability/trajectories
system properties	network properties as a whole	global activity, integrated info
self-modeling	internal representation of self	body image, virtual machines
relational properties	cognitive entity relations	symbol grounding, HOT theories
cognitive control	managing cognitive processes	directing attention, neural gating

4.1 Representation Properties

One class of computational correlates of consciousness is based on intrinsic properties of the neural representation of information used in cognitive states/processes. For example, a number of theoretical arguments and several early recurrent neural network models of consciousness have been based on hypothesizing that attractor states, certain specific neural activity patterns (“bubbles”), the stability/strength/distinctiveness of activity patterns, trajectories through activation state space, activity resonance, or even the explicit encoding of information in general, are associated with or characterize consciousness (Aleksander, 1996; Cleeremans, 2005; Fekete and Edelman, 2011; Grossberg, 1999; Mathis and Mozer, 1995; O’Brien and Opie, 1999; Taylor, 1997).

The difficulty with identifying computational correlates of consciousness based on representation properties is that there are many physical systems that are generally not viewed as conscious that have properties such as attractor states or activity patterns, leading to past suggestions that just representing consciousness in such a fashion is not helpful: It may relate to the contents of consciousness, but it is arguably not useful because it involves no “internal experience” (Taylor, 2003) or because it implies panpsychism (Cleeremans, 2005).

4.2 System Properties

Another class of computational correlates of consciousness is based on system-level properties of a neurocomputational system. The term “system-level properties” is used here to mean those computational properties that characterize a neurocomputational system as a whole rather than, as with representational properties, depend on the specific nature of neural activity patterns that represent/encode cognitive states and processes. Two examples are given here: global information processing and information integration.

The terms local and global, when applied to neurocognitive architectures, generally refer to the spatial extent of information processing. A number of empirical studies have suggested that information processing during conscious mental activities is widely distributed (“global”) across cerebral cortex regions and also associated with increased inter-regional communication (Baars et al., 2003; Massimini et al., 2005; Tagliazucchi et al., 2016). Inspired by this, a number of neural network models have been studied in which conscious/effortful information processing is associated with computational processes involving the recruitment of hidden layers and hence widespread network activation (Connor and Shanahan, 2010; Dehaene et al., 1998; Newman et al., 1997; Raffone and

Pantani, 2010; Shanahan, 2008). An example neural architecture of this type is illustrated schematically in Figure 3. This specific example concerns a Stroop task where one must either read aloud each seen word or name the color of the ink used in the word, the latter being more difficult and effortful with incongruous words. Significant widespread activation of the “global workspace” occurs when the model is dealing with a difficult naming task that requires conscious effort by humans, and not with the easier, more automatic word reading task. It has been suggested that similar global information processing provides a plausible neurocomputational basis for phenomenal and access consciousness (Raffone and Pantani, 2010).

Figure 3: A schematic representation of the architecture of a neural global workspace model for a word-color Stroop task (Dehaene et al., 1998). Each rounded box represents a set/layer of neurons. Local processing occurs in the input and output layers, while with global processing, substantial activation also occurs throughout the workspace layer. ■

The hypothesis that global information processing is a computational correlate of conscious mind is consistent with substantial scientific evidence that widespread cortical activation correlates with conscious brain states, and it provides a partial intuitive account for the unity of consciousness and a clear statement about what distinguishes conscious from unconscious brain activity. However, neural global workspace models of consciousness have not yet shed significant light on *why* global information processing should lead to phenomenal consciousness. Further, there are other multi-layer neural networks with architectures similar to global workspace models that include low-level information processing that is generally not viewed as conscious (van der Velde & de Kamp, 2006).

A second system property that has been postulated to be associated with phenomenal consciousness is the ability to integrate information. Information integration theory hypothesizes that consciousness is equivalent to the capacity of a system to integrate information into a unified experience and to differentiate among a large collection of different conscious states (Tononi, 2004, 2008). The theory provides a quantitative measure ϕ of a neural network’s ability to integrate information, and the larger that ϕ is, the more phenomenally conscious a network is. For example, estimated ϕ values associated with different modules in a neural control system have suggested that modules that gate other modules’ activity have the highest ϕ values and thus are more closely associated with phenomenal consciousness (Gamez, 2010). Information integration theory claims that a system’s subjective experience is equivalent to its capacity to integrate information, that the

“informational relations” of a system’s components define a space of qualia, and that a limited form of panpsychism holds (Tononi, 2004, 2008).

Information integration theory is attractive in providing a well-defined quantitative measure of the extent to which a system, natural or artificial, is conscious. However, at the present time it remains an open question as to whether the measure ϕ has anything to do with the subjective experience that we associate with conscious mind rather than with the information processing properties of a system, conscious or not (Manzotti, 2012). The ϕ score seems more directly related to a system’s potential to exhibit intelligent information processing rather than subjective experience. For example, it has been shown that, with a suitable selection of synaptic weights, one can construct simple, fully connected neural networks having arbitrarily high ϕ values (Seth et al., 2006). According to integrated information theory, at some (unspecified) ϕ value such networks would therefore exhibit phenomenal consciousness, a conclusion that is difficult for some to accept, and which does not appear to clarify the nature of consciousness (Seth, 2009). This theory arguably does not have some basic properties needed for well-formedness (Ceruleo, 2011). On the practical side, a major limitation of information integration theory is that one cannot compute ϕ directly for neural networks of appreciable size, and it is extremely difficult to interpret what any given value of ϕ actually means (i.e., ϕ is just a relative measure, not an absolute measure).

4.3 Self-Modeling

A third class of computational correlates of consciousness involves *self-modeling*. Self-awareness has long been viewed as a central aspect of a conscious mind (Dawkins, 1976; Samsonovich and Nadel, 2005; Searle, 2004). For example, analysis of the first-person perspective suggests that our mind includes a *self-model* (virtual model, body image) that is supported by neural activation patterns in the brain. The argument is that the subjective, conscious experience of a “self” arises because this model is always there due to the proprioceptive, somatic and other sensory input that the brain constantly receives, and because our cognitive processes do not recognize that the body image is a virtual representation of the physical body (Metzinger, 2000a). Accordingly, AI and cognitive science researchers have made a number of proposals for how subjective conscious experience could emerge in embodied agents from such self-models. One suggested possibility is that self-models could lead to “strong self-reference” whereby an artificial agent that reasons about itself can reason about its own reasoning about itself, leading to a “strongly self-referring loop” (Perlis, 1997). Another possibility is that phenomenal consciousness could emerge from introspective reasoning mechanisms about perception (McDermott, 2007). It has alternatively been suggested that introducing a self-concept during learning is the key event in creating “computational consciousness” in suitable cognitive architectures (Samsonovich, 2007).

One specific type of internal model that is claimed to capture various features of consciousness is the *virtual machine architecture* (Sloman&Chrisley, 2003). This approach views a conscious human mind as a virtual machine (software) that is being executed by the brain (hardware). The “mental” states of the virtual machine are generally but not necessarily causally connected to one another, consistent with a functional view of consciousness. Self-modeling occurs in a suitable virtual machine when the machine develops concepts for categorizing and labeling its own states as sensed by internal monitors. In this context, qualia are defined to be what virtual machines “refer to when referring to the objects of internal self observation” (Sloman&Chrisley, 2003). In other words, qualia are viewed as

being a side-effect of having an executive or meta-management component as part of the virtual machine that allows it to examine its own internal functioning and representations.

Another example of this type of work has studied “conscious robots” that are controlled by recurrent neural networks (Takeno, 2013). A central focus of this work is handling temporal sequences of stimuli, predicting the event that will occur at the next time step from past and current events. The predicted next state and the actual next state of the world are compared at each time step, and when these two states are the same a robot is declared to be conscious (“consistency of cognition and behavior generates consciousness”). In support of this concept of consciousness, these expectation-driven robots have been shown to be capable of self-recognition in a mirror (Takeno, 2008), effectively passing the “mirror test” of self-awareness used with animals (Gallup, 1970) and allowing the investigators to claim that it represents self-recognition (Takeno, 2008). The predictive value of a self-model like that used here has previously been argued to be associated with consciousness (Ascoli, 2005).

While the concept that an internal self-model is a crucial aspect of consciousness is quite appealing, a significant limitation of past self-modeling work is that most of the neurocomputational self-models actually developed to date are quite simple and deal primarily with low level sensorimotor processing. Predictions have focused on what sensory input to anticipate next, or the outcome to be expected from taking a certain action, and thus they do not capture the richness of human meta-knowledge about one’s higher cognitive processes.

4.4 Relational Properties

A fourth class of potential computational correlates of consciousness is based on relationships between entities. Symbol grounding models and higher-order thought theories are examples that fall within this category.

The term *symbol grounding* is often used to indicate that the symbols (such as words) used in information processing systems must be grounded (associated with entities in the external world) in order for those symbols to have meaning. It has been suggested that symbol grounding is a computational correlate of consciousness. For example, one claim is that any system that maintains a correspondence between high-level, symbolically represented concepts and low-level data stream entities, and that has a reasoning system which makes use of these grounded symbols, has true subjective experiences corresponding to qualia and a sense of self-awareness (Kuipers, 2005). A physical machine that satisfies these requirements would therefore be phenomenally conscious (Kuipers, 2008). However, not all would agree with this viewpoint: It has recently been argued that the dimensionality reduction implied by symbol grounding is just the opposite of what is needed to create phenomenal machine consciousness (Chella and Gaglio, 2012).

The hypothesis that symbol grounding is a computational correlate of consciousness makes neurocomputational systems that learn to ground auditory symbols by associating them with visual images potentially relevant to the issue of consciousness (Weems and Reggia, 2006). Such models, while quite limited at present, can learn to produce the meaning of sentences describing a very simple external world that is observed, including for novel objects and novel descriptive sentences (Monner and Reggia, 2011). They sometimes learn/discover a *latent symbol system* – a system for symbol processing that was not built into the structure of the network but instead emerged as part of its learned distributed representation (Monner and Reggia, 2012). While the network’s internal symbols

are latent from our perspective as outside observers, they are very real in the sense that they are accessed, manipulated and inspected by the neural architecture itself when interpreting sentences, including novel ones.

A second potential computational correlate of consciousness involving relational properties is based on *higher-order thought theories* (HOT theories) of consciousness (Rosenthal, 1996). HOT theory is relational: it postulates that a mental state S is conscious to the extent that there is some other mental state (a “higher-order thought”) that directly refers to S or is about S . In other words, consciousness occurs when there is a specific relationship like this between two mental states. It has been suggested that artificial systems with higher-order representations consistent with HOT theory could be viewed as conscious, even to the extent of having qualia, if they use grounded symbols (Rolls, 1997, 2007). Further, HOT theory has served as a framework for developing *metacognitive networks* in which some components of a neurocomputational architecture, referred to as higher-order networks, are able to represent, access information about other lower-order networks (Cleeremans et al., 2007), as illustrated in Figure 4. It is the second-order network’s learned reference to the first-order network’s internal representations that makes the latter functionally conscious, according to HOT theory. The claim is that “conscious experience occurs if and only if an information processing system has learned about its own representation of the world” (Cleeremans et al., 2007), and that the higher-order representations formed in metacognitive models capture the core mechanisms of HOT theories of consciousness (Pasquali et al., 2010).

Figure 4: A caricature of a typical metacognitive neural architecture. Boxes indicate network layers, and gray arrows indicate all-to-all adaptive connections between layers. The first-order network on the left is used to learn pattern classification tasks. The second-order network on the right monitors the first-order network’s hidden layer representation and outputs a high or low estimate as to the correctness of the first-order network’s output for each input pattern. The correctness of this output over time reflects the extent to which the second-order network has learned a meta-representation of the first-order network’s representation. ■

Models of conscious information processing based on HOT theory face some important challenges. For example, contemporary computers are generally believed by many people not to be conscious, and yet modern computers have monitoring systems that keep track of their internal states, indicating that they are conscious according to HOT theory (Rey, 1983). A recent review of higher-order representation theories of consciousness has discussed several objections to such theories (Lycan, 2009).

4.5 Cognitive Control

A fifth class of computational correlates of consciousness is based on *cognitive control*, the process of managing other cognitive processes (Schneider and Chein, 2003). Examples of cognitive control include directing attention, maintaining working memory, goal setting, and inhibiting irrelevant signals. The limited ability of current neurocomputational methods to support higher-level cognitive control is somewhat surprising in that the human brain handles such issues routinely, establishing that neural computations clearly have the capacity to do so. Here we consider two examples of how cognitive control models can suggest computational correlates of consciousness: via attention control and network gating.

At each moment in time, a person is only consciously aware of a fraction of the ongoing stream of visual, auditory, touch and other sensory information being received. The attention control mechanisms that actively select what we attend to have the ability to focus on behaviorally important aspects of sensory information, selecting, for example, which specific seen object to attend to based on its location in space, what type of object it is, and/or its general properties. While attention and conscious awareness are not equivalent processes, normally they are closely linked and highly correlated (Graziano and Webb, 2014; Koch and Tsuchiya, 2006; Lamme, 2003; Treisman, 2009). It is therefore not surprising that some computational studies have viewed attention mechanisms as computational correlates of consciousness.

For example, Haikonen (2007) has proposed a conscious machine based on a number of properties, including inner speech and the machine's ability to report its inner states. The machine is composed of a collection of inter-communicating modules that communicate with one another. The distinction between a topic being conscious or not is based on whether the machine as a whole attends to that topic. At each moment of time, each specialized module of the machine broadcasts information about the topic it is addressing to the other modules. Each broadcasting module competes to recruit other modules to attend to the same topic that it is processing. The subject of this unified collective focus of attention is claimed to enter consciousness: "The machine becomes 'conscious' about a topic when the various modules cooperate in this way in unison and focus their attention on the percepts about this topic." It is the collective, cooperative attending to a single topic that represents a computational correlate of consciousness. Recently implemented in a robot (Haikonen, 2012), this design suggests that a machine can only be conscious of one topic at a time, and like global workspace theory, this model is consistent with empirical evidence that conscious brain states are associated with widespread communication between cerebral cortex regions.

Researchers developing other attention-based models have suggested that the efference copy (or corollary discharge) associated with top-down attention control mechanisms can be viewed as a computational correlate of consciousness. An efference signal or corollary discharge is a copy of the output of a control system that is fed back to input regions of the control system. In some theories these efference copies in the brain that are sent to posterior sensory cortex are hypothesized to become conscious, and this has formed the basis of neurocomputational models (Cotterill, 2003). Similarly, so-called Corollary Discharge of Attention Movement computational models simulate control of changes in the focus of attention (Taylor, 2003, 2007). In these latter models an executive system sends top-down control signals (corollary discharges) that selectively modulate posterior sensory cortex input, and it is postulated that consciousness is specifically due to the corollary discharge signal, making the latter a computational correlate of consciousness (Taylor, 2007). The central claim is that this signal is

directly responsible for the conscious experience of ownership of actions that control one's attention to topics, providing a sense of agency.

Another type of computational correlate of consciousness that relates to cognitive control involves *network gating*. In neural networks that are gating architectures, the modules not only exchange information via neural pathways, but they also gate the actions of one another. A module may turn on/off the activity or connections of other modules, and may determine when they retain/discard their activity state, when they learn/unlearn information, and when they generate output. Unlike with HOT theory, which is concerned with networks that monitor one another, gating architectures are concerned with modules that control one another's actions.

Network gating architectures often focus on general cognitive control, an umbrella term for goal-directed executive cognitive systems that manage other cognitive processes. As noted above, developing neural architectures capable of modeling cognitive control processes in a general sense has proven to be surprisingly challenging, in part because they struggle to represent the goals and rules of various cognitive tasks, and in part because they consist of a set of neural modules that learn to perform tasks involving temporal sequences. A recent example of this type of neurocognitive architecture, illustrated in Figure 5, is composed of attractor neural networks that gate the actions of other modules and themselves (Sylvester et al., 2013). A control module learns to turn on/off the actions of other modules, thereby determining when input patterns are saved or discarded, when to learn/unlearn information in working memory, when to generate outputs, and even when to update its own states. Arguably these gating operations, driven by itinerant attractor states of the model's modules, correspond to consciously reportable cognitive activities, and thus they are potential computational correlates of consciousness. This specific architecture illustrated in Figure 5 produced accuracy and timing results reminiscent of those seen in humans performing similar tasks, and made testable predictions (Sylvester et al, 2013). Subsequent related work has used long short-term memory networks with associative memory to learn simple procedures (Graves et al., 2014), and attractor networks with gating for problem-solving tasks involving "game playing" (Sylvester and Reggia, 2016). Gating networks like these may relate to the concept of top-down causation discussed in the free will literature.

continues on next page

Figure 5: A gating network model for the n-Back task (a widely used test in psychology). Thin, solid arrows are one-to-one connections. A working memory module is recurrently connected (broad arrow). Dashed lines are the gating outputs of a control module (on the right) that control the flow of activity via pathways in the architecture, determine when output occurs, and decide when and where learning occurs. ■

5 Discussion

Existing computational models have successfully captured a number of neurobiological, cognitive and behavioral aspects of conscious information processing as machine simulations. Put simply, it has been possible to develop limited forms of what we have called simulated consciousness. This achievement is extremely important: These computational studies are providing a way to test whether theories about key neural, cognitive and/or behavioral features of consciousness, when implemented as computer models, can produce results in agreement with experimental data. Computational modeling also represents important progress towards producing machines that can exhibit external behaviors that are associated with human consciousness, and thus may benefit future artificial agents by providing increased functionality and that lets them interact with people in more natural ways (Aleksander, 2013; Charkaoui, 2005; McCauley, 2007; Sanz et al., 2012). Put simply, work on simulating consciousness has become a recognized and increasingly accepted methodology for the scientific study of consciousness, especially within the framework of functionalism.

In contrast, at the present time our understanding of the phenomenology of consciousness remains incredibly limited. This is true both for human consciousness and for considerations about the possibility of machine consciousness. Our central argument here is that this apparent lack of progress towards a deeper understanding of phenomenal consciousness is largely due to the computational explanatory gap: our current lack of understanding of how higher-level cognitive algorithms can be mapped onto neurocomputational algorithms/states. While those versed in mind-brain philosophy may

generally be inclined to dismiss this gap as just part of the “easy problem”, such a view is at best misleading. This gap has proven surprisingly intractable to over half a century of research on neurocomputational methods, and existing philosophical works have provided no deep insight into *why* such an “easy problem” has proven to be so intractable. The view offered here is that the computational explanatory gap is a fundamental issue that needs a much larger collective effort to clarify and resolve.

It is in this context that the concepts of cognitive phenomenology are so significant to work on biologically-inspired cognitive architectures. These concepts imply that substantial past and ongoing computational work outside of the artificial consciousness literature is of potential relevance to consciousness studies. As illustrated above, a great deal of past work on neurocognitive architectures has attempted to map higher cognitive functions into neurocomputational processes, thereby directly relating to the computational explanatory gap. This includes work that does not explicitly consider its relevance to consciousness studies.

The examples presented above, and results from many other related studies, is very encouraging in suggesting that there now exist several potential computational correlates of consciousness. However, an important difficulty remains: Many hypothesized computational mechanisms identified so far as correlates of consciousness are either known to also occur in neurocomputational systems supporting non-conscious activities, or their occurrence has not yet been definitively excluded in such settings. In other words, the main difficulty in identifying computational correlates of consciousness at present is in establishing that proposed correlates are *not* also present with unconscious aspects of cognitive information processing. For example, the metacognitive neural networks that HOT theory has inspired suggest that information processing in such neural architectures can be taken to be a computational correlate, but there are analogous types of information processing in biological neural circuits at the level of the human brainstem and spinal cord that are associated with apparently unconscious processing. Thus, “second-order” information processing in neural architectures *alone* would not fully satisfy our criteria for being a computational correlate of consciousness. Similarly, global/widespread neural activity per se does not appear to fully satisfy our criteria because such processing also occurs in apparently unconscious neural systems (e.g., interacting central pattern generators in isolated lamprey eel spinal cord preparations produce coordinated movements involving widespread distributed neural interactions (Ijspeert, 2008), but it is improbable that an isolated eel spinal cord should be viewed as being conscious). Very similar points can be made about gating mechanisms, integrated information processing, and self-modeling.

For this reason, it remains unclear at present whether any of the currently hypothesized candidates for neurocomputational correlates of consciousness will ultimately prove to be satisfactory without further qualifications/refinements. This is not because they fail to capture important aspects of conscious information processing, but primarily because similar computational mechanisms have not yet been proven to be absent in unconscious information processing situations. These computational mechanisms are thus not yet *specifically* identifiable only with conscious cognitive activities. Future work is needed to resolve such issues, and to further consider other types of potential computational correlates (quantum computing, massively parallel symbol processing, etc.).

Acknowledgement: This work was supported by ONR award N000141310597.

References

- Aleksander, I. (1996) *Impossible Minds*, Imperial College Press.
- Aleksander, I. (2013) Phenomenal Consciousness and Biologically Inspired Systems, *International Journal of Machine Consciousness*, 5, 3 – 9.
- Appel, H. and Cocroft, R. (2014) Plants Respond to Leaf Vibrations Caused by Insect Herbivore Chewing, *Oecologia*, 175, 1257-1266.
- Arico, A, Fiala, B, Goldberg, R, Nichols, S. (2011) The Folk Psychology of Consciousness, *Mind and Language*, 26, 327-352.
- Ascoli, G. (2005) Brain and Mind at the Crossroads of Time, *Cortex*, 41, 619-620.
- Baars, B. (1988) *A Cognitive Theory of Consciousness*, Cambridge University Press.
- Baars, B. (2002) The Conscious Access Hypothesis, *Trends in Cognitive Sciences*, 6, 47-52.
- Baars, B., Ramsey, T., Laureys, S. (2003) Brain, Conscious Experience, and the Observing Self, *Trends in Neuroscience*, 26, 671-675.
- Bayne, T., Montague, M. (2011a), Cognitive Phenomenology: An Introduction, in Bayne, T., Montague, M. (eds.), *Cognitive Phenomenology*, Oxford University Press, 1-34.
- Bayne, T., Montague, M.; eds. (2011b), *Cognitive Phenomenology*, Oxford University Press
- Bishop, M. (2009) Why Computers Can't Feel Pain, *Minds and Machines*, 19, 507-516.
- Block, N. (1995) On a Confusion about a Function of Consciousness, *The Behavioral and Brain Sciences*, 18, 227-247.
- Bringsjord, S. (2007) Offer: One Billion Dollars for a Conscious Robot; If You're Honest, You Must Decline, *Journal of Consciousness Studies*, 14, 28-43.
- Cerullo, M. (2011) Integrated Information Theory, *J. of Consciousness Studies*, 18, 45-58.
- Charkaoui, N. (2005) A Computational Model of Minimal Consciousness Functions, *Proc. World Academy of Science, Engineering and Technology*, 9, 78-85.
- Chalmers, D. (1996) *The Conscious Mind*, Oxford University Press.
- Chalmers, D. (2000) What is a Neural Correlate of Consciousness? In: Metzinger, T. (ed.), *Neural Correlates of Consciousness*, MIT Press, 17—39.
- Chella, A. (2007) Towards Robot Conscious Perception, in Chella, A. & Manzotti, R. (eds.), *Artificial Consciousness*, Imprint Academic, 124-140.
- Chella, A. and Gaglio, S. (2012) Synthetic Phenomenology and High-Dimensional Buffer Hypothesis, *International Journal of Machine Consciousness*, 4, 353-365.
- Chudnoff, E. (2015) *Cognitive Phenomenology*, Routledge Press.
- Cleeremans, A. (2005) Computational Correlates of Consciousness, in S. Laureys (ed.), *Progress in Brain Research*, 150, 81-98.
- Cleeremans, A., Timmermans, B., Pasquali, A. (2007) Consciousness and Metarepresentation: A Computational Sketch, *Neural Networks*, 20, 1032-1039.
- Connor, D., Shanahan, M. (2010), A Computational Model of a Global Neuronal Workspace with Stochastic Connections, *Neural Networks*, 23, 1139-1154.
- Cotterill, R. (2003) CyberChild – A Simulation Test-Bed for Consciousness Studies, *Journal of Consciousness Studies*, 10, 31-45
- Crick, F., Koch, C. (1990) Towards a Neurobiological Theory of Consciousness, *Seminars in the Neurosciences*, 2, 263-275.

- Cvrckova, F., Lipavska, H., and Zarsky, V. (2009) Plant Intelligence, *Plant Signaling and Behavior*, 4, 394-399.
- Dawkins, R. (1976) *The Selfish Gene*, Oxford University Press, page 63.
- Dehaene, S., Kerszberg, M., Changeux, J. (1998) A Neuronal Model of a Global Workspace in Effortful Cognitive Tasks, *Proc. National Academy of Sciences*, 95, 14529-14534.
- Dehaene, S., Naccache, L. (2001) Towards a Cognitive Neuroscience of Consciousness, *Cognition*, 79, 1-37.
- Fekete, T., Edelman, S. (2011) Towards a Computational Theory of Experience, *Consciousness and Cognition*, 20, 807-827.
- Gallup, G. (1970) Chimpanzees: Self-Recognition, *Science*, 167, 86-87.
- Gamez, D. (2010) Information Integration Based Predictions about the Conscious States of a Spiking Neural Network, *Consciousness and Cognition*, 19, 294-310.
- Graves, A., Wayne, G., and Danihelka, I. (2014) Neural Turing Machines, arXiv preprint arXiv:1410.5401
- Graziano, M., Webb, T. (2014) A Mechanistic Theory of Consciousness, *International Journal of Machine Consciousness*, 6, 163-175.
- Grossberg, S. (1999) The Link between Brain Learning, Attention, and Consciousness, *Consciousness and Cognition*, 8, 1-44.
- Haikonen, P. (2007) Essential Issues of Conscious Machines, *Journal of Consciousness Studies*, 14, 72-84.
- Haikonen, P. (2012) *Consciousness and Robot Sentience*, World Scientific.
- Huynh, T., Reggia, J. (2012) Symbolic Representation of Recurrent Neural Network Dynamics, *IEEE Transactions on Neural Networks and Learning Systems*, 23, 1649-1658.
- Ijspeert, A. (2008) Central Pattern Generators for Locomotion Control in Animals and Robots, *Neural Networks*, 21, 642-653.
- Kitamura, T., Tahara, T. & Asami, K. (2000) How Can a Robot Have Consciousness? *Advanced Robotics*, 14, 263-275.
- Koch, C., Tsuchiya, N. (2006) Attention and Consciousness: Two Distinct Brain Processes, *Trends in Cognitive Sciences*, 11, 16-22.
- Kuipers, B. (2005) Consciousness: Drinking from the Firehose of Experience, *Proc. 20th National Conference on Artificial Intelligence*, AI Press, 1298-1305.
- Kuipers, B. (2008) Drinking from the Firehose of Experience, *Artificial Intelligence in Medicine*, 44, 155-170.
- Lamme, V. (2003) Why Visual Attention and Awareness are Different, *Trends in Cognitive Sciences*, 7, 12-18.
- Levine, J. (1983) Materialism and Qualia: The Explanatory Gap, *Pacific Philosophical Quarterly*, 64, 354-361.
- Lycan, W. (2009) Higher-Order Representation Theories of Consciousness, in Bayne, T., Cleeremans, A., & Wilken, P., *The Oxford Companion to Consciousness*, Oxford University Press, 346-350.
- Manzotti, R. (2012). The Computational Stance is Unfit for Consciousness, *International Journal of Machine Consciousness*, 4, 401-420.
- Marder, M. (2012) Plant Intentionality and the Phenomenological Framework of Plant Intelligence, *Plant Signaling and Behavior*, 7, 1-8.

- Massimini, M., Ferrarelli, F., Huber, R., et al. (2005) Breakdown of Cortical Effective Connectivity During Sleep, *Science*, 309, 2228-2232.
- Mathis, D., Mozer, M. (1995) On the Computational Utility of Consciousness, in G. Tesaro, D. Touretzky, T. Leen (eds.), *Adv. in Neural Information Processing Systems*, 7, 10-18.
- McCauley, L. (2007) Demonstrating the Benefit of Computational Consciousness, *Proc. of the AAAI Fall 2007 Symposia*, 108-114.
- McCulloch, W., Pitts, W. (1943) A Logical Calculus of the Ideas Immanent in Nervous Activity, *Bulletin of Mathematical Biophysics*, 5, 115-133.
- McDermott, D. (2007) Artificial Intelligence and Consciousness, in P. Zelazo, M. Moscovitch, E. Thompson (eds.), *Cambridge Handbook of Consciousness*, Cambridge University Press, 117-150.
- McGinn, C. (2004) *Consciousness and Its Origins*, Oxford University Press.
- Metzinger, T. (2000a) The Subjectivity of Subjective Experience, in *Neural Correlates of Consciousness*, T. Metzinger (editor), MIT Press, 285-306.
- Metzinger, T. (2000b) *Neural Correlates of Consciousness*, MIT Press.
- Monner D, Reggia J. (2011) Systematically Grounding Language Through Vision in a Deep Recurrent Neural Network, *Proc. Fourth Intl. Conf. on Artificial General Intelligence*, Schmidhuber J, Thorision K, Looks M (eds.), Springer, 112-121.
- Monner D, Reggia J. (2012) Emergent Latent Symbol Systems in Recurrent Neural Networks, *Connection Science*, 24, 193-225.
- Nagel, A. (1997) Are Plants Conscious?, *Journal of Consciousness Studies*, 4, 215-230.
- Newman, J., Baars, B., Cho, S. (1997) A Neural Global Workspace Model for Conscious Attention, *Neural Networks*, 10, 1195-1206.
- O'Brien, G., Opie, J. (1999) A Connectionist Theory of Phenomenal Experience, *Behavioral and Brain Sciences*, 22, 127-196.
- Pasquali, A., Timmermans, B., Cleeremans, A. (2010) Know Thyself: Metacognitive Networks and Measures of Consciousness, *Cognition*, 117, 182-190.
- Perlis, D. (1997) Consciousness as Self-Function, *Jour. of Consciousness Studies*, 4, 509-525.
- Piper, M. (2012) You Can't Eat Causal Cake with an Abstract Fork, *Journal of Consciousness Studies*, 19, 154-180.
- Raffone, A., Pantani, M. (2010) A Global Workspace Model for Phenomenal and Access Consciousness, *Consciousness and Cognition*, 19, 580-596.
- Reggia, J. (2013) The Rise of Machine Consciousness, *Neural Networks*, 44, 112-131.
- Reggia, J., Huang, D., Katz, G. (2015) Beliefs Concerning the Nature of Consciousness, *Journal of Consciousness Studies*, 22, 146-171.
- Reggia, J., Monner, D., Sylvester, J. (2014) The Computational Explanatory Gap, *Journal of Consciousness Studies*, 21 (9), 153-178.
- Rey, G. (1983) A Reason for Doubting the Existence of Consciousness, in Davidson, R., Schwartz, G., & Shapiro, D., *Consciousness and Self-Regulation*, Vol. 3, Springer.
- Rolls, E. (1997) Consciousness in Neural Networks?, *Neural Networks*, 10, 1227-1240.
- Rolls, E. (2007) A Computational Neuroscience Approach to Consciousness, *Neural Networks*, 20, 962-982.
- Rosenthal, D. (1996) A Theory of Consciousness. In N. Block, O. Flanagan, and G. Guzeldere (eds.), *The Nature of Consciousness*, MIT Press, 729-753.

- Samsonovich, A. (2007) Universal Learner as an Embryo of Computational Consciousness, in Chella, A. & Manzotti, R., editors, *AI and Consciousness: Theoretical Foundations and Current Approaches*, AAAI Fall Symposium, AAAI Technical Report FS-07-01, 129-134.
- Samsonovich, A., Nadel, L. (2005) Fundamental Principles and Mechanisms of the Conscious Self, *Cortex*, 41, 669-689.
- Sanz, R., Hernandez, C., Sanchez-Escribano, M. (2012) Consciousness, Action Selection, Meaning and Phenomenic Anticipation, *International Journal of Machine Consciousness*, 4, 383-399.
- Schlagel R (1999) Why not Artificial Consciousness or Thought? *Minds and Machines*, 9, 3-28.
- Schneider, W., Chein, J. M. (2003). Controlled and Automatic Processing: Behavior, Theory, and Biological Mechanisms, *Cognitive Science*, 27(3), 525-559.
- Searle, J. (2004) *Mind*, Oxford University Press.
- Seth, A. (2009) The Strength of Weak Artificial Consciousness, *International Journal of Machine Consciousness*, 1, 71-82.
- Seth, A., Izhikevich, E., Reeke, G., and Edelman, G. (2006) Theories and Measures of Consciousness, *Proceedings of the National Academy of Sciences*, 103, 10799-10804.
- Shanahan, M. (2008) A Spiking Neuron Model of Cortical Broadcast and Competition, *Consciousness and Cognition*, 17, 288-303.
- Siewert, C. (2011) Phenomenal Thought, in Bayne, T., Montague, M. (eds.), *Cognitive Phenomenology*, Oxford University Press, 236-267.
- Slooman, A., Chrisley, R. (2003) Virtual Machines and Consciousness, *Journal of Consciousness Studies*, 10, 133-172.
- Sun, R. (1999) Accounting for the Computational Basis of Consciousness, *Consciousness and Cognition*, 8, 529-565.
- Sun, R. (2002) *Duality of the Mind*, Hillsdale, NJ: Erlbaum.
- Sylvester J, Reggia J, Weems S, Bunting M (2013) Controlling Working Memory with Learned Instructions, *Neural Networks*, 41, 23-38.
- Sylvester J, Reggia J (2016) Engineering Neural Systems for High-Level Problem Solving, *Neural Networks*, 79, 37-52.
- Tagliazucchi, E., Chialvo, D., Siniatchkin, M. et al. (2016) Large-Scale Signatures of Unconsciousness are Consistent with a Departure from Critical Dynamics, *Jrnl. of the Royal Society Interface*, 13.
- Takeno, J. (2008) A Robot Succeeds in 100% Mirror Image Cognition, *International Journal on Smart Sensing and Intelligent Systems*, 1, 891-911.
- Takeno, J. (2013) *Creation of a Conscious Robot*, Pan Stanford.
- Taylor, J. (1997) Neural Networks for Consciousness, *Neural Networks*, 10, 1207-1225.
- Taylor, J. (2003) Neural Models of Consciousness, in M. Arbib (editor), *Handbook of Brain Theory and Neural Networks*, MIT Press, 263-267.
- 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.
- Tononi, G. (2008) Consciousness as Integrated Information, *Biological Bulletin*, 215, 216-242.
- Treisman, A. (2009) Attention: Theoretical and Psychological Perspectives, in Gazzaniga, M. (editor), *The Cognitive Neurosciences*, MIT Press, 189-204.
- van der Velde, F., de Kamps, M. (2006) Neural Blackboard Architectures of Combinatorial Structures in Cognition, *Behavioral and Brain Sciences*, 29, 37-108.

Weems, S., Reggia, J. (2006) Simulating Single Word Processing in the Classic Aphasia Syndromes, *Brain and Language*, 98, 291-309.

Acknowledgement: This work was supported by ONR award N000141310597.