The Computational Explanatory Gap

January 2014

James A. Reggia, Derek Monner, and Jared Sylvester
University of Maryland

Abstract

Efforts to study consciousness using computational models over the last two decades have received a decidedly mixed reception. Investigators in mainstream AI have largely ignored this work, and some members of the philosophy community have argued that the whole endeavor is futile. Here we suggest that very substantial progress has been made, to the point where the use of computational simulations has become an increasingly accepted approach to the scientific study of consciousness. However, efforts to create a phenomenally conscious machine have been much less successful. We believe that a major reason for this is a computational explanatory gap: our inability to understand/explain the implementation of high-level cognitive algorithms in terms of neurocomputational processing. Contrary to prevailing views, we suggest that bridging this gap is not only critical to further progress in the area of machine consciousness, but is also an important step towards understanding the hard problem. We briefly describe some recent progress that has been made towards bridging this gap, and assess whether any computational correlates of consciousness have been identified.

Correspondence: James Reggia. Dept. of Computer Science, A.V. Williams Bldg., University of Maryland, College Park, MD 20742, USA (reggia@cs.umd.edu)

1 Introduction

Is it possible to create a conscious machine? While this is not a particularly new question (Butler, 1872), it is only over the last two decades that it has motivated sustained work on developing computational models of the conscious mind, either via software on computers or by using physical robots. Such studies concerning artificial consciousness have been intended to advance our understanding of human consciousness and its relationship to cognition, to contribute to increased functionality in future AI systems, and (at times) to design a phenomenally conscious machine.

At present, efforts to study artificial consciousness remain highly controversial. Researchers in mainstream AI (with a few exceptions) have largely ignored work in this area. In philosophy, while a variety of opinions have been expressed, a significant number of these would make pursuit of machine consciousness appear to be a rather fruitless task. For example, it has been argued that, in general, the objective methods of science cannot shed light on consciousness due to its subjective nature (McGinn, 2004), making computational investigations a moot point. Further, a broad range of more specific arguments have been presented in recent years that phenomenal machine consciousness is simply not possible (Bishop, 2009; Bringsjord, 2007; Manzotti, 2012; Schlagel, 1999). Individuals who advocate or study the possibilities of machine consciousness have so far not found such arguments persuasive (Aleksander, 2005; Koch & Tononi, 2008).

In a recent review of work in this area, we found that, in contrast to what one might expect based on such negative viewpoints, very substantial progress has been made over the last twenty years in the field of artificial consciousness (Reggia, 2013). Here we show that by distinguishing between simulated correlates of consciousness and instantiated consciousness, it is possible to clearly delineate where significant progress is being made, and where the jury is still out. We then ask, from a purely computational/engineering viewpoint: What is the main practical barrier to further progress on creating phenomenal/instantiated machine consciousness? Our answer is that it is a *computational explanatory gap*: our current lack of understanding concerning how high-level cognitive computations can be captured in low-level neural computations. The significance of this gap is that bridging it may be a critical step not only in developing a future conscious machine, but also in addressing the original philosophical explanatory gap, in gaining a better understanding of the neural correlates of consciousness, and in making advances on the mind-brain problem in general during coming years. We argue that bridging this gap has the potential to identify new *computational correlates of consciousness*, i.e., computational processes that are exclusively associated with conscious information processing (Cleeremans, 2005). A summary is given of some (small) steps that have been taken recently towards bridging the computational explanatory gap. We

find that while a number of candidate correlates have been identified, at present none of these can yet fully meet our criteria for being computational correlates of consciousness.

2 The Nature of Progress in Artificial Consciousness

To understand the sense in which recent work on artificial consciousness has made progress, it is useful to distinguish between two possible objectives for such work: simulation of correlates of consciousness versus instantiation of consciousness. Such a distinction parallels the distinction often made between information processing aspects of consciousness (relates to functionalism, access consciousness (Block, 1995)) and subjective experience (phenomenal consciousness).

With *simulated correlates of consciousness*, the goal is to capture some aspect of the neural/behavioral/cognitive correlates of consciousness in a computational model, much as is done in using computers to simulate other natural processes (e.g., models of weather/climate). There is nothing particularly mysterious about such work; just as we would not expect that a computer used to simulate a thunderstorm would become wet inside, we should not expect that a computer used to model some aspect of information processing associated with consciousness would become "conscious inside". There is no claim that phenomenal consciousness is actually present in this situation. The results of a simulation are often assessed based on the extent to which they correspond to experimentally verified correlates of consciousness such as neurophysiological measures, or on the extent to which they may contribute increased functionality to future artificial systems. In contrast, with *instantiated consciousness*, the issue is the extent to which an artificial system actually has phenomenal consciousness. Does it experience qualia and does it have subjective experiences? This is a much more difficult and controversial question. The dichotomy between simulated correlates and instantiated consciousness is reminiscent of the distinction between weak AI (behavioral criteria) and strong AI (artificial mind) (Seth, 2009).

Recognizing the difference between simulated correlates of consciousness and instantiated machine consciousness clarifies the nature of the progress that has been made in artificial consciousness research over the last two decades. From the perspective of simulated correlates, neurocomputational modeling has successfully captured a number of neurobiological, cognitive and behavioral correlates of conscious information processing as machine simulations. To give just a few examples:

- Neurocomputational models that increase activation of their global workspace when performing difficult tasks associated with conscious effort in people (Dehanene et al., 1998), supporting global workspace theories of consciousness (Baars, 1988, 2002).
- The unexpected finding that information integration theory (Tononi, 2004) identifies gating modules as the most conscious components of a neurocontroller (Gamez, 2010),

linking gating mechanisms in cognitive control (Sylvester et al., 2013) to consciousness studies.

- Demonstration that expectation-driven robots can recognize themselves in a mirror (Takeno, 2013), essentially passing the well-known mirror test used to identify self-recognition in animals.
- Showing that second-order neural networks can match behavioral data from human blindsight subjects during post-decision wagering tasks (Pasquali et al., 2010), supporting some aspects of higher order thought (HOT) theories of consciousness (Carruthers, 2005).
- Establishing that corollary discharge signals in neurocomputational models of human top-down attention control mechanisms can account for some human data involving conscious information processing (Taylor et al., 2007).

Clearly, these and other computational models of correlates of consciousness have provided useful information for advancing consciousness studies. As a result, they are increasingly being viewed as an acceptable approach to furthering the scientific investigation of consciousness, and they may help constrain the large number of theories of consciousness that exist (Katz, 2013) by clarifying their implications via modeling. A more detailed discussion of these and many other related models is given in (Reggia, 2013).

The situation is quite different from the perspective of instantiated consciousness. Several investigators have claimed to know how to create phenomenally conscious artifacts. To give just a few examples:

- 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).
- Within the framework of global workspace theory, it is possible to develop models that show in a transparent fashion plausible neurocomputational bases for phenomenal and access consciousness (Raffone and Pantani, 2010).
- Adaptive resonance theory predicts that all conscious states are resonant states (Grossberg, 1999).
- A system has subjective experience to the extent that it has the capacity to integrate information (Tononi, 2004).
- Computational systems supporting higher order syntactic thoughts experience qualia (Rolls, 2007).

On the other hand, substantial arguments have also been presented that instantiated machine consciousness is simply not possible. A few examples are:

- Machines cannot be conscious due to their non-organic nature (Schlagel, 1999).
- Phenomenal machine consciousness would imply panpsychism (Bishop, 2009).
- The absence of a formal definition of consciousness precludes conscious machines (Bringsjord, 2007).
- Computation is insufficient to underpin consciousness (Manzotti, 2012).

Our intent here is not to explore in depth the claims and counter-claims above, but only to demonstrate that the issue of instantiated machine consciousness remains controversial.

We believe that, at the present time, no existing computational approach to artificial consciousness has yet presented a compelling demonstration or design of instantiated consciousness in a machine, or even clear evidence that instantiated machine consciousness will eventually be possible. Yet at the same time, no compelling refutation of the possibility of a phenomenally conscious machine has been generally accepted. From a history-of-technology perspective, the current situation resembles discussions concerning the possibility of heavier-than-air machine flight during the late 1800s. During this period, as primitive efforts were underway to create such machines, very plausible arguments were being made about the impossibility of machine flight, and these arguments were only put to rest when the Wright brothers' took wing at Kitty Hawk[1]. It remains to be seen whether work on artificial consciousness will have an analogous outcome. In the mean time, the current situation raises the issue of what can be done to resolve whether or not instantiated machine consciousness is possible. Resolution of this issue depends on clearly identifying the main barriers to further progress that are tractable, or at least amenable to scientific/technical investigation, and then hopefully overcoming them.

3 The Computational Explanatory Gap

What is the main practical barrier at present to creating instantiated machine consciousness? Clearly, 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 we possibly know whether or not a machine is conscious?). While each of these difficulties is substantial, from our 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

[1] This analogy with flight is limited in that consciousness involves the other-minds problem, unlike flight. The analogy also raises the question of whether consciousness is an all-or-nothing property or lies on a continuum in the same way that "flying" does (person jumping, flying squirrel, hang gliding, bird flight, sustained plane flight, orbiting satellite, etc.). Our primary point here is just that prior to 1900, many informed people argued that heavier-than-air flight was impossible.

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 argument here is that there is another less recognized barrier, the *computational explanatory gap*, that we would argue is also of critical importance. We define the computational explanatory gap very specifically: It is the current lack of understanding of how high-level cognitive information processing can be mapped onto low-level neural computations. By "high-level cognitive information processing", we mean aspects of cognition such as goal-directed problem solving, executive decision making, planning, language, 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 we will argue below that 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, it is a purely computational issue (Figure 1).

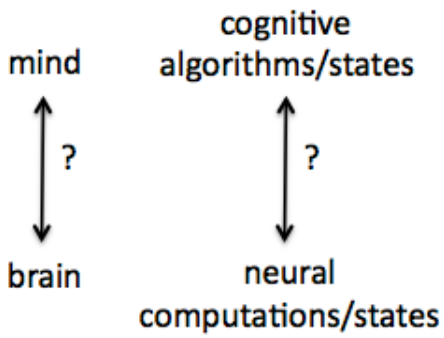

Figure 1: The well-known philosophical (left) and less recognized computational (right) gaps. Our argument is that the latter may ultimately prove to be the more fundamental problem, and that focusing on solving it rather than dismissing it may be the key to advancing future work on instantiated machine 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. If this abstraction is significant, then one would expect to

find that it can be related to past consciousness studies occurring across a broad range of disciplines. Is this the case? We suggest in the following that not only is this true for disciplines such as philosophy, AI, cognitive psychology, and neuroscience, but also that the computational explanatory gap provides a prism through which to potentially unify past work relevant to conscious information processing across these disciplines.

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 (Chalmers, 1996). This characterization 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, we conjecture that, with high probability, the unfortunate terminology of the "easy" and "hard" problems that dominates contemporary philosophical thought will ultimately turn out to be precisely backwards. In other words, 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. We elaborate on this issue in the next section.

In AI the computational explanatory gap relates to the long-standing debate concerning the relative values of top-down (symbolic, numerical, etc.) vs. bottom-up (neural, swarm, etc.) approaches to creating machine intelligence (Franklin, 1995). This (in)famous and continuing debate has largely missed the point that these two approaches are not so much competing alternatives as complementary in what they have each captured about intelligence. Top-down symbolic methods have excelled at modeling high-level cognitive tasks such as goal-directed reasoning, metacognition, problem-solving, decision making, "understanding" natural language, and planning, but they have been much less successful at pattern recognition and low level motor control. In other words, they have been relatively successful in capturing aspects of high-level deliberative reasoning and sequential behavioral control that, in a person, are associated with conscious, reportable components of cognition. Top down methods have also generally been found to be brittle, for example, failing dramatically in the

context of noise, minor unexpected events, or small changes to the contents of memory. In contrast, bottom-up neurocomputational methods have roughly the opposite strengths and weaknesses. They are remarkably effective and robust in learning low-level pattern classification ("input") and low-level control ("output") tasks, but are not nearly as effective for high-level cognitive tasks. For example, neurocomputational systems have proven to be very effective in learning procedural knowledge that, in a person, is largely carried out in an automatic unconscious manner. Examples include robotic arm movement neurocontrollers (Gentili et al., 2011) and steering of an autonomous automobile (Pomerleau, 1993). Compared to top-down symbolic AI methods, neurocomputational methods are less brittle in the context of external noise or substantial random changes to stored information (random weight changes, random loss of simulated neurons, etc.). In short, past work in mainstream symbolic AI has so far done relatively little to bridge the computational explanatory gap. The field of AI largely remains split into two opposing camps today, in part directly due to the computational explanatory gap.

In cognitive psychology too, the computational explanatory gap becomes evident when one considers past efforts to characterize explicitly the properties that distinguish human conscious versus unconscious cognitive information processing. For example, human conscious information processing has been characterized as being serial, relatively slow, and largely restricted to one task at a time (attempting to carry out multiple tasks requiring conscious direction simultaneously leads to interference and errors) (Baars, 1988, 2002; Dehaene & 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,[2] 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 (no explanatory gap here). 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

[2] 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).

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. This is particularly remarkable because identifying the contrasting properties of conscious and unconscious aspects of cognition was not an explicit effort to characterize the nature of the computational explanatory gap.

Finally, the computational explanatory gap is also evident in neuroscience, where substantial efforts 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). This approach to understanding consciousness has identified several candidate correlates, such as specific activity patterns in the brain's electrical activity (e.g., widespread 40 hz cortical oscillations), activation of specific neural structures (e.g., thalamic intralaminar nuclei), global brain activation, etc. (Metzinger, 2000b). However, in spite of such work and an enormous neuroscientific endeavor in general over more than a century, there remains a large difference between our understanding of unconscious versus conscious information processing in the brain. Consider unconscious information processing, such as the automatic motor control mechanisms underlying leg movements during walking, or the mechanisms that store information in long-term memory. There is nothing really mysterious about these mechanisms today in the sense that we can identify neural circuits and computations that plausibly account for these functions: central pattern generators for oscillatory movements, Hebbian synaptic changes for associative memory, and so forth. In contrast, for high-level cognitive tasks that relate more closely to conscious cognition, such as goal-directed problem solving and understanding the meaning of spoken natural language, we remain relatively lost in neurocomputational terms. This is in spite of the fact that we know a great deal about high-level cognition and the brain today. For example, a lot 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 these high-level cognitive functions into computations over the low-level neural circuits that are present. Once again, we encounter the computational explanatory gap. It remains unclear why this mapping from cognitive processes to neural computations is so opaque today given the enormous resources that have been poured into understanding these issues. This widely-recognized situation has led to a recent call by prominent neuroscientists for a "brain activity map initiative" that would develop the technology for bridging this gap (Alivisatos et al., 2013). 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 to one another is, at least in part, also a manifestation of the underlying computational explanatory gap.

4 Implications of Bridging The Gap

Why is bridging the computational explanatory gap of critical importance in addressing the possibility of instantiated machine consciousness? The reason is that bridging this gap would allow us to do something that is currently beyond our reach: 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, it 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 (Cleeremans, 2005). By "computational correlate of consciousness", we mean minimal computational processing mechanisms that are specifically associated with conscious aspects of cognition but *not* with unconscious aspects. In the context of the computational explanatory gap, we are specifically interested in *neuro*computational 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. Neurocomputational correlates are abstractions that may well find implementation in the brain and thus can sometimes also be candidates for neural correlates of consciousness, but as abstractions they are intended to be independent of the physical substrate that implements them (brain, silicon, etc.).

We believe that if convincing neurocomputational correlates can be identified, many individuals concerned with the mind-brain problem would concede that this could at least provide insight into the nature of access consciousness. However, our suggestion here is that, should neurocomputational correlates of consciousness be discovered, they may also provide a direct route to investigating the possibility of instantiated machine consciousness, to identifying candidate properties that could serve as objective criteria for the presence/absence of phenomenal consciousness in machines and people, and perhaps even to a better understanding of the fundamental nature of consciousness. Pursuit of neurocomputational correlates of consciousness is a worthy endeavor regardless of the ultimate outcome of such work: Even if no differences between the neurocomputational implementation of conscious and unconscious cognitive functions can be found, that too would have tremendous implications for the modern functionalist viewpoint of the mind-brain problem, for example lending support to theories that incorporate some aspect of dualism (e.g., naturalistic dualism (Chalmers, 1996)).

In other words, we are suggesting that a complete characterization of high-level cognition in neurocomputational terms may, along the way, show us how subjective experience arises mechanistically. We make this claim in the context of the increasing discussion of cognitive phenomenology in philosophy during recent years (reviewed in (Bayne and Montague, 2011)). Cognitive phenomenology emphasizes that our subjective first-person experience is not restricted to just traditional qualia involving perception and affect, but also 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, 2011), 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.

Further, even if bridging the computational explanatory gap does not provide a full mechanistic account of subjective experiences, doing so would still provide more meaningful conditions for investigating phenomenal consciousness in an artifact. At the least, we believe that bridging the computational explanatory gap could make the hard problem largely fade away. A familiar historical analogy with vitalism may help in understanding why this latter point is at least plausible.[3] The concept of life seemed just as mysterious to many scientists during the early 1800's as the concept of consciousness does to us today. As a result, many scientists at the time accepted the philosophical doctrine of *vitalism* (Garrett, 2006). Vitalists attributed some non-physical "vital spirit" or "life force" to living entities that was not possessed by inanimate objects. In other words, vitalists believed that the laws of physics and chemistry by themselves would never be able to account fully for living processes. In effect, there was an apparent philosophical explanatory gap between the property of being alive and what could be accounted for mechanistically (i.e., there existed a biological version of the "hard problem"), similar to the philosophical explanatory gap concerning consciousness today. However, at present we believe that much of the mystery underling this philosophical explanatory gap concerning life was really due to a "biological explanatory gap", i.e., to the limited scientific understanding two hundred years ago of how processes associated with living entities (metabolism, reproduction, inheritance, etc.) could be implemented by biochemical and biophysical mechanisms. Today, even though much is still not understood about the physics and chemistry of living processes, and there is still not even a generally agreed upon definition of life (Regis, 2008; Wolfram, 2002), vitalism and much of the mysteriousness of life that gave rise to it has faded away. This has occurred due to scientific advances, including the ability to synthesize organic molecules from inorganic ones, a mechanistic understanding of cellular energy metabolism, our knowledge of molecular

[3] While some philosophers support the analogy with vitalism that we use here (e.g., Dennett, 1996), others have argued that it is inadequate because consciousness involves the mysteries of subjective experience while life does not (e.g., Chalmers, 1996, 2007). We believe that the latter view simply reflects the demystification of the concept of life that has occurred for contemporary authors due to scientific advances in the biosciences. See (Garrett, 2006) for further recent arguments supporting the merit of this analogy.

genetics and evolution governed by DNA, and the recognition that self-organization based on simple rules can generate high-level behaviors as emergent properties in complex systems (Wolfram, 2002). If we replace "life" with "consciousness", and "biological explanatory gap" with "computational explanatory gap", the analogy is clear. Our expectation, based on this historical analogy, is that even if bridging the computational explanatory gap does not fully account for the hard problem of subjective experience or produce a crisp definition of phenomenal consciousness, it will at least greatly demystify it, and consequently the hard problem may fade away.

5. Identifying Computational Correlates of Consciousness

If one allows the possibility that the "easy problem", represented in part by the computational explanatory gap, is important and a substantial barrier to instantiated machine consciousness, then the immediate research program becomes determining how we can bridge this gap. Encouragingly, there has been a substantial and increasing effort over the last two decades by researchers who are examining issues that relate to such a program. What computational correlates of consciousness has this work proposed? If one surveys past work in the area of artificial consciousness (Reggia, 2013), one is left with two immediate observations that are relevant to the computational explanatory gap. First, influenced by the distinction made in AI between top-down symbolic computations and bottom-up neural computations, a number of past models that are explicitly intended to explain consciousness are based on the distinction between local symbolic representations versus distributed neural representations (Chella, 2007; Kitamura et al, 2000; Sun, 1999, 2002). These hybrid models consist of a high-level symbolic module that interacts with one or more lower-level neurocomputational modules. Such models start with the *claim* that symbolic information processing per se in the high-level module is the basis of conscious information processing, and thus they essentially build in the computational explanatory gap. However, while such models implicitly recognize this gap, they do not attempt to provide a solution to it. The critical issue raised by the computational explanatory gap is how to *replace* the symbolic modules of such models with neurocomputational implementations. How such replacement could be done remains stubbornly mysterious today, and is the essence of the computational explanatory gap.

The second observation that emerges from reviewing past models in artificial consciousness is that many of these models can be interpreted as being founded upon specific candidates for computational correlates of consciousness, and thus are more directly related to addressing the computational explanatory gap. Considering these models, Cleeremans (2005) proposed two candidate computational correlates of consciousness a number of years ago: the quality of a representation (its stability over time, strength, and distinctiveness), and the extent to which a representation is referred to by other representations. The latter hypothesized computational correlate has especially received attention in subsequent metacognitive neural

network models where higher-order networks interpret lower level networks' activities (Cleeremans et al., 2007; Pasquali et al., 2010), linking work on machine consciousness to a rich history of philosophical discussions about higher-order thought (HOT) theories. Other studies of artificial consciousness have been founded on hypotheses that can be viewed as potential computational correlates of consciousness, including global information processing (Dehaene et al., 1998), the ability of a neural architecture to integrate information into a unified experience (Gamez, 2010; Tononi, 2004), resonating neural activity (1999), neural computations that implement self-models and self-recognition (Takeno, 2013), the grounding of symbols in sensory data streams (Kuipers, 2005), and various aspects of attention control (Coward and Gedeon, 2009; Haikonen, 2007, 2012; Tinsley, 2008; Starzyk & Prasad, 2011; Taylor, 2007).

The hypothesized computational correlates above have emerged from work explicitly directed at understanding consciousness. The viewpoint produced by the computational explanatory gap thus suggests that there is a third important observation to be made: Additional computational correlates of consciousness may be generated by studies of *neurocognitive architectures*. Unlike work explicitly studying artificial consciousness, this latter work attempts to map higher cognitive functions into neurocomputational mechanisms, *independently of any explicit relationship between these functions and consciousness*. Has this work (perhaps unintentionally) identified neurocomputational mechanisms that are specifically associated with conscious aspects of higher cognition? Past studies in this area far exceed what we can summarize here, so we focus on just two specific examples of our own work related to higher-level cognition, goal-directed cognitive control of working memory and the grounding of language, to support the claim that such studies are also generating hypotheses for computational correlates of consciousness.

Our first example involves cognitive control, an umbrella term for goal-directed executive cognitive systems that manage other cognitive processes, such as working memory, planning, attention, and action selection. While developing neural architectures capable of modeling cognitive control processes is recognized as an important research direction today (Roy, 2008), actually doing so has proven to be surprisingly challenging, consistent with the concept of a computational explanatory gap. Neural systems struggle to represent the goals and rules of various cognitive tasks (Marcus, 2001), in striking contrast to symbolic AI systems (Simen et al., 2010). Growing interest in this issue has led to the development of pioneering neural models that explicitly incorporate aspects of cognitive control, such as for planning solutions to the Towers of London Problem (Dehaene & Changeux, 1997). However, for working memory (short-term memory having very limited capacity (Baddeley, 2000)), the vast majority of neurocomputational models have no endogenous control mechanisms at all, with control being managed externally by a human. Further, many are quite specific to their given task, being unable to generalize to variations without major re-implementation (for exceptions, see Rougier et al., 2005).

In our own work over the last few years, we have developed a number of attractor neural network models of working memory that learn temporal sequences, and have shown that their performance can match empirical data from human behavioral experiments (Winder et al., 2009; Sylvester et al., 2010). But like most past working memory models, these did not include cognitive control mechanisms. To address this deficit, we have recently been studying a neural architecture called GALIS that focuses on learned cognitive control via gating mechanisms (Sylvester et al., 2013). GALIS is inspired by existing knowledge of the regions and pathways of prefrontal cortex. Cortical regions are represented as sequence-learning attractor neural network modules. The same neurocomputational mechanisms that support working memory in one model region during executing a specific task are used in other regions for autonomously learning the temporal instruction sequence needed to control that task's performance (Figure 2). The key functional enhancement that was needed to make this work effectively was to allow GALIS' cortical control region to gate the actions of other cortical regions and itself. In other words, 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. GALIS has been applied successfully to learn to perform n-back working memory tasks (a standard psychological task) simultaneously for different values of n, producing accuracy and timing results reminiscent of those seen in humans performing similar tasks, and making testable predictions (Sylvester et al, 2013).

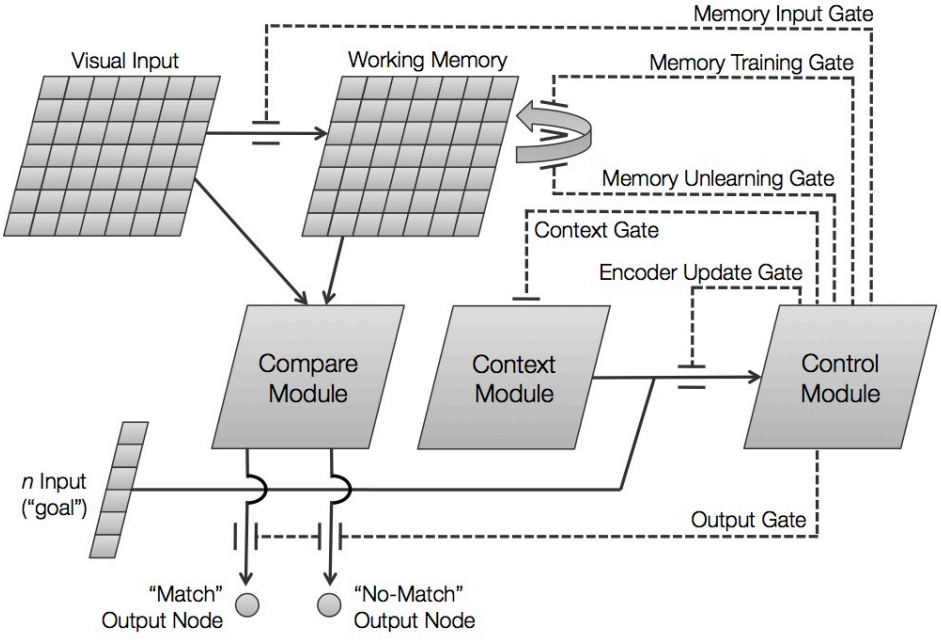

Figure 2: The GALIS model as used for the n-Back task. Thin, solid arrows are one-to-one connections. The working memory layer is recurrently connected (broad arrow). Dashed lines are the gating outputs of the control module. The number of boxes pictured in each layer is an approximation only, and does not faithfully represent the number of nodes used in the model.

GALIS is a small step towards bridging the computational explanatory gap: its ability to learn temporal sequences of discrete "instructions" in a "computer-like fashion" produces a level of executive functionality that is more typically associated with top-down symbolic AI systems. Does the theory of cognitive control embedded in GALIS's neural networks suggest any neurocomputational correlates of consciousness? The two critical computational mechanisms in GALIS that might be viewed as candidate correlates are the ability to learn temporal sequences of discrete instructions that are needed to perform a task, and the ability of a higher-level cortical control region to learn to gate other cortical regions performing the task (i.e., to turn other regions on/off).

Our second example of how work on neurocognitive models is suggesting computational correlates of consciousness involves the grounding of language: the binding of internally-represented symbols to external/internal phenomena, thereby giving these symbols meaning (Roy, 2005). Language has long been argued to be a key aspect of consciousness (Jaynes, 1976; Rolls, 2007), and the grounding of symbols in particular has been suggested as a critical aspect of both phenomenal and functional consciousness (Kuipers, 2005; van der Velde, 2013), making neurocomputational models of language grounding especially relevant to the computational explanatory gap. As with cognitive control, most past efforts to create neural systems that capture various aspects of natural language processing have found this to be a challenging task, and the neurocomputational models that currently exist do not come close in performance to what can be achieved by contemporary AI machine learning approaches.

Our own work over the last several years has focused on creating distributed-representation models of natural language grounding that are inspired by the neuroanatomical organization of human cortical language regions and pathways. Our initial model focused on learning to process single words (Weems and Reggia, 2006). Sequences of auditory phonemes representing the names of physical objects were grounded by associating them with simple visual images of the corresponding objects during learning. After training, the initially language-naïve model had learned to represent the "meaning" of heard words in terms of their visual representation as well as to name seen objects. Interestingly, simulated localized damage to model components (Wernicke's area, arcuate fasciculus, etc.) produced distinct behavioral deficits dependent on damage location; these deficits were reminiscent of the classical aphasia syndromes seen in people with brain damage. This model was very limited though in only processing single words.

To address this issue, we recently developed and studied a similar neural architecture having interacting auditory and visual pathways, but now processing multi-word sentences (see Figure 3). Input sentences consisted of phoneme sequences without inter-word breaks, so the model needed to learn to segment the phoneme sequences into words/morphemes as well as to associate the resulting tokens with objects in simple scenes that it observed. Using deep recurrent networks and a variant of long short-term memory (Monner and Reggia, 2012a), the

model learned to produce the meaning of sentences describing the simple external world that it observed, recovering appropriate meanings even for novel objects and novel descriptive sentences (Monner and Reggia, 2011). It also learned to correctly answer questions by observing question-answer pairs (Monner and Reggia, 2012b). Most importantly in the context of our current discussion is that formal analysis of the trained model found that it had learned/discovered 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 the learned distributed representation adopted by the network (Monner and Reggia, 2014). While the network's internal symbols are latent from our perspective as outside observers, they can be shown to be very real in the sense that they are accessed, manipulated and inspected by the neural architecture itself when interpreting sentences, including novel ones.

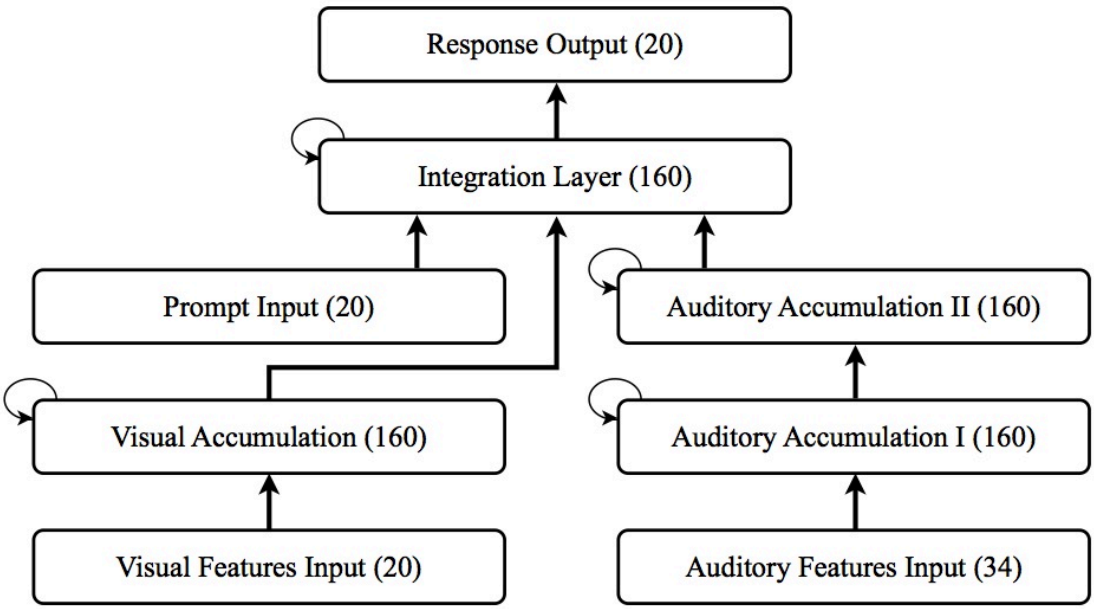

Figure 3: The architecture of our language-grounding model. Boxes represent layers of units (number of units in parentheses) and straight arrows represent banks of trainable connection weights between units of the sending and receiving layers. Layers of memory cells are denoted with curved arrows, representing the self-recurrence of units in these layers.

This grounding of language in neurocomputational models described above is quite limited in its scope and is only a small step towards bridging the computational explanatory gap. However, the demonstrated ability of a neurocomputational architecture to learn to ground symbols/words in terms of visual objects provides a kind of functionality that is usually associated with top-down symbolic AI systems. Does this suggest any neurocomputational correlates of consciousness? Learned gating of neural circuitry is again being used, but this time at a more fine-grained level of individual simulated neurons gating the input, output, or activity persistence of other individual neurons, rather than at the more

macroscopic level of gating of whole pathways and regions as was done in GALIS. Further, the emergence of a latent symbol system that exhibits combinatorial computations and systematic behaviors using a learned distributed representation of information is a key result of this work. Such a finding becomes relevant to consciousness studies to the extent that past theories concerning the role of language and symbol grounding in consciousness (Jaynes, 1976; Kuipers, 2005; Rolls, 2007) are ultimately found to be correct.

6 Discussion

Existing computational models have successfully captured a number of neurobiological, cognitive and behavioral correlates of conscious information processing as machine simulations. Put simply, it has been possible to develop limited forms of what we have called simulated correlates of consciousness. This achievement is extremely important: These computational studies are providing a way to test whether theories about key neural, cognitive and/or behavioral correlates 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 lead to future artificial agents that have increased functionality and are able to interact with people in more natural ways (Aleksander, 2013; Charkaoui, 2005; McCauley, 2007; Sanz et al., 2012). Put simply, work on simulated 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 no existing approach to artificial consciousness has presented a compelling demonstration of instantiated (phenomenal) consciousness in a machine, or even clear evidence that instantiated machine consciousness will eventually be possible. While some investigators have made intriguing claims that the approach they are using is or could be the basis for a phenomenally conscious machine, none is currently generally accepted as having done so. In our opinion, none of the past studies of which we are aware, even when claimed otherwise, has yet provided a convincing case for how a given methodology would eventually lead to instantiated artificial consciousness.

Our central argument here is that this apparent lack of progress towards a deeper understanding of instantiated machine 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 be inclined to dismiss this gap as just part of the “easy problem”, we think 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 (to our knowledge) provided no deep insight into *why* such an “easy problem” has proven to

be so intractable. On the contrary, we would argue that the computational explanatory gap is a fundamental issue that needs a much larger collective effort to clarify and resolve.

Further, recognizing this gap emphasizes that substantial past computational work reported outside of the artificial consciousness literature is of direct relevance to consciousness studies. As described above, a great deal of past work on neurocognitive architectures has attempted to map higher cognitive functions into neurocomputational processes, thereby directly addressing the computational explanatory gap, without explicitly considering the relevance of this work to machine consciousness or consciousness studies in general. As can be seen from the examples that we gave above, the results from this work, and from many other studies, suggest that there are some potential computational correlates of consciousness beyond those presented previously (Cleeremans, 2005) or identified in general by investigators working on artificial consciousness, such as the ability to learn sequential attractor states, the use of adaptive gating mechanisms to exert top-down influences, and the presence of a learned emergent symbol system.

Our current ability to identify several candidate computational correlates of consciousness is very encouraging. However, a fundamental difficulty remains: Most hypothesized computational mechanisms identified so far as correlates are either known to also occur in neurocomputational systems supporting non-conscious activities, or their occurrence has not yet been excluded in such settings. In other words, the main difficulty in identifying computational correlates of consciousness to date has been 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 a hierarchical neural architecture 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, hierarchical information processing 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 even self-modeling.

For this reason, it remains unclear whether any of the hypothesized candidates for computational correlates of consciousness, including those we suggest from our own studies, will ultimately prove to be satisfactory without further qualifications: 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 neurocomputational mechanisms are thus not *specifically* identifiable only with human conscious cognitive activities. In short, while a great deal of progress has been made by studies of machine consciousness and neurocognitive architectures over the last two decades, only limited progress has been made by this work that is relevant to resolving the computational explanatory gap and identifying computational correlates of consciousness.

As described earlier in this paper, we believe that bridging the computational explanatory gap will make bridging the philosophical explanatory gap more tractable, and that it may ultimately lead to an operational test for the presence of phenomenal consciousness. In other words, we are suggesting that a complete characterization of high-level cognition in neurocomputational terms, even if it does not explicitly show us how subjective experience arises mechanistically, will substantially demystify the nature of phenomenal consciousness and perhaps even provide necessary and sufficient tests for phenomenal consciousness in an artifact. Such a view is compatible with and complementary to recent arguments made by others in the expanding literature of cognitive phenomenology (Bayne & Montague, 2011). We believe that bridging the computational explanatory gap is possibly *the* most critical step we could take during the next decade to advance prospects for a phenomenally conscious artifact and a deeper understanding of the mind-brain problem. Perhaps if progress can be made in this way, the insights provided will reveal that the "hard problem" is ultimately much easier, or at least more understandable, than it currently appears.

References

Aleksander, I. (2005) *The World in my Mind, My Mind in the World*, Imprint Academic.

Aleksander, I. (2013) Phenomenal Consciousness and Biologically Inspired Systems, *International Journal of Machine Consciousness*, 5, 3 – 9.

Alivisatos, A., et al. (2013) The Brain Activity Map. *Science*, 339, 1284-1285.

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.

Baddeley, A. (2000) Short-term and Working Memory. In E. Tulving & F. Craik (Eds.), *The Oxford Handbook of Memory*, Oxford Univ. Press.

Bayne, T., Montague, M. (2011), Cognitive Phenomenology: An Introduction, in Bayne, T., Montague, M. (eds.), *Cognitive Phenomenology*, Oxford University Press, 1-34.

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.

Butler, S. (1872) *Erewhon*, 1872.

Carruthers, P. (2005) *Consciousness – Essays from a Higher-Order Perspective*, Oxford University Press.

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.

Chalmers, D. (2007) The Hard Problem of Consciousness, in Velmans, M. and Schneider, S. (eds.), *The Blackwell Companion to Consciousness*, Blackwell, 225-235.

Chella, A. (2007) Towards Robot Conscious Perception, in A. Chella & R. Manzotti (eds.), *Artificial Consciousness*, Imprint Academic, 124-140.

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.

Coward, L, Gedeon, R. (2009) Implications of Resource Limitations for a Conscious Machine, *Neurocomputing*, 72, 767-788.

Crick, F., Koch, C. (1990) Towards a Neurobiological Theory of Consciousness, *Seminars in the Neurosciences*, 2, 263-275.

Dehaene, S., Changeux, J.-P. (1997). A Hierarchical Neuronal Network for Planning Behavior, Proc. *National Academy of Sciences USA*, 94, 13293-8.

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.

Dennett, D. (1996) Facing Backwards on the Problem of Consciousness, *Journal of Consciousness Studies*, 3, 4-6.

Franklin, S. (1995) *Artificial Minds*, MIT Press.

Gamez, D. (2010) Information Integration Based Predictions about the Conscious States of a Spiking Neural Network, *Consciousness and Cognition*, 19, 294-310.

Garrett, B. (2006) What the History of Vitalism Teaches us about Consciousness and the "Hard Problem", *Philosophy and Phenomenological Research*, 72, 616-628.

Gentili RJ, Oh H, et al. (2011) Neural Network Models for Reaching and Dexterous Manipulation. In V. Cutsuridis, Hussain A., Taylor, J. (Eds.). *Perception-Action Cycle: Models*, Springer, New York, 187-217.

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.

Jaynes, J. (1976) *The Origin of Consciousness in the Breakdown of the Bicameral Mind*, Houghton Mifflin.

Katz, B. (2013) An Embarrassment of Theories, *Journal of Consciousness Studies*, 20, 43-69.

Kitamura, T., Tahara, T., & Asami, K. (2000) How Can a Robot Have Consciousness? *Advanced Robotics*, 14, 263-275.

Koch, C., Tononi, G. (2008) Can Machines be Conscious, *IEEE Spectrum*, June, 55-59.

Kuipers, B. (2005) Consciousness: Drinking from the Firehose of Experience, *Proc. 20th National Conference on Artificial Intelligence*, AI Press, 1298-1305.

Levine, J. (1983) Materialism and Qualia: The Explanatory Gap, *Pacific Philosophical Quarterly*, 64, 354-361.

Manzotti, R. (2012). The Computational Stance is Unfit for Consciousness, *International Journal of Machine Consciousness*, 4, 401-420.

Marcus, G. (2001) *The Algebraic Mind: Integrating Connectionism and Cognitive Science*, Cambridge, MA: MIT Press.

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.

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. (2012a) A Generalized LSTM-like Training Algorithm for Second Order Recurrent Neural Networks, *Neural Networks*, 25, 70-83.

Monner D, Reggia J. (2012b) Neural Architectures for Learning to Answer Questions, *Biologically Inspired Cognitive Architectures,* 2, 37-53.

Monner D, Reggia J. (2014) Emergent Latent Symbol Systems in Recurrent Neural Networks, *Connection Science*, in press.

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.

Pomerleau, D. (1993) Knowledge-based Training of Artificial Neural Networks for Autonomous Robot Driving. In J. Connell, S. Mahadevan (Eds.), *Robot Learning*, Kluwer, 19-43.

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.

Regis, E. (2008) *What is Life?,* Farber, Strauss and Giroux (New York).

Rolls, E. (2007) A Computational Neuroscience Approach to Consciousness, *Neural Networks*, 20, 962-982.

Rougier, N., Noelle, D., Braver, T., Cohen, J., O'Reilly, R. (2005) Prefrontal Cortex and Flexible Cognitive Control. *Proc. National Academy of Sciences USA*, 102, 7338-43.

Roy, A. (2008). Connectionism, Controllers, and a Brain Theory, *IEEE Transactions on Systems, Man and Cybernetics,* Part A, 38, 1434-1441.

Roy, D. (2005) Grounding Words in Perception and Action: Computational Insights, *Trends in Cognitive Sciences*, 9, 389-396.

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.

Seth, A. (2009) The Strength of Weak Artificial Consciousness, *International Journal of Machine Consciousness*, 1, 71-82.

Simen, P., Vugt, M., Balci, F., Freestone, D., Polk, T. (2010) Toward an Analog Neural Substrate for Production Systems, *Proceedings 10th International Conference on Cognitive Modeling*, 223-228.

Starzyk, J., Prasad, D. (2011) A Computational Model of Machine Consciousness, *International Journal of Machine Consciousness*, 3, 255-281.

Sun, R. (1999) Accounting for the Computational Basis of Consciousness, *Consciousness and Cognition*, 8, 529-565.

Sun, R. (2002) *Duality of the Mind*, Erlbaum.

Sylvester, J., Reggia, J., Weems, S., Bunting, M. (2010) A Temporally Asymmetric Hebbian Network for Sequential Working Memory. In D. Salvucci, G. Gunzelmann (Eds.), *Proceedings 10th International Conference on Cognitive Modeling*, 241-246.

Sylvester J, Reggia J, Weems S, Bunting M (2013) Controlling Working Memory with Learned Instructions, *Neural Networks*, 41, 23-38.

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

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.

van der Velde, F. (2013) Consciousness as a Process of Queries and Answers in Architectures Based on in situ Representations, *Intl. Journal of Machine Consciousness*, 5, 27-45.
Weems, S., Reggia, J. (2006) Simulating Single Word Processing in the Classic Aphasia Syndromes, *Brain and Language*, 98, 291-309.
Winder, R., Reggia, J., Weems, S., & Bunting, M. (2009) An Oscillatory Hebbian Network Model of Short-Term Memory, *Neural Computation*, 21, 741-761.
Wolfram, S. (2002) *A New Kind of Science*, Wolfram Media.