Computer Modeling in Philosophy

James A. Reggia*, Garrett E. Katz and Gregory P. Davis

Modeling Working Memory To Identify Computational Correlates of Consciousness

https://doi.org/10.1515/opphil-2019-0022
Received April 06, 2019; accepted August 09, 2019

Abstract: Recent advances in philosophical thinking about consciousness, such as cognitive phenomenology and mereological analysis, provide a framework that facilitates using computational models to explore issues surrounding the nature of consciousness. Here we suggest that, in particular, studying the computational mechanisms of working memory and its cognitive control is highly likely to identify computational correlates of consciousness and thereby lead to a deeper understanding of the nature of consciousness. We describe our recent computational models of human working memory and propose that three computational correlates of consciousness follow from the results of this work: itinerant attractor sequences, top-down gating, and very fast weight changes. Our current investigation is focused on evaluating whether these three correlates are sufficient to create more complex working memory models that encompass compositionality and basic causal inference. We conclude that computational models of working memory are likely to be a fruitful approach to advancing our understanding of consciousness in general and in determining the long-term potential for development of an artificial consciousness specifically.

Keywords: computational correlates, computational explanatory gap, cognitive control, working memory, cognitive phenomenology, mereology, machine consciousness, artificial consciousness, mind-brain problem

1 The computational investigation of consciousness

It is easy to dismiss most work done in artificial intelligence (AI) and computational modeling in general as irrelevant to the mind-body problem and, more specifically, to the study of consciousness or to assessing the prospects for ultimately creating a conscious machine. This view appears to be fairly widespread in the philosophy literature. It includes arguments that the subjective nature of consciousness (the "hard problem") precludes study via objective scientific methods[1], and that the lack of a formal definition of consciousness makes a conscious machine unobtainable[2]. These views have resonated with many investigators in AI. For example, a survey of AI researchers found that many of them agreed that the problem was too ill-defined to be of interest in AI[3].

The term "consciousness" has indeed proven difficult to define. One way to approach this issue is simply to define consciousness to be the subjective experiences and awareness that a person has when

1 McGinn, "Can We Solve the Mind-Body Problem?" 357.
2 Bringsjord, "One Billion Dollars for a Conscious Robot", 32.
3 McDermott, "Artificial Intelligence and Consciousness", 120.

***Corresponding author: James A. Reggia,** University of Maryland, Maryland, United States of America;
Email: reggia@cs.umd.edu
Garrett E. Katz, Syracuse University, Syracuse, New York, United States of America; Email: gkatz01@syr.edu
Gregory P. Davis, University of Maryland, Maryland, United States of America; Email: vicariousgreg@gmail.com

awake, consistent with past suggestions that a more precise definition should be deferred, given our inadequate scientific understanding of the nature of consciousness.[4] Many philosophers use the term *phenomenal consciousness* to refer to this notion of consciousness, emphasizing that one is referring specifically to the phenomena of subjective experience (qualia; or "something it is like"[5]). This can be contrasted with the term *access consciousness* that refers to the availability of information for conscious processing. Access consciousness has been defined in different ways. Here we will take a perceptual state to be access conscious "if its content gets to the Executive System, whereby it can be used to control reasoning and behavior."[6] Our work described in this paper is most closely related to this functional concept of access consciousness.

Our primary concern in the following is with identifying computational aspects of high-level cognition that are associated with consciousness. We are more optimistic about the eventual prospects of such an approach than some others. Part of this optimism comes from recent developments in contemporary philosophical thinking that we believe imply a more positive viewpoint about using computational models to investigate the nature of consciousness, and perhaps even its eventual instantiation in machines. Two of these developments, cognitive phenomenology and mereological analysis, are especially encouraging in this regard to those of us in computer science, as follows.

The first development, *cognitive phenomenology*, asserts that our conscious experiences extend beyond traditional sensory qualia to include deliberative thought and high-level cognitive processes.[7] This assertion has proven to be controversial among philosophers. While philosophers generally agree that some aspects of cognition are accessible to consciousness, there is substantial disagreement beyond that point. Specifically, a number of philosophers argue that *all* phenomenology is fundamentally sensory, including that associated with cognitive states.[8] In contrast, advocates of cognitive phenomenology argue that there are additional phenomenal aspects of cognition that cannot be accounted for by traditional sensory qualia and mental imagery. For example, an advocate of cognitive phenomenology might assert that abstract thoughts provide an example of this, or that there are different non-sensory subjective qualities associated with hearing a sentence in a foreign language when one understands that language versus when one does not.

While there are those who do not agree with the concept of cognitive phenomenology, we find the arguments of advocates to be compelling. Here we will simply assume that cognitive phenomenology exists and ask what that might imply. In other words, we assume that there are distinct non-sensory subjective mental experiences associated with at least some aspects of cognition, and ask what this might signify about the value of computational modeling as an investigative tool concerning the nature of consciousness. Our answer is that cognitive phenomenology makes computational studies based on modeling cognitive processes potentially much more relevant to studying consciousness. Most past modeling work in AI and in cognitive science more generally involves computational mechanisms that do not in any meaningful way have associated sensory representations. This makes such models irrelevant to understanding important aspects of consciousness if one assumes a priori that all phenomenology is sensory based. For example, computational models that use abstract symbols that are not grounded in the environment in any way would, in the absence of any sensory representation, be viewed as irrelevant to consciousness studies. From our perspective in computer science, assuming the existence of non-sensory cognitive phenomenology changes the possibilities: It implies that computational states in such models that are lacking in sensory representations might still be associated with conscious cognitive states, and thus it greatly broadens the range of computational mechanisms that might reasonably be found to correlate with consciousness. In other words, cognitive phenomenology substantially expands the potential for mechanistic computational

4 Crick, "The Astonishing Hypothesis", 20.
5 Nagel, "What is it like to be a bat?", 435.
6 Block, "On a Confusion about a Function of Consciousness", 230.
7 Bayne and Montague, "Cognitive Phenomenology", 12; Chudnoff, "Cognitive Phenomenology"; Jorba and Vincente, "Cognitive Phenomenology", 74.
8 Prinz, "The Conscious Brain", 149; Carruthers, "The Centered Mind", 15.

models of cognitive processes to provide insight into computational correlates of consciousness, a concept that we describe in Section 2.[9]

A second development in philosophy that strikes us as relevant to computational studies of consciousness is the assertion that phenomenal consciousness can be approached effectively via *mereology*. Mereology focuses on formally studying the part-whole relations of a system. Prentner has recently argued within the framework of process metaphysics that re-conceiving phenomenal consciousness in the context of mereology could be an effective way to elucidate its nature.[10] In other words, rather than viewing consciousness as composed of qualia that are non-structured properties of subjective experience, consciousness and its constitution should be understood in terms of a mereological analysis of internally-structured processes. This mereological approach provides a potential bridge between philosophical issues surrounding the hard problem and methodologies already used in AI cognitive models, and more generally suggests to us that investigating the structure of cognitive processes might lead to useful insights about consciousness.

Philosophical perspectives such as cognitive phenomenology and mereology open the door to much more widespread consideration of computational methods for investigating the fundamental nature of consciousness and the mind-brain problem.[11] In Section 2, we first briefly explain the concept of computational correlates of consciousness in general terms and review some past related work that has been done searching for them. We also identify three practical barriers that make identifying computational correlates very challenging. These barriers include a clearly identifiable "computational explanatory gap" that can productively serve as the focus of computational investigations. In Section 3 we address the question of which aspects of cognition might be most productive to examine in searching for computational correlates of consciousness. We argue that investigating the mechanisms that underlie working memory, a part of the human short-term memory system that is widely accepted as involving conscious aspects of human cognition, should provide an especially fertile subject for investigation in this context (Section 3.1). To support our argument, we then describe two examples of our recent work implementing computational models of working memory that begin to investigate this question. The first of these models deals with a simple card matching task and is implemented using neural computational methods (Section 3.2). Based on our work with this model we conclude that three specific computational correlates of consciousness could be identified, and we briefly describe each of these. Our second model examines more complex compositional and inference aspects of working memory and is implemented using more traditional symbolic AI methods (Section 3.3). We are examining whether, when this latter model is converted to purely neurocomputational form, using the three computational correlates of conscious that we identified will be sufficient to support the more advanced working memory mechanisms that are involved. In Section 4, we summarize our results and their implications.

2 Computational correlates of consciousness

Given the lack of a generally-accepted formal definition of consciousness and the subjective nature of consciousness, it is especially important to be clear as to what the goal is in any effort to study consciousness computationally. In the work described here, we are not trying to create a phenomenally conscious machine. Instead we are focused on a more tractable issue: exploring whether there exist identifiable computational aspects of high-level cognitive processes that are associated with the presence of consciousness. According to cognitive phenomenology, such underlying computational mechanisms may be directly relevant to elucidating the nature of consciousness.

9 However, we note that whether phenomenology is solely sensory or not does not impact our main purpose and results in this paper. Even if consciousness is purely sensory-based, identifying computational correlates of consciousness remains relevant (e.g., identifying the contribution of non-conscious cognitive processes to conscious sensory phenomena).

10 Prentner, "Process Metaphysics of Consciousness", 9.

11 Chella & Manzotti, "Machine Consciousness: A Manifesto for Robotics", 14.

We take a *computational correlate of consciousness* to be any aspect of information processing that is associated with conscious cognitive activities but not with unconscious cognitive processes[12]. This could include, in theory, both representations and processing mechanisms based on the symbolic methods of traditional AI. For example, variable binding is an aspect of information processing in consciously-reportable reasoning that is performed in symbolic AI systems. However, our interest here is primarily on correlates based on neurocomputational modeling. Specifically, we are interested in *neurocomputational correlates of consciousness*, the representation and processing of information in neural networks, because they address lower-level, more fundamental (in our opinion) mechanisms than symbolic AI, and they can be directly compared to contemporary neuroscience knowledge. Neurocomputational correlates can be related to but in general differ from the more widely known neural correlates of consciousness involving biological systems, the latter of which include biochemical phenomena, neuroanatomical structures, patterns of brain electrical/metabolic activity, and other inherently biological phenomena related to the brain[13]. Neurocomputational correlates are more abstract and are independent of the physical hardware on which they occur, be it the biological brain, silicon-based electronic circuitry, or bio-molecular systems. We emphasize that in all of this, the word "correlates" neither implies nor precludes causality.

A substantial number of past computer modeling studies have been done that can be viewed as exploring various aspects/implications of potential neurocomputational correlates of consciousness[14]. Some examples illustrate this point, as follows. Viewing recurrent neural networks as mathematical dynamical systems, it has been proposed that the *activity attractor states*[15] of these networks characterize conscious states[16]. Another suggested correlate is widespread activity over multi-region neural networks that form a *global workspace*[17], something that is consistent with functional imaging studies of the human brain during conscious versus unconscious tasks. It has also been hypothesized that having a *self-model* embedded in a robot's internal model of its environment could be the basis of a conscious robot. This has been investigated using a recurrent neural network control system in a physical robot that, impressively, was able to pass the mirror test used by ethologists to assess self-awareness in animals[18]. Other previously hypothesized examples of potential computational correlates include *higher-order neural networks* that can represent the information in other lower-order neural networks and are related to higher-order thought theory in philosophy[19], the reportable *collective processing of shared information*[20], and various other aspects of top-down *attention mechanisms*[21]. These potential correlates are not mutually exclusive, and no doubt additional ones will be proposed and studied computationally over the next several years. The hope is that identification of a sufficient set of such correlates could ultimately lead to a better understanding of the structure and functionality of the conscious mind.

In practice, identifying neurocomputational correlates of high-level cognition can be very challenging for at least three reasons. First, it is often not clear precisely which aspects of cognition are conscious and which are not. It is widely recognized that substantial portions of cognition occur at a sub-conscious level. In practice, experimental cognitive psychologists have often taken subjects to be conscious of an event if they can verbally report that event's occurrence. Such a criterion, however imperfect, leads one to characterize unconscious cognitive processes as involving relatively fast parallel processing where multiple

12 Cleeremans, "Computational Correlates of Consciousness", 1032.
13 Chalmers, "What is a Neural Correlate of Consciousness?" 17; Metzinger, "Neural Correlates of Consciousness".
14 Reggia, "The Rise of Machine Consciousness", 116-127.
15 The term "attractor" here usually refers to fixed-point attractors, but could also refer to other types of attractors, such as limit cycles and chaotic attractors.
16 Fekete and Edelman, "Towards a Computational Theory of Experience", 815; Taylor, "Neural Networks for Consciousness", 1209.
17 Baars, "A Cognitive Theory of Consciousness", 86; Connor and Shanahan, "A Computational Model of a Global Neuronal Workspace", 1140.
18 Takeno, "Creation of a Conscious Robot", 203.
19 Cleeremans et al., "Consciousness and Metarepresentation", 1034.
20 Haikonen, "Consciousness and Robot Sentience", 187.
21 Taylor, "CODAM: A Neural Network Model of Consciousness", 987; Perlis and Brody, "Operationalizing Consciousness", 4.

tasks can occur simultaneously with only limited interference between them, while in contrast, conscious cognitive processes are characterized as being slow and serial, and attempting to simultaneously carry out multiple tasks that require conscious involvement often leads to errors because the tasks interfere with each other[22]. However, the limited ability of investigators in psychology to definitively discriminate between conscious versus unconscious cognitive processes more generally means that, in searching for computational correlates of consciousness, one needs to focus on aspects of cognition where there is relatively clear and widely accepted agreement on this issue.

A second practical challenge is that there is currently only very limited understanding in AI of how high-level cognition (logical reasoning, goal-directed problem solving, planning, metacognition, etc.) that, at least in part, is widely accepted as involving conscious mental activity, can be instantiated as low-level neural computations (artificial neural networks). We have referred to this previously as the *computational explanatory gap*[23]. The computational explanatory gap is an abstraction, a purely computational issue, concerning how the symbolic-level algorithms occurring with high-level cognition can be mapped into the distributed representation and parallel computations occurring in "low-level" neural networks. As such it is distinct from the traditional explanatory gap in philosophy associated with the "hard" mind-brain problem. From our viewpoint in AI, and in contrast to many in philosophy who would view this as part of the "easy" problem, the computational explanatory gap is somewhat mysterious and fundamental because attempts to solve it over more than half a century have found it to be largely intractable, this in spite of the fact that the human brain provides a proof that a solution exists. To our knowledge past philosophical work has not provided deep insight into why this intractability exists.

The third practical challenge to identifying computational correlates of consciousness within the scope of cognitive phenomenology is more technical. Specifically, one must not only establish that a computational mechanism is associated with some aspect of conscious mental activities, but also that it is *not* associated with other non-conscious mental information processing. For example, one might question whether a neural activity attractor state in general is a computational correlate of consciousness since it is easy to imagine neural networks with attractor states that appear to be associated with non-conscious functionality, such as central pattern generators for motor control in the spinal cord. In this case it is necessary to go deeper and clarify what specific types of activity attractor states would qualify as neurocomputational correlates of consciousness, something that is an open question to our knowledge.

3 Modeling working memory and its cognitive control

If one accepts that identifying computational correlates of consciousness could lead to a deeper understanding of consciousness, the immediate question becomes: What aspects of cognition would be most fruitful to examine in searching for such correlates? Our answer is that studying the computational representations and processes that support, interact with, and control working memory are especially likely to be productive, for the reasons that we give below.

Cognitive psychologists have long viewed human memory as partitioned into a variety of systems based on behavioral and neuroscientific data[24]. For example, long-term memory systems include both declarative semantic and episodic memory, and non-declarative procedural memory. Here our interest is instead in the short-term memory system referred to as *working memory*. We propose that working memory mechanisms form an ideal context for exploring ideas concerning computational correlates of consciousness. In particular, based on our recent work modeling working memory, we hypothesize that three important

22 Baars, "A Cognitive Theory of Consciousness", 74; Dehaene and Naccache, "Towards a Cognitive Neuroscience of Consciousness", 5; These distinctions relate to and reinforce the reality of the computational explanatory gap discussed in the following sections in the sense that the properties of unconscious cognitive processes, such as fast parallel processing, are a great match to what occurs in artificial neural networks, while properties of conscious cognition are not.

23 Reggia et al., "The Computational Explanatory Gap", 158.

24 Squire and Dede, "Conscious and Unconscious Memory Systems", 3.

computational correlates of consciousness can be identified: itinerant attractor sequences, top-down gating, and very fast Hebbian weight changes during learning. We elaborate on each of these three ideas as we discuss our computational models of working memory in subsequent sections.

More specifically, in the following we will describe two recent computer models that we have been studying that incorporate simulations of working memory and its control via cognitive processes, and relate these studies to the three hypothesized correlates listed above. The first model uses a neurocomputational framework and is trained to solve simple card matching problems. Our second modeling effort is based on a more traditional AI symbol-processing framework that controls the behavior of a physical robot. It is more complex but of special interest in that it incorporates composition and inference involving working memory. We are using this second model to determine whether the three hypothesized computational correlates of consciousness are sufficiently powerful to incorporate compositional and inference aspects of working memory when rendered in a purely neurocomputational framework. It is these models that led us to the three computational correlates stated above. Before describing these models, we first briefly consider the nature of working memory and its relationship to consciousness.

3.1 Human working memory and consciousness

Human working memory is the memory system that transiently stores and manipulates information over a short time period[25]. For example, suppose someone were to ask you verbally to "Subtract 196 from 425 and tell us what you get without writing anything down." In doing this you would have to retain in working memory the numbers involved, and manipulate this information in various ways ("Let's see, I need to borrow a 1 from the 10's column, ..."). Working memory is characterized by retention of information over a period of seconds to minutes. If you were consecutively solving multiple arithmetic problems like the above, the information about each problem is quickly cleared from working memory in a controlled fashion and replaced by new problem-specific information as you work on each problem. Working memory is characterized by very severe restrictions on the amount of information that it can retain. Experimental studies by psychologists have found that human working memory capacity is approximately four independent items of information under laboratory conditions[26], in marked contrast to the enormous capacity of human long-term memory. Items stored in working memory may be lost because they interfere with each other or because they "decay" over time, reflecting its limited storage capacity. Further, working memory can be compositional: The operations acting on stored information can construct structured representations, such as the three-digit answer to the subtraction problem above. While working memory has historically been most closely associated with prefrontal cortex in the human brain, recent studies present a more nuanced view with widespread involvement of other cortical regions[27].

We argue here that working memory and the cognitive control mechanisms associated with it provide an excellent context in which to search for computational correlates of consciousness. To see this, we next consider how modeling working memory addresses in part the three barriers to identifying neurocomputational correlates of consciousness that we gave in Section 2: discriminating conscious from non-conscious cognitive processes, instantiating high-level cognitive processes as low-level distributed neural processing, and establishing that computational correlates are uniquely involved with conscious cognition.

First, working memory and the operations on it are widely considered by both psychologists and philosophers to involve conscious and reportable cognitive processes.[28] This perspective is supported by

25 Baddeley, "Working Memory and Conscious Awareness", 22.

26 Cowan et al., " On the Capacity of Attention", 53.

27 Lara and Wallis, "The Role of Prefrontal Cortex in Working Memory".

28 Baars and Franklin, "How Conscious Experience and Working Memory Interact", 166; Baddeley, "Working Memory: Theories, Models and Controversies", 15; Block, "Perceptual Consciousness Overflows Cognitive Access", 567; Carruthers, "The Centered Mind", 75; Courtney et al., "The Role of Prefrontal Cortex in Working Memory", 1819; Persuh et al., "Working Memory and Consciousness".

recent neurobiological evidence that the neocortical mechanisms underlying working memory are fairly widespread and not limited to prefrontal cortex[29]. It is also consistent with other evidence that conscious cognitive processes are associated with widespread cortical intercommunication and activity[30], and with the global workspace theory of consciousness mentioned earlier. Thus, according to cognitive phenomenology, working memory and the cognitive control of working memory provide an appropriate context in which to search for computational correlates of consciousness.

Second, instantiating the cognitive processes associated with working memory in distributed neural computations, although challenging, appears to be reasonably feasible. Several neural models of aspects of working memory have been developed and studied in recent years, for example[31]. These models often include simulating presumed cognitive control of working memory via top-down "executive processes" that, in the brain, is believed to be mediated by regions of the prefrontal cortex. The circumscribed nature of working memory, relative to human long-term memory and cognition in general, is advantageous in making such models computationally tractable. Our card-matching model below illustrates this tractability.

Finally, the top-down, goal-directed control of working memory is very different from much of the unconscious sensory processing and low-level motor control that occurs in the brain. This, and its partial dependence on symbol manipulation, makes it highly likely that the neurocomputational mechanisms underlying conscious information processing in working memory will ultimately be found to differ from those associated with unconscious information processing. Our suggestion below is that this top-down goal-driven control process is implemented by gating mechanisms where one or more neural modules as a whole control the functionality of other modules/networks as a whole. While we are focused primarily on working memory, we suspect that gating of other conscious, goal-directed cognitive functions occurs widely in the brain. For example, suppose a person is listening to a series of a few spoken words. If the goal is simply to remember those words, then top-down gating mechanisms would activate their retention in working memory and de-activate speech output modules such as Broca's area. In contrast, if the goal is to repeat aloud the words with no expectation of recalling them later, then Broca's area and other speech output modules would, via gating, be turned on. What is special here is that, in general, the neural modules involved not only exchange information via interconnecting pathways, but they also control one another's actions in a very broad sense. A module may, via gating, turn the activity or connections of other modules on or off, determine when other modules discard or retain their activity state, when they learn or forget information, and when they generate output. There is substantial evidence that top-down gating operations occur in the brain, although the precise mechanisms involved remain unclear[32].

Given this rationale for developing computational modeling of working memory as the basis for identifying computational correlates for consciousness, we now turn to describing our recent investigation of working memory models for two tasks, a type of problem solving involving a card matching task, and imitation learning of simple procedures. Most past work investigating computational correlates of consciousness has used a paradigm in which one starts with a presumed correlate (global information processing, self-modeling, etc.) and then investigates its plausibility and/or implications via computational experiments. In contrast, our approach here can be characterized as starting with a model of conscious information processing (i.e., a model of working memory), and asking what distinct, core neurocomputational mechanisms were needed to instantiate that model. We then hypothesize that these key mechanisms are computational correlates of consciousness.

29 Lara and Wallis, "The Role of Prefrontal Cortex in Working Memory".

30 Massimini et al., "Breakdown of Cortical Effective Connectivity During Sleep", 2231.

31 Pascanu and Jaeger, "A Neurodynamical Model for Working Memory", 201; Sylvester et al., "Controlling Working Memory with Learned Instructions", 25; Verduzco-Flores et al., "Modeling Neuopathologies as Disruption", 21.

32 Frank et al., "Interactions Between Frontal Cortex and Basal Ganglia", 139; Sherman & Guillery, "Exploring the Thalamus and Its Role in Cortical Function", 303; Singer, "Dynamic Formation of Functional Networks by Synchronization", 191.

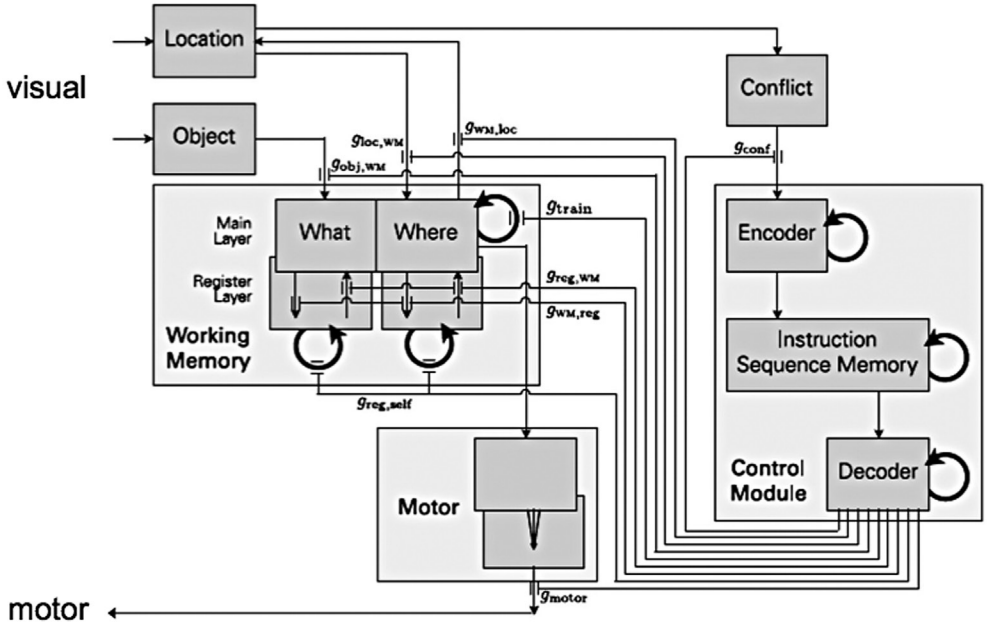

Figure 1: The neurocomputational architecture used for the card matching task. Input to the system (upper left) consists of images of a set of cards on a table. Information about the location and face patterns of these cards is extracted from these images. This visual information is used by the agent to select which card to turn over next (lower left motor output). The rest of the diagram shows the architecture of a neural model that serves as the agent's "brain" that controls its behavior. Each dark grey box represents a neural region, with arrows indicating pathways connecting the regions. Key to our discussion here is that the agent has a working memory (lightly shaded larger box in the center left of the illustration) where it retains the identities and locations of a few recent cards that it has seen (subject to decay), and an executive control module (lightly shaded box on the right) that directs the agent's working memory and output motor actions via gating. Further details are given in the text.

3.2 A computer model of working memory during a card matching task

Our first investigation implemented and studied a multi-component neurocomputational model of working memory and the cortical executive processes that control it[33]. The model includes computational mechanisms allowing one module to gate the actions of other modules, determining when environmental patterns are stored or removed in working memory, when learning occurs, and when to update the state of the model in general. The model has been applied to simulate human behavior in both an n-back task (a standard test of working memory used by psychologists) and in a card-matching task. We consider just the latter, more challenging card-matching task here.

Figure 1 shows the components and structure of the card-matching task's neural network. The neurocognitive agent controlled by the neural circuitry shown in Figure 1 can see the current configuration of a set of cards on a table (visual input image at the upper left) and selects specific cards on the table to be turned over by pointing at them (motor output at lower left). The agent's goal is to remove all of the cards from the table in as few steps as possible. At each time step, the agent can select two face-down cards to turn over, revealing the identity (the patterns on their faces) of those two cards. If the two cards match, then they are removed and progress is made towards the goal of clearing the table. If the selected cards reveal different patterns then they are turned back over and remain face down on the table. For this latter case, the agent's working memory records which patterns have just been seen on these two cards and their locations. Continuing this example, if the agent next turns over a different card and discovers that it has a pattern on its face matching that on one of the first two cards, then based on its working memory contents the agent would also select the (currently face down) appropriate previous card to get a matching pair that is removed from the table.

33 Sylvester et al., "Engineering Neural Systems", 39.

Many of the details of the neural architecture shown in Figure 1 are described elsewhere[34], and are not germane to our discussion. Two aspects of the model are however particularly important here. First, the *working memory* (on the left in Figure 1) is a recurrent neural network that stores information about which cards have been observed previously by learning, at appropriate times, the location and identity of those cards. This allows the system subsequently to choose pairs of cards based on its past experience, much as a person does. Working memory is implemented as an auto-associative neural network that uses one-step Hebbian learning. It stores object-location pairs as attractor states, activity states to which the memory network will evolve over time. This allows the system to retrieve complete pairs when given just the object information alone or just the location information alone as needed during problem solving. This functionality can be viewed as a solution to the binding problem[35]. Learned information in working memory decays with time, allowing stored information to be displaced as new information arrives.

The second relevant point about our model is that the control module's *instruction sequence memory* (on the right in Figure 1) is "programmed" a priori via temporally-asymmetric one step Hebbian learning so that it has learned temporal sequences of attractor states. Each of these states represents an "instruction" that is part of an algorithm that performs the card matching task. The recurrent network used in the instruction sequence memory thus implements a procedural memory that is capable of simultaneously storing multiple instruction sequences that are used to perform card matching tasks. Each instruction indicates to the control module which gates should be opened at that point in time during problem solving. In other words, the instruction sequence memory allows the system to learn simple "programs" (procedures) for what actions to take in situations where there are zero, one, or two cards face up. The instruction sequence memory uses Hebbian learning to both store individual instructions as attractor states of the network, and to transition between these attractors/instructions during problem solving.

The model's instruction sequence memory serves as an "executive system" for working memory. It is inspired by current knowledge of biological prefrontal cortex functionality, and directs and controls the actions of the rest of the system. This executive control module is initially trained to carry out the card removal task and acts via nine top-down gating connections to control the sequence of operations that are performed by the agent. Its outgoing gating connections (bottom right in Figure 1; these connections are pictured ending with -|| to suggest their valve-like gating functions) act on the various operational components of the system to control when information can flow over pathways and when information is to be learned/deleted by working memory. For example, the rightmost gating connection's activity g_{motor} turns on/off the output from the motor module, determining when and where the agent points at a card to indicate that it should be turned over. As other examples, the activation $g_{obj,WM}$ and $g_{loc,WM}$ of two other gating connections determine when a seen card's identity and/or location are stored in working memory. A mathematical description of how these gating connections work in the model can be found elsewhere[36]. Their functionality is inspired by what neuroscientists refer to as multiplicative modulation in the brain[37].

When given an appropriate set of parameter values, our model exhibited accuracy and timing results reminiscent of those we observed experimentally in humans performing similar card matching tasks. For example, it successfully solved every one of hundreds of randomly generated tasks on which it was tested, and the number of rounds the model required to solve card matching problems as the number of cards involved varied was qualitatively similar to what we observed in having human subjects carry out the same task. This supports the idea that our model captures some important aspects of human control of working memory.

---

34 Sylvester and Reggia, "Engineering Neural Systems for High-Level Problem Solving".

35 Feldman, "The Neural Binding Problem".

36 Sylvester and Reggia, "Engineering Neural Systems for High-Level Problem Solving", 42.

37 Akam and Kullmann, "Oscillatory Multiplexing of Population Codes for Selective Communication in the Mammalian Brain", 111.

Given the above results and the tenets of cognitive phenomenology, the key question becomes: What core computational mechanisms were required to implement the card matching task that might therefore be hypothesized to be computational correlates of consciousness? In answering this, we considered representation, processing, control and learning mechanisms. Our answer is that there are three critical neurocomputational mechanisms that were essential to make our card matching model function effectively and that we consider to be potential computational correlates of consciousness. They are

i) *itinerant attractor sequences* representing learned cognitive states in working memory,
ii) *top-down gating* mechanisms associated with the cognitive control of working memory, and
iii) *very fast weight changes* that support immediate learning/unlearning in working memory.

We now discuss each of these in turn.

The first proposed computational correlate of consciousness is *itinerant attractor sequences* representing learned cognitive states in working memory. Each learned cognitive state in such a sequence is an attractor of the underlying recurrent neural network that drives the instruction sequence memory as it controls the overall system's functioning during card matching or other tasks. It has previously been hypothesized that an activity-space trajectory might serve as a computational correlate of consciousness[38], but here we are specifically emphasizing the fact that the trajectory is composed of an attractor sequence, that it is specific to working memory control, that it involves *learned* rather than built-in states and control mechanisms (in other words, it represents "software" of the system rather than pre-wired circuitry or "hardware" that is genetically pre-determined), and that it specifically involves cognitive states in higher level reasoning and problem solving.

The second potential computational correlate of consciousness in our model is the use of *top-down gating* of working memory that controls what is learned and manipulated by working memory. Arguably these gating operations (see discussion of Figure 1 above), driven by the sequences of attractor states in the executive control module, correspond to consciously reportable cognitive activities involving working memory, and thus they are a potential computational correlate of consciousness. Such intimate control of working memory learning and manipulation conveys a sense of ownership or agency to a system and thus forms an important part of conscious awareness. It also relates to the concept of mental causation discussed in the philosophy of mind literature concerning free will.

The third potential computational correlate suggested by our model is the use of *very fast weight changes* that support immediate learning/unlearning in working memory. Working memory is able to reliably store a new piece of information immediately upon a single presentation of that information. Our modeling work implements such fast weight changes, or "one step learning", that supports immediate learning using Hebbian weight change rules to acquire both information about the environment and temporal behavioral sequences. These two types of information are learned/stored in working memory using temporally symmetric and asymmetric versions of Hebbian learning, respectively.

3.3 Modeling working memory during imitation learning

The card-matching model described in the previous section provides an answer to the question of how one can use neurocomputational methods to implement (at least) a circumscribed part of working memory and its cognitive control. In doing so, we argued that it suggests the three potential computational correlates of consciousness characterized in the preceding section. However, the card-matching model is a very limited simulation of working memory and its control. For example, it does not support the compositional aspects of working memory, nor the ability to manipulate and more generally to reason with the structured contents of working memory. Even more troublesome, while it provides for very fast weight changes that learn information from a single stimulus presentation, it does not adequately allow for the very fast and

38 Fekete and Edelman, "Towards a Computational Theory of Experience", 815.

definitive controlled erasure of items that are no longer needed in working memory. While such aspects of working memory are readily implemented using traditional symbol-processing methods in AI, they are largely beyond the practical capabilities of the card matching model and, more generally, existing neurocomputational methods.

To address these issues, we are currently investigating whether the same three principles that we have postulated are computational correlates of consciousness are sufficient to implement these much more powerful cognitive processes associated with working memory. Our ongoing research program involves the following steps:

1. select an existing *target cognitive system* that uses symbolic AI methods which support a compositional working memory and the ability to make inferences based on the contents of that working memory;
2. develop a practical framework, a *neural virtual machine*, capable of implementing universal computation in purely neurocomputational systems;[39] and
3. use the neural virtual machine to re-implement the symbolic AI target cognitive system as a *compositional working memory* based on purely neurocomputational methods.

Our fundamental research hypothesis that is being challenged in this work is that the same three mechanisms that we are taking to be computational correlates of consciousness will prove sufficient to create this more advanced and complex neurocognitive model, supporting our claim that they are important computational correlates of consciousness, according to the principles of cognitive phenomenology and mereology. We now briefly describe each of the three steps in our research program.

The first step is to identify a state-of-the-art *target cognitive system* that, implemented using the methods of symbolic AI, incorporates a compositional working memory. For this we have selected a recently developed cognitive robotic system named CERIL[40] that learns from human demonstrations (imitation learning)[41]. CERIL was initially entirely based on conventional AI programming methods and software (no neural networks), but we recently replaced its low-level sensorimotor components with neurocomputational control methods, making it a hybrid system. Its cognitive-level components for problem solving and learning remain expressed solely as algorithms that are within the rubric of top-down symbolic AI. CERIL learns to perform bimanual procedures based on representing the demonstrator's intentions in its working memory, rather than on trying to replicate the observed actions verbatim. The robot's high-level reasoning during imitation learning is based on a knowledge base of cause-effect relations. As illustrated in Figure 2, our bi-manual Baxter robot learns and represents in working memory a plausible high-level *explanation* ("Intentions") of the demonstrator's goals in performing the specific low-level actions that are actually observed by the robot (this is abductive rather than deductive reasoning). The constructed explanation allows the robot to use the same cause-effect relations to generate its own, possibly different hierarchical plan for carrying out the procedure in similar situations. Further, it allows the robot to learn and often to subsequently generalize to new similar situations from just a single demonstration, much as a person does in learning from imitating others. It also enables the robot to explain its actions to a human based on its inferred high level intentions and their causal relations. We did a theoretical analysis of CERIL's learning algorithms, including establishing their soundness, completeness and complexity, and measured their effectiveness across multiple applications via a systematic experimental evaluation. A critically important result is that, having learned a high-level representation of the demonstrator's actions from a single demonstration, CERIL is often capable of performing the learned task in the physical world. It uses its own plan, inspired by but modified from that of the human demonstrator, to successfully generalize from a single demonstration. The critical point for us here is that CERIL makes use of a structured *compositional working memory*: In learning to imitate the demonstrator's behavior, CERIL constructs and maintains in its

39 While universal computation has been implemented in some previous neural systems, these past implementations have not been based on the three computational correlates of consciousness described above.

40 CERIL is an acronym for Cause-Effect Reasoning for Imitation Learning.

41 Katz, Huang, et al., "A Novel Parsimonious Cause-Effect Reasoning Algorithm for Robot Recognition", 177.

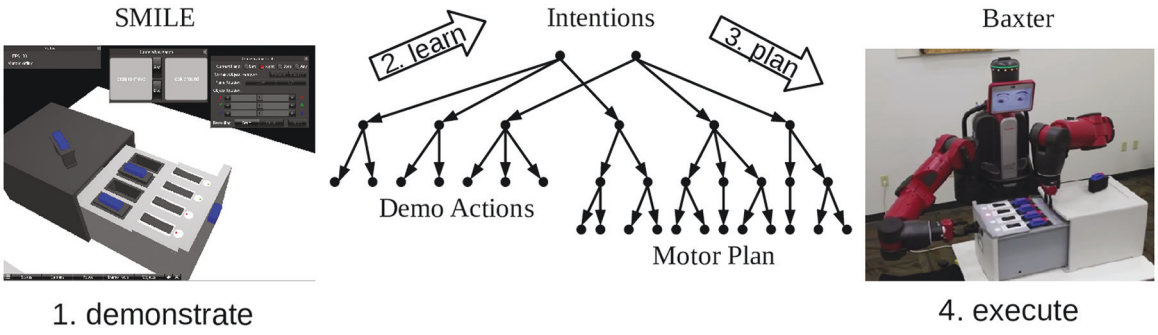

Figure 2: A sketch of CERIL's approach to imitation learning. **1.** On the left, a person demonstrates a procedure within an artificial physics-enabled world named SMILE on a typical desktop/laptop computer. Here a disk drive dock is used where a faulty disk is indicated by a red LED turning on. The demonstrator shows a sequence of steps: open the cabinet, flip the toggle switch, remove the faulty disk, discard it, pick up the new disk, etc. that restore the disk drive dock to its normal operational state. **2.** A robot controlled by CERIL then uses its cause-effect relational knowledge to construct in working memory a hierarchical explanation/interpretation of the low-level actions observed in the demonstration. The high-level sequence ("Intentions" in the middle panel) learned from a single demonstration represents the goals of the demonstrator, not how to copy the demonstrator's actions verbatim. **3.** Subsequently, when presented with a mock-up disk drive dock in the real world, the robot matches its internal models against objects on the table and then constructs a hierarchical plan that, when performed by the robot, will achieve the same goals as those of the demonstrator. **4.** The executed plan (on the right) typically differs from that of the human demonstrator. For example, the disk to be replaced may be in a different slot, the new disk in a different location, and the robot may need to transfer a disk from one gripper to another to facilitate its insertion in the slot. The robot learns both to carry out the demonstrator's intentions rather than copy the observed actions, and to successfully generalize from a single demonstration in this context.

working memory a hierarchical representation of its explanation and a hierarchical representation of its planned actions, and makes inferences involving both of these hierarchies. CERIL's hierarchical plan can be viewed abstractly as a part-whole relationship network, making the ideas of mereology relevant to its understanding and analysis.

CERIL's existing symbolic AI algorithms for representing and controlling working memory, cognitive-level cause-effect reasoning, planning and learning provide a concrete, circumscribed target functionality for a purely neurocomputational implementation of modeled human cognitive processes. Further, the current symbolic AI implementation of CERIL, along with human experimental work that we are currently conducting, provides a standard against which our neurocomputational results can be compared.

The second step in our research program is to develop a practical framework, a *neural virtual machine* (NVM), that is capable of implementing universal computation in systems of neural networks. The NVM, which has been implemented over the last two years[42], provides a purely neurocomputational framework for instantiating cognitive-level algorithms that are currently readily implemented via more traditional symbolic AI methods, but much less so via existing neural network methods. It provides many of the tools needed to "program" the behavior of neural systems in much the same way that conventional computers can be programmed. It encompasses and qualitatively extends the methods used in our card-matching model. For example, unlike that previous application-specific model, the NVM is completely general purpose in nature. Further, it introduces a novel fast store-erase synaptic learning rule that, in a single time step, both stores and erases associations simultaneously using Hebbian and anti-Hebbian weight changes. This allows the possibility that top-down working memory control mechanisms can quickly remove information from working memory. Thus, in theory, the NVM should be able to implement the same working memory and reasoning algorithms that are currently readily done using traditional symbolic AI methods, such as those in CERIL, but now in a solely neurocomputational framework. Use of the NVM is outlined in Figure 3.

42 Katz et al., "A Programmable Neural Virtual Machine", 10.

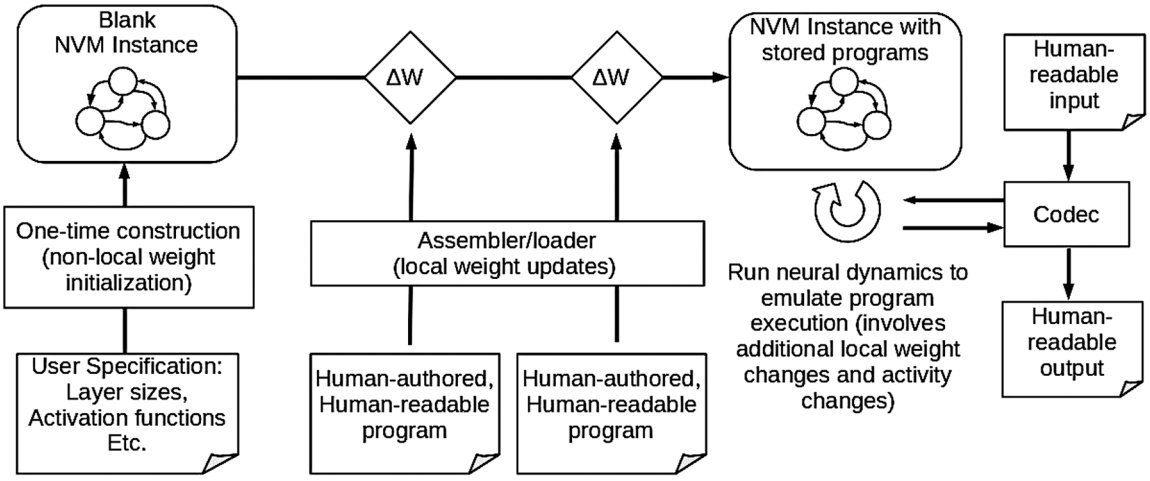

Figure 3: Using the NVM. After an initial description of a neurocomputational model's architecture (leftmost), human-authored programs are learned by that neural architecture (middle left) using local learning rules. Execution of the stored program(s) is emulated by running the dynamics of the underlying recurrent neural networks (middle right), which involves additional fast local weight updates. During execution a "codec" (coder-decoder) converts between the neural representations and human-readable input/output (rightmost).

The NVM supports all of the functionality that is standard in conventional computers but now instantiated as neural networks, with a clear architectural distinction between the domain-independent neural system itself and any associated domain-specific information. The NVM's functionality includes instruction operands, conditional branching, pointers, compositionality, and heterogeneous associative memory modules. The NVM can learn and use neural attractor sequences to represent not only arbitrary procedures, but also to remember arbitrary contiguous memory items in general. The constituent activity patterns of such sequences can be fed through a hetero-associative decoder to retrieve the corresponding memory contents. When appropriate, retrieved memory contents can effectively be treated as pointers and "dereferenced" by feeding them through the decoder again, allowing the NVM to represent trees and other hierarchical data structures in addition to sequences. Understanding how this structured representation occurs and its implications are important issues for mereological analysis.

Critically important to our research program here, the NVM's instruction set is sufficiently expressive so that the NVM can be programmed to represent the high-level symbolic concepts and algorithms relevant to imitation learning in CERIL. For example, its instruction set is sufficiently expressive so that the NVM can represent and learn cause-effect relations and can subsequently manipulate them in its working memory (compositionality) during problem solving. Learning is based on a one-step Hebbian learning rule that supports the ability to simultaneously store one pattern while erasing another. NVM input code is in human-readable form and capable of representing the procedures acquired during learning. It comes with an assembler and loader that can transparently convert these human-readable programs into the corresponding NVM encodings that are actually learned by the recurrent neural networks in the resulting neural system. Our work on the NVM promotes human-understandable knowledge representation and transparency by supporting a direct mapping between symbolic human-readable programs and system dynamics encoded in the NVM, and by using recent analytical methods that we developed for analyzing recurrent attractor neural network dynamics[43].

While the NVM is not intended to be a veridical model of the detailed microscopic neural circuitry of the brain, the NVM implementation that we have developed is potentially biologically relevant at the macroscopic neuroanatomical level. Applications using the NVM produce neural networks composed of simulated brain regions and pathways between them, and these recurrently-connected regions and pathways can be designed to mirror those existing in the brain. For example, in a separate application from

43 Katz and Reggia, "Using Directional Fibers to Locate Fixed Points", 3636.

our work here modeling brain regions relevant to post-traumatic stress disorder[44], there are NVM modules that are intended to be analogous to dorsolateral prefrontal cortex (stores procedures as neuro-dynamical itinerant attractor sequences) and basal ganglia (control instruction execution via a gating mechanism). Each module in this system is a recurrently connected network of continuously valued neurons (rate encoding) whose adaptive synaptic weights are modified during learning using various forms of one-step Hebbian learning (no use of iterative error backpropagation). Each instruction is performed with multiple gating operations that coordinate stepping through program memory and exchanging information between model regions.

Finally, the third step in our research program is to use the NVM to re-implement the symbolic AI target cognitive system CERIL as a *compositional working memory* based on only neurocomputational methods. This final step in our work is currently in progress. As noted above, our central hypothesis is that the three computational correlates of consciousness that we have identified will be sufficient to support the compositional and reasoning abilities needed without introducing additional innovations. If this proves to be the case, it will support the generality and plausibility of these three correlates by showing that they can address the broader mechanisms that are associated with working memory. This exploration may also lead to other novel insights and suggestions for other computational correlates of consciousness.

This third part of our research program is truly challenging in the context of current neural network technology. For example, it requires representing hierarchical cause-effect relations. Roughly speaking, the underlying knowledge is a set of temporal cause-effect relationships: "X can cause Y1 then Y2 then Y3, Y1 can cause Z1 then Z2, ...," where each cause or effect is an intention, sub-intention, or observable action that influences the external environment as was done in CERIL, but now represented as distributed activation patterns in neural networks. Such sequences can be stored as a sequence in NVM memory, where each pattern in the sequence is a "pointer" to another sequence also stored in a different area of the NVM memory's dynamical state space. These sequences each contain the individual cause and ordered effects for a particular relationship. Sequences of tokens are encoded by neural dynamics that transit through the corresponding patterns. These neural dynamics are ultimately determined by learned connection weights. All of this causal knowledge can be rapidly stored in working memory using the NVM via temporally symmetric and temporally asymmetric forms of one-step Hebbian learning as was done in our card-matching model. The key point here is that *the modeled causal knowledge and cognitive processes are a learned virtual machine,* represented by the distributed patterns of activity over the underlying neural substrate, and are not built into this underlying neural "hardware". Another major challenge is to implement CERIL's causal inference algorithm used during learning and its planning algorithm used after learning. The critical compositional aspect of these algorithms is based on constructing a problem-specific explanation of observations, i.e., abductive reasoning, based on the stored causal knowledge.

4 Summary and discussion

Past work that has developed computational models related to consciousness, either to deepen our understanding of its nature or to explore the possibility of artificial/machine consciousness, has often been based on hypothesized computational correlates of consciousness[45]. These include a global workspace, higher-order neural networks, and aspects of attention mechanisms (see Section 2). This past work has generally *started* with the assumption that a certain computational correlate exists and then developed a model based on that assumption. In the work described here we have taken a somewhat different approach. Noting that working memory and its cognitive control are widely viewed as associated with conscious thought, we developed a model of working memory. Only subsequently do we ask as to which core neurocomputational mechanisms were required to do this: What underlying computational mechanisms

44 Davis et al., "A Neurocomputational Model of Increased Saccade Latency and BOLD Changes in Posttraumatic Stress Disorder".

45 Reggia, "The Rise of Machine Consciousness", 116.

are critically needed to realize such a model and distinguish it from other contemporary neural networks in general? Our answer to this question is that at least three key mechanisms were needed in our model, and as such are candidates for computational correlates of consciousness: itinerant attractor sequences, top-down gating, and very fast weight changes.

The first of these, itinerant attractor sequences, deals with how learned information is represented and processed in recurrent neural networks. It asserts that conscious cognitive activity is represented as a sequence of learned cognitive states in working memory, where each cognitive state is represented as a distributed pattern of neural activity. The term "itinerant" here refers to a behavior in which the neural system gravitates towards a fixed activity state but that this state is unstable, leading to a learned transition to another attractor state, and so forth. While it has been suggested previously that individual activity attractor states may be relevant to consciousness[46], such suggestions have been of limited utility because they do not address the issue of transitions between attractor states and the concept of an attractor state in general is fairly generic. The computational correlate we are proposing here differs in defining mechanisms for transitions between attractor states and in associating these states specifically with learned representations of conceptual information. Itinerant attractor sequences mesh well with the idea that individual thoughts must persist for a significant period of time (on the order of 100 milliseconds or so) to become conscious, while still permitting a "train of thought" as a sequence of transient attractor states.

Our second proposed computational correlate, the use of *top-down gating* associated with the cognitive control of working memory, deals with how working memory is controlled by so-called executive functions. Such gating relates high-level cognitive processes for reasoning, problem-solving, and planning, to events in working memory. Top-down gating differs substantially from previously proposed computational correlates of consciousness. For example, unlike past computational models inspired by higher-order thought theory that are concerned with metacognitive states which *monitor* one another, gating involves modules that *control* each other's actions. Top-down gating also differs from, but complements and might ultimately include, past suggestions that top-down attention mechanisms (the efference copy associated with attention[47], multiple neural modules attending to the same topic[48], etc.) are computational correlates of consciousness. Of interest in this context is that direct and indirect top-down gating of neural activity is widely accepted by neuroscientists to occur in the brain, but remains only partially understood. Several hypotheses have been put forward as to the underlying mechanisms responsible for this functionality in the brain, including well-documented direct backwards connections between cortical regions[49], indirect influences via the network of basal ganglia and thalamic nuclei pathways[50], or even via functional mechanisms such as synchronization of cortical oscillations[51]. While our model necessarily makes a commitment to a specific computational mechanism to permit its implementation, our assertion that top-down gating of working memory is a computational correlate of consciousness is intended to be neutral regarding which of these *biological* mechanisms is/are ultimately found to be responsible for implementation of top-down gating in the primate brain.

Our third proposed computational correlate, *very fast weight changes*, deals with the speed with which information changes in working memory. Unlike long-term memory that requires a consolidation process and may require repeated presentation/rehearsal of information before it is retained, working memory is remarkably fluid. Immediate learning and retention occurs with a single presentation of information, and stored information can quickly disappear as it competes with new incoming information that needs to be retained. Our working memory model handles this issue using one-step Hebbian and anti-Hebbian learning mechanisms, respectively. This approach to weight changes is a major departure from the

46 Fekete and Edelman, "Towards a Computational Theory of Experience", 807; Taylor, "Neural Networks for Consciousness", 1207.

47 Taylor, "CODAM: A Neural Network Model of Consciousness", 987.

48 Haikonen, "Consciousness and Robot Sentience", 187.

49 Van Essen, "Corticocortical and Thalamocortical Information Flow", 173.

50 Chatham et al., "Corticostriatal Output Gating During Selection from Working Memory", 930; Sherman and Guillery, "Exploring the Thalamus and its Role in Cortical Function", 253.

51 Akam and Kullmann, "Oscillatory Multiplexing of Population Codes for Selective Communication in the Mammalian Brain", 111; Singer, "Dynamic Formation of Functional Networks by Synchronization", 191.

learning mechanisms used in most contemporary neurocomputational systems based on gradient descent methods, including in modern deep learning systems[52], although a few recent studies are beginning to incorporate fast weights into such models[53]. Hebbian learning, including temporally-asymmetric versions that are potentially important for learning sequential information, has been experimentally established as occurring in the mammalian nervous system. Further, recent empirical results in neuroscience have hypothesized that rapid synaptic changes are a neural correlate of working memory, consistent with our hypothesis[54].

The results presented here are, of course, predicated on our assumptions about the relationships of working memory and consciousness. In a strict sense, they are only relevant to a cognitive consciousness[55] or functionally-defined consciousness, such as access consciousness as we defined that term earlier[56]. They also do not consider, for example, possibilities such as panpsychism. However, in spite of these limitations, our results and other work on computational correlates of consciousness may ultimately prove very useful in informing fundamental philosophical theories of consciousness in unexpected ways. For example, we speculate that bridging the computational explanatory gap via the identification of computational correlates of consciousness may even eventually help with demystifying the hard problem in philosophy. This opinion is supported by a historical analogy with vitalism. The concept of life was just as mysterious to many scientists during the 1800's as the concept of consciousness is to us today. Many scientists at that time accepted the philosophical doctrine of vitalism[57]. Vitalism postulates a non-physical "life force" or "vital spirit" to living beings that is not possessed by non-living objects. Vitalists believed that the laws of physics and chemistry alone would never be able to fully account for living processes: there was an apparent philosophical explanatory gap between being alive and what could be explained mechanistically, similar to the philosophical explanatory gap concerning phenomenal consciousness today. Currently we believe that much of the mysteriousness underlying this philosophical explanatory gap concerning life was actually due to a "biophysical explanatory gap" involving the limited scientific understanding two hundred years ago of how processes associated with life (reproduction, inheritance, metabolism, and so forth) could be implemented by biophysical mechanisms. Today, while there is still no generally accepted definition of life[58], much of the mystery surrounding life that led to vitalism has faded away. This has occurred due to scientific advances in molecular genetics and evolution (e.g., discovery of DNA), an improved understanding of self-organization and emergent behaviors, a mechanistic explanation of cellular energy metabolism, and our ability to synthesize organic molecules from inorganic ones. We believe that, analogously, much of the mystery surrounding consciousness today will diminish as we develop a better understanding of the computational explanatory gap and gain a deeper understanding of the computational correlates of consciousness.

In summary, while much remains to be done, the computational correlates of consciousness that we have proposed here as well as those postulated by previous investigators provide encouragement for the potential of computational models to clarify issues related to the mind-brain problem. Philosophical theories about consciousness involving cognitive phenomenology and mereological analysis provide encouragement to those of us in computer science about this potential. Computational modeling work may in turn contribute back to the further development of such theories by exploring the (sometimes unexpected) implications of these theories when they are explored in detail. This is not only true for understanding human consciousness, but also in considering the difficult theoretical issues that arise when one analyzes the possibility of animal and/or machine consciousness more broadly.

Acknowledgements: This work was supported in part by ONR award N00014-19-1-2044.

52 Goodfellow et al., "Deep Learning", 161.

53 Ba, et al., "Using Fast Weights to Attend to the Recent Past"; Munkhdalai and Trischler, "Metalearning with Hebbian Fast Weights".

54 Bhandari & Badre, "A Nimble Working Memory", 503; Mongillo et al., "Synaptic Theory of Working memory", 1543.

55 Chalmers, "The Conscious Mind", 25.

56 Block, "On a Confusion about a Function of Consciousness", 230.

57 Garrett, "What the History of Vitalism Teaches Us about Consciousness", 616.

58 Regis, "What is Life?"; Wolfram, "A New Kind of Science", 1178.

References

Akam, Thomas, Kullmann, Dimitri. "Oscillatory Multiplexing of Population Codes for Selective Communication in the Mammalian Brain". *Nature Reviews Neuroscience*, 15 (2014), 111-122.

Ba, Jimmy, Hinton, Geoffrey, Mnih, Volodymyr, Leibo, Joel and Ionescu, Catalin. "Using Fast Weights to Attend to the Recent Past". Retrieved from arXiv :1610.06258v3, Dec. 2016.

Baars, Bernard. *A Cognitive Theory of Consciousness*, NY: Cambridge University Press, 1988.

Baars, Bernard, Franklin, Stan. "How Conscious Experience and Working Memory Interact". *Trends in Cognitive Science,* 7 (2003), 166-172.

Baddeley, Alan. "Working Memory and Conscious Awareness". In *Theories of Memory*, edited by A. Collins, S. Gattercole, et al., NY: Erlbaum (1993), 11-28.

Baddeley, Alan. "Working Memory: Theories, Models and Controversies". *Annual Review of Psychology*, 63 (2012), 1-29.

Bayne, Tim, Montague, Michelle, eds., *Cognitive Phenomenology*, Oxford: Oxford University Press, 2011.

Bhandari, Apoorva and Badre, David. "A Nimble Working Memory". *Neuron,* 91 (2016), 503-505.

Block, Ned. "On a Confusion about a Function of Consciousness". *Behavioral and Brain Sciences*, 18 (1995), 227-287.

Block, Ned. "Perceptual Consciousness Overflows Cognitive Access". *Trends in Cognitive Sciences*, 15:12 (2011), 567-575.

Bringsjord, Selmer. "Offer: One Billion Dollars for a Conscious Robot; If You're Honest, You Must Decline". *Journal of Consciousness Studies*, 14 (2007), 28-43.

Carruthers, Peter. *The Centered Mind*, Oxford: Oxford University Press, 2015.

Chalmers, David. *The Conscious Mind*, Oxford: Oxford University Press, 1996.

Chalmers, David. "What is a Neural Correlate of Consciousness?" In *Neural Correlates of Consciousness*, edited by T. Metzinger, Cambridge: MIT Press, 17-39, 2000.

Chatham, Christopher, Frank, Michael and Badre, David. "Corticostriatal Output Gating during Selection from Working Memory". *Neuron*, 81 (2014), 930-942.

Chella, Antonio and Manzotti, Riccardo. "Machine Consciousness: A Manifesto for Robotics". *International Journal of Machine Consciousness*, 1 (2009), 33-51.

Chudnoff, Elijah. *Cognitive Phenomenology*, Routledge Press, 2015.

Cleeremans, Axel. "Computational Correlates of Consciousness". *Progress in Brain Research*, 150 (2005), 81-98.

Cleeremans, Axel, Timmermans, Bert, Pasquali, Antoine. "Consciousness and Metarepresentation: A Computational Sketch". *Neural Networks*, 20 (2007), 1032-1039.

Connor, Dustin, Shanahan, Murray. "A Computational Model of a Global Neuronal Workspace with Stochastic Connections". *Neural Networks*, 23 (2010), 1139-1154.

Courtney, Susan, Petit, Laurent, Haxby, James and Ungerleider, Leslie. "The Role of Prefrontal Cortex in Working Memory: Examining the Contents of Consciousness". *Philosophical Transactions of the Royal Society of London B*, 353 (1998), 1819-1828.

Cowan, Nelson, Elliott, Emily, Saults, J. Scott, Morey, Candice, Mattox, Sam, Hismjatullina, Anna, Conway, Andrew. "On the Capacity of Attention: Its Estimation and Its Role in Working Memory and Cognitive Aptitudes". *Cognitive Psychology*, 51 (2005), 42-100.

Crick, Francis. *The Astonishing Hypothesis*, NY: Charles Scribner's Sons, 1994.

Davis, Greg, Katz, Garrett, Soranzo, Daniel, Costanzo, Michelle, Reinhard, Matthew, Gentili, Rodolphe, Reggia, James. "A Neurocomputational Model of Increased Saccade Latency and BOLD Changes in Posttraumatic Stress Disorder", submitted, 2019.

Dehaene, Stanislas, Naccache, Lionel. "Towards a Cognitive Neuroscience of Consciousness". *Cognition*, 79 (2001), 1-37.

Fekete, Tomer, Edelman, Shimon. "Towards a Computational Theory of Experience". *Consciousness and Cognition*, 20 (2011), 807-827.

Feldman, Jerome. "The Neural Binding Problem". *Cognitive Neurodynamics*, 7:1 (2013), 1-11.

Frank, Michael, Loughry, Bryan and O'Reilly, Randall. "Interactions Between Frontal Cortex and Basal Ganglia in Working Memory: A Computational Model". *Cognitive, Affective and Behavioral Neuroscience,* 1 (2001), 137-160.

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

Goodfellow, Ian, Bengio, Yoshua and Courville, Aaron. *Deep Learning*, NY: MIT Press, 2016.

Haikonen, Pentti. *Consciousness and Robot Sentience*, Singapore: World Scientific, 2019.

Jorba, Marta, Vincente, Agustin. "Cognitive Phenomenology, Access to Contents, and Inner Speech. *Journal of Consciousness Studies*, 21:9-10 (2014), 74-99.

Katz, Garrett, Davis, Greg, Gentili, Rodolphe, Reggia, James. "A Programmable Neural Virtual Machine Based on a Fast Store-Erase Learning Rule". *Neural Networks*, 119 (2019), 10-30.

Katz, Garrett, Huang, Di-Wei, Hauge, Theresa, Gentili, Rodolphe and Reggia, James. "A Novel Parsimonious Cause-Effect Reasoning Algorithm for Robot Imitation and Plan Recognition". *IEEE Transactions on Cognition and Developmental Systems*, 10 (2018), 177-193.

Katz, Garrett, Reggia, James. "Using Directional Fibers to Locate Fixed Points of Recurrent Neural Networks". *IEEE Transaction on Neural Networks and Learning Systems*, 29 (2018), 3636-3646.
Lara, Antonio and Wallis, Jonathan. "The Role of Prefrontal Cortex in Working Memory: A Mini Review". *Frontiers in Systems Neuroscience.* Retrieved from https://doi.org/10.3389/fnsys.2015.00173. Accessed December 18, 2015.
Massimini, Marcello, Ferrarelli, Fabio, Huber, Reto, Esser, Steve, Singh, Harpreet, Tononi, Giulio. "Breakdown of Cortical Effective Connectivity During Sleep". *Science*, 309 (2005), 2228-2232.
McDermott, Drew. "Artificial Intelligence and Consciousness." In *Cambridge Handbook of Consciousness*, edited by M. Moscovitch and E. Thompson, 117-150, Cambridge: Cambridge University Press, 2007.
McGinn, Colin. Can We Solve the Mind-Brain Problem? *Mind*, 98 (1989), 349-366.
Metzinger, Thomas. *Neural Correlates of Consciousness*, NY: MIT Press, 2000.
Mongillo, Gianluigi, Barak, Omri and Tsodyks, Mish. "Synaptic Theory of Working Memory". *Science*, 319 (2008), 1543-1346.
Munkhdali, Tsenduren and Trischler, Adam. "Metalearning with Hebbian Fast Weights". Retrieved from arXiv: 1807.05076v1, 2018.
Nagel, Thomas. "What is it Like to be a Bat?" *Philosophical Review*, 4 (1974), 435-450.
Pascanu, Razvan and Jaeger, Herbert. "A Neurodynamical Model for Working Memory". *Neural Networks*, 24 (2011), 199-207.
Perlis, Don and Brody, Justin. "Operationalizing Consciousness". *AAAI Spring Symposium*, Stanford CA, 2019.
Persuh, Marjan, LaRock, Eric and Berger, Jacob. "Working Memory and Consciousness: The Current State of Play". *Frontiers in Human Neuroscience*. Retrieved from https://doi.org/10.3389/fnhum.2018.00078. Accessed March 2018.
Prentner, Robert. "Process Metaphysics of Consciousness". *Open Philosophy*, 2 (2018) 3-13.
Prinz, Jesse. *The Conscious Brain*, Oxford: Oxford University Press, 2012.
Reggia, James. "The Rise of Machine Consciousness". *Neural Networks*, 44 (2013), 112-131.
Reggia, James, Monner, Derek, Sylvester, Jared. "The Computational Explanatory Gap". *Journal of Consciousness Studies*, 21:9 (2014), 153-178.
Regis, Ed. *What is Life?*, NY: Farber, Strauss and Giroux, 2008.
Sherman, S. Murray, Guillery, R. *Exploring the Thalamus and its Role in Cortical Function*, NY: MIT Press, 2006.
Singer, Wolf. "Dynamic Formation of Functional Networks by Synchronization". *Neuron*, 69 (2011), 191-193.
Squire, Larry and Dede, Adam. "Conscious and Unconscious Memory Systems". *Cold Spring Harbor Perspectives in Biology, 7* (2015). Retrieved from DOI:10.1101/cshperspect.a021667.
Sylvester, Jared, Reggia, James, Weems, Scott, Bunting, Michael. "Controlling Working Memory with Learned Instructions". *Neural Networks*, 41 (2013), 23-38.
Sylvester, Jared, Reggia, James. "Engineering Neural Systems for High-Level Problem Solving". *Neural Networks, 79* (2016), 37-52.
Takeno, Junichi. *Creation of a Conscious Robot*, Singapore: Pan Stanford, 2013.
Taylor, John. "Neural Networks for Consciousness". *Neural Networks*, 10 (1997), 1207-1225.
Taylor, John. "CODAM: A Neural Network Model of Consciousness". *Neural Networks*, 20 (2007), 983-992.
Van Essen, David. "Corticocortical and Thalamocortical Information Flow in the Primate Visual System". *Progress in Brain Research*, 149 (2005), 173-185.
Verduzco-Flores, Sergio, Ermentrout, Brad and Bodner, Mark. "Modeling Neuropathologies as Disruption of Normal Sequence Generation in Working Memory Networks". *Neural Networks*, 27 (2012), 21-31.
Wolfram, Stephen. *A New Kind of Science*, *Champaign*, *Illinois:* Wolfram Media, 2002.