The Computational Explanatory Gap

James A. Reggia, Derek Monner, and Jared Sylvester
University of Maryland, College Park, US
{reggia,dmonner,jared}@cs.umd.edu

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 this is due to a computational explanatory gap: our inability to understand/explain the implementation of high-level cognitive algorithms in terms of neurocomputational algorithms. Contrary to prevailing views, we will suggest that bridging this gap is not only critical to further progress in the area of machine consciousness, but is also a fundamental step towards solving the hard problem. We briefly describe some small steps we have taken recently to make progress in this area.

1 Introduction

While the idea of a conscious machine is not a particularly new one (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 in 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 conscious appear to be a rather fruitless task. For example, it has been suggested 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. More specific arguments have been presented in recent years that phenomenal machine consciousness is simply not possible. Examples along these lines include analyses indicating that phenomenal machine consciousness would imply panpsychism (Bishop, 2000), that computation is insufficient to underpin consciousness (Manzotti, 2012), and that machines cannot be conscious due to their non-organic nature (Schlagel, 1999). Individuals who advocate or study the possibilities of machine consciousness have so far not found such arguments persuasive.

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 several years in the field of artificial consciousness (Reggia, 2013). Here we show that by distinguishing between simulated 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 argue that a major and fundamental barrier to further progress on creating phenomenal/instantiated machine consciousness is

a *computational explanatory gap*: our current lack of understanding concerning how high-level cognitive computations can be captured in low-level neural algorithms. The significance of this gap is that bridging it may be a critical step in addressing the original philosophical explanatory gap, and thus in making advances on the mind-brain problem during coming years. As evidence that bridging the computational explanatory gap may be tractable, a brief summary is given of some (small) steps we and others have taken recently towards implementing high-level cognitive processing in neurocomputational models.

2 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 versus instantiation of consciousness. Such a distinction parallels the distinction between information processing aspects of consciousness (functionalism) and subjective experience (phenomenal consciousness).

With *simulated consciousness*, the goal is to capture some aspect of consciousness or its neural/behavioral correlates 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 conscious information processing would be "conscious inside". There is no real claim that phenomenal consciousness is actually present in this situation. The results of a simulation are 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 experiences 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 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 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 consciousness, 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 just 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, 2008) 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, 2008), essentially passing the well-known mirror test used to identify self-recognition in animals (Gallup, 1970).
- Showing that second-order neural networks can match behavioral data from human blindsight subjects during post-decision wagering tasks (Pasquali et al., 2010), supporting core aspects of higher order thought (HOT) theories of consciousness (Rosenthal, 1996; 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 simulated consciousness have provided useful information for advancing consciousness studies; we do not consider them further in this paper.

The situation is quite different from the perspective of instantiated consciousness. Several investigators have claimed to have either designed or created phenomenally conscious artifacts. To give just a few examples of the claims that have been made:

- Any system that maintains a correspondence between high-level, symbolically represented concepts and low-level data stream entities, and that has a reasoning system which makes use of these grounded symbols, has true subjective experiences corresponding to qualia and a sense of self-awareness (Kuipers, 2005).
- A 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).

While such claims are intriguing, at the present time it seems reasonable to conclude that no existing computational approach to artificial consciousness has yet presented a compelling demonstration of instantiated consciousness in a machine, or even clear evidence that instantiated machine consciousness will eventually be possible. This conclusion does *not* preclude the possibility of ultimately creating a phenomenally conscious machine any more than the inability to produce machine-powered flight prior to the Wright brothers showed that mechanical flying machines were impossible (as some scientists argued at the time). But it does raise 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 investigation.

3 The Computational Explanatory Gap

There are a number of well-recognized barriers to creating instantiated machine consciousness. 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?). There is another less recognized barrier, the *computational explanatory gap*, that we would argue is also of critical importance: our current lack of understanding of how high-level cognitive information processing can be mapped onto low-level neural computations. This gap can be contrasted with the widely recognized *philosophical explanatory gap* between a successful functional/computational account of consciousness and the subjective experiences that accompany it (Levine, 1983). The computational explanatory gap is not a mind-brain issue per se. Rather, it is a gap in our understanding of how computations/algorithms at a high level of cognitive information processing can be mapped into computations/algorithms at the low level of neural networks. In other words, it is a purely computational issue (Figure 1). Contemporary philosophical thought tends to largely dismiss solving the computational explanatory gap as the "easy problem", while bridging the philosophical explanatory gap is viewed as the "hard problem" (Chalmers, 1996). In contrast, we conjecture that, with high probability, this perspective will turn out to be precisely backwards; the computational explanatory gap is actually the more fundamental issue, and that once it is bridged, the philosophical explanatory gap will be found to be tractable and fade away.

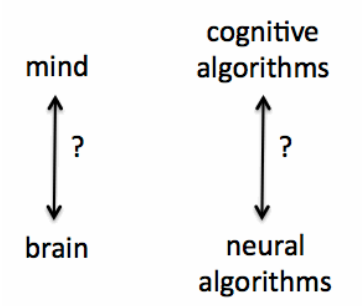

Figure 1: Mind the gap: the well-known philosophical (left) and computational (right) gaps. Our argument is that the latter may ultimately prove to be the more fundamental problem, and that focusing on solving it rather than dismissing it may be the key to advancing future work on instantiated machine consciousness.

The computational explanatory gap has influenced work in a number of disciplines, such as AI and neuroscience. In AI this gap is reflected in the long-standing debate concerning the relative values of top-down (symbolic) vs. bottom-up (neural, swarm, etc.) approaches to creating machine intelligence (Franklin, 1995). This (in)famous debate has largely missed the point that these two approaches are not so much competing alternatives as complementary in what they each capture about intelligence. Top-down symbolic methods have excelled at modeling high-level cognitive tasks such as reasoning, decision making, "understanding" natural language, and planning, but they have been much less successful at pattern recognition and low level control. They have generally been found to be brittle, for example, failing in the context of noise or novel situations. In contrast, 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. Similarly, the computational explanatory gap is evident in neuroscience, where a lot is 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 microscopic functions of neural circuitry all the way down to the molecular and genetic levels, but it remains unclear how to put those two types of information together. 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 gap in our neuroscientific knowledge is, at least in part, a manifestation of the underlying computational explanatory gap: how are the "algorithms" associated with large-scale brain regions (e.g., Granger, 2006) mapped into computations performed by microscopic biological neural nets?

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) computational mechanisms associated with conscious/reportable high-level cognitive activities, and (ii) computational 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. If these correlates can be identified, then it would provide a direct route to investigating the possibility of instantiated machine consciousness. If no correlates/differences between the neurocomputational implementation of conscious and unconscious cognitive functions can be found, that too would have tremendous implications for the modern functionalist/computationalist viewpoint of the mind-brain problem.

4 Steps Towards Bridging The Gap

If one allows the possibility that the "easy problem" represented by the computational explanatory gap is important, and perhaps even a fundamental barrier to instantiated machine consciousness, then the immediate research program becomes determining how we can bridge this gap. Encouragingly,

there has been a substantial effort over the last two decades by researchers unconcerned with the issue of machine conscious to understand how higher cognitive functions (reasoning, language, etc.) can be implemented in neurocomputational substrates. While these models have not yet reached the effectiveness of models based on top-down, symbolic AI systems, they have clearly shown that substantial progress can be made in bridging the computational explanatory gap.

In addition, a number of investigators explicitly studying artificial consciousness have examined hypotheses that relate to the computational explanatory gap. For example, some studies have created hybrid symbolic-connectionist systems, consisting of a high-level cognitive module implemented as a symbolic, top-down architecture, plus a lower-level cognitive module implemented using bottom-up neurocomputational methods (Chella, 2007; Kitamura et al, 2000; Sun, 1999; these and others are reviewed in Reggia, 2013). Such models can be viewed as starting with the interesting *claim* that symbolic information processing per se is the basis of conscious information processing, and investigate the extent to which a model can account for some neural and cognitive correlates of consciousness. In contrast, the computational explanatory hypothesis as described here suggests that the critical issue is how to replace the symbolic (often production rules or predicate logic) modules of such models with neurocomputational implementations, and then to determine what (if any) differences exist between the properties of the needed computational mechanisms at the two levels. From this viewpoint, symbolic-connectionist hybrid architectures (regardless of their potential validity as a theory of consciousness) obscure the central issue of the computational explanatory gap by introducing a separate confounding factor (i.e., the a priori introduction of symbolic vs. connectionist information processing methods in addition to conscious vs. unconscious information processing).

More directly related to the computational explanatory gap is recent work that has tried to directly extend neurocomputational methods to high-level cognitive tasks associated with consciousness. Metacognitive neural networks provide a good example of this. Tied to philosophical concepts from higher-order thought (HOT) theories (Rosenthal, 1986; Carruthers, 2005), such studies have been based on second-order neural networks that interpret the behavior of first-order networks (Cleeremans et al., 2007; Pasquali et al., 2010).

We have been taking a somewhat different approach to diminishing the computational explanatory gap by trying to reverse-engineer functional properties of cerebral cortex, including its large-scale architecture, with respect to language and working memory. This work is inspired in part by contemporary evidence that higher cognitive functions are implemented by a large-scale network of cortical regions; these regions interact directly via well-known neuroanatomical pathways, and indirectly via pathways between cortical regions and subcortical centers (thalamic nuclei, basal ganglia, hippocampus, etc.). With respect to language, we have established that it is possible for neurocomputational models of the cortical language areas (Fig. 2, left) to learn to perform simple word processing tasks while grounding words in "seen" images, and that such models break down in ways similar to what is observed in people following localized cortical damage (Weems and Reggia, 2006). Further, we recently extended this investigation, showing that a neurocomputational model could learn to perform simple question answering at a sentence level (Monner & Reggia, 2012a,b), a task that has only been modeled previously using the methods of symbolic AI. Interestingly, analysis of these and similar models has discovered that they learn a grounded latent symbol system that supports combinatorial computations using distributed representations (Monner & Reggia, 2011, 2013) – a clear step towards bridging the computational explanatory gap. With respect to working memory, we have developed attractor neural network models that learn temporal sequences and shown that their performance can match empirical data from human behavioral experiments (Winder et al., 2009; Sylvester et al., 2010). Most recently we have studied a region-pathway network model, inspired by prefrontal cortex (Fig. 2, right), where regions in the model are each sequence-learning attractor neural network modules (Sylvester et al., 2013). This latter model captures some aspects of human cognitive control of both working memory and the learning of task procedures autonomously, produces accuracy and timing results that correlate with those of human subjects performing similar

tasks, and makes testable predictions. These language and working memory models suggest to us that latent symbol systems, and the use of cortical modules that not only exchange information but also gate one another, are potential candidates for computational correlates of consciousness. All of these results related to language and working memory represent small steps towards bridging the computational explanatory gap, encouraging further work in this direction.

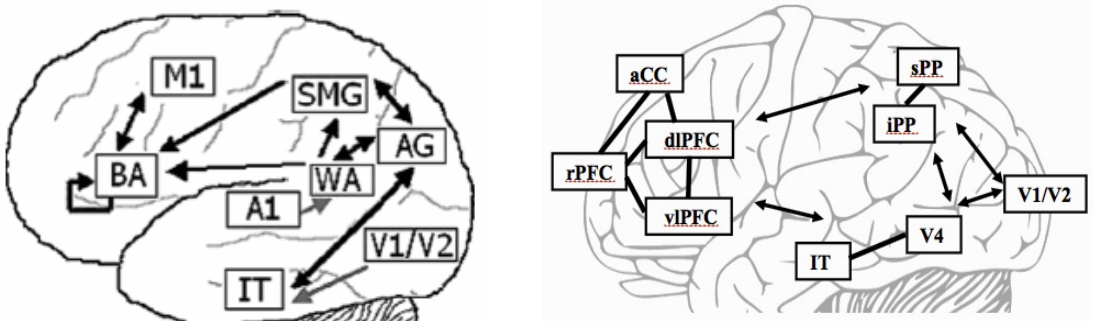

Figure 2: Cerebral cortex can be viewed as a large-scale network of cortical regions connected by pathways. This is illustrated here for language (left) and cognitive control of working memory (right).

5 Discussion

Distinguishing between simulated and instantiated machine consciousness clarifies both the progress and limitations of past work in the field of artificial consciousness. 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 what we have called simulated artificial consciousness. This is extremely important; it is 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. It 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 can reason more effectively and interact with people in more natural ways. It appears likely that simulated consciousness will play a significant role in future work on creating an artificial general intelligence. Put simply, work on simulated consciousness has become an effective and 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 instantiated machine consciousness is largely due to the computational explanatory gap: our current lack of understanding how higher-level cognitive algorithms can be mapped onto neurocomputational algorithms. 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 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. It is possible that bridging the computational explanatory gap will make bridging the philosophical explanatory gap tractable, and that it may lead to an operational test for the presence of phenomenal consciousness. Doing so, and bridging this 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 than it currently appears.

References

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.

Bishop, M. (2009) Why Computers Can't Feel Pain, *Minds and Machines*, 19, 507-516.

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

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

Chalmers, D. (1996) *The Conscious Mind*, Oxford University Press.

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

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

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.

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

Gallup, G. (1970) Chimpanzees: Self-Recognition, *Science*, 167, 86-87.

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

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

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

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

Massimini, M., Ferrarelli, F., Huber, R., et al. (2005) Breakdown of Cortical Effective Connectivity During Sleep, *Science*, 309, 2228-2232.

McGinn, C. (2004) *Consciousness and Its Origins*, Oxford University Press.

Pasquali, A., Timmermans, B., Cleeremans, A. (2010) Know Thyself: Metacognitive Networks and Measures of Consciousness, *Cognition*, 117, 182-190.

Reggia, J. (2013). The Rise of Machine Consciousness, *Neural Networks*, in press.

Rodrigues, A. et al. (2004) Derivation and Analysis of Basic Computational Operations of Thalamocortical Circuits, *J. Cognitive Neuroscience*, 16, 856-877.

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

Rosenthal, D. (1996) A Theory of Consciousness. In N. Block, et al. (eds.), *The Nature of Consciousness*, MIT Press, 729-753.

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.
Sun, R. (1999) Accounting for the Computational Basis of Consciousness, *Consciousness and Cognition*, 8, 529-565.
Sylvester J, Reggia J, Weems S, Bunting M (2013) Controlling Working Memory with Learned Instructions, *Neural Networks*, 41, 23-38.
Takeno, J. (2008) A Robot Succeeds in 100% Mirror Image Cognition, *International Journal on Smart Sensing and Intelligent Systems*, 1, 891-911.
Taylor, J. (2007) CODAM: A Neural Network Model of Consciousness, *Neural Networks*, 20, 983-992.
Tononi, G (2004) An Information Integration Theory of Consciousness, *BMC Neuroscience*, 5:42.
Tononi, G (2008) Consciousness as Integrated Information, *Biological Bulletin*, 215, 216-242.