Neural Networks 44 (2013) 112–131

Contents lists available at SciVerse ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

The rise of machine consciousness: Studying consciousness with computational models

James A. Reggia *

Department of Computer Science, A. V. Williams Building, University of Maryland, College Park, MD 20742, United States

ARTICLE INFO

Article history:
Received 10 November 2012
Received in revised form 13 March 2013
Accepted 14 March 2013

Keywords:
Artificial consciousness
Machine consciousness
Review

ABSTRACT

Efforts to create computational models of consciousness have accelerated over the last two decades, creating a field that has become known as *artificial consciousness*. There have been two main motivations for this controversial work: to develop a better scientific understanding of the nature of human/animal consciousness and to produce machines that genuinely exhibit conscious awareness. This review begins by briefly explaining some of the concepts and terminology used by investigators working on machine consciousness, and summarizes key neurobiological correlates of human consciousness that are particularly relevant to past computational studies. Models of consciousness developed over the last twenty years are then surveyed. These models are largely found to fall into five categories based on the fundamental issue that their developers have selected as being most central to consciousness: a global workspace, information integration, an internal self-model, higher-level representations, or attention mechanisms. For each of these five categories, an overview of past work is given, a representative example is presented in some detail to illustrate the approach, and comments are provided on the contributions and limitations of the methodology. Three conclusions are offered about the state of the field based on this review: (1) computational modeling has become an effective and accepted methodology for the scientific study of consciousness, (2) existing computational models have successfully captured a number of neurobiological, cognitive, and behavioral correlates of conscious information processing as machine simulations, and (3) no existing approach to artificial consciousness has presented a compelling demonstration of phenomenal machine consciousness, or even clear evidence that artificial phenomenal consciousness will eventually be possible. The paper concludes by discussing the importance of continuing work in this area, considering the ethical issues it raises, and making predictions concerning future developments.

> "There is no security … against the ultimate development of mechanical consciousness, in the fact of machines possessing little consciousness now".
>
> [Samuel Butler, *Erewhon*, 1872]

1. Introduction

What is the nature of consciousness? Can a machine be conscious? Over the last two decades there have been increasing efforts to shed light on these questions in the field that is becoming known as *artificial consciousness*, or synonymously, *machine consciousness*. Work in this area focuses on developing computational models of various aspects of the conscious mind, either with software on computers or in physical robotic devices. Such efforts face substantial barriers, not the least of which is that there is no generally agreed-upon definition of consciousness. In this review of past computational studies that model consciousness, we will simply define consciousness to be the subjective experiences/awareness that one has when awake, following earlier suggestions that a more precise definition, given our currently inadequate scientific understanding of consciousness, is best left to the future (Crick, 1994; Searle, 2004). Some recent discussions of more formal definitions and the philosophical difficulties that these raise can be found, for example, in Block (1995), Sloman (2010) and Zeman (2001).

What would be the value of studying consciousness with machines? There are two main answers to this question. First and foremost would be to improve our scientific understanding of the nature of consciousness. At the present time, our understanding of how a physical system such as the human brain can support the subjective experiences that are at the core of a conscious mind are largely pre-scientific, and some philosophical discussions have argued that, almost by definition, the objective methods of science will never be able to shed light on consciousness due to its

* Tel.: +1 301 405 2686; fax: +1 301 405 6707.
E-mail address: reggia@cs.umd.edu.

0893-6080/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.neunet.2013.03.011

subjective nature (McGinn, 2004). Individuals working on artificial consciousness obviously take a much more optimistic view of this issue, observing that computer models of specific aspects of consciousness (simulated consciousness) may prove to be useful in advancing our understanding of conscious information processing, just as computer models are useful in many other fields. The second primary motivation for work in artificial consciousness is the technological goal of creating a conscious machine (instantiated consciousness). As Samuel Butler's comment above indicates, this is not a particularly new idea, but it is one that has only been pursued through concrete steps during the last few decades (at least since the early 1980's; see Culbertson (1982) and Wilks (1984)). It is this aspect of work on machine consciousness that has proven most controversial. In part, interest in designing conscious machines comes from the limitations of current artificial intelligence (AI) systems. Most AI researchers and developers today take the viewpoint that one determines whether an artifact possesses intelligence based on behavioral criteria, e.g., the Turing Test (Turing, 1950). By such criteria, at present we have only ''weak AI'' that in most ways does not yet come close to the general intellectual abilities of the human mind. It is possible that adding conscious awareness, or information processing capabilities associated with the conscious mind, would open the door to a much more powerful and general AI technology.

Past computational models are considered for inclusion in this review if and only if the investigators doing the research, or others discussing it, have explicitly claimed that the work is studying, modeling, or in some fashion examining a significant aspect of consciousness. Such a criterion is by no means perfect, but it has the advantage of bypassing definitional difficulties, effectively taking the field of artificial consciousness to be ''that which its investigators are studying''. To the author's knowledge, while there have been several past introductions to machine consciousness (Aleksander, 2007; Holland, 2009; Sun & Franklin, 2007; Taylor, 2003a), there has only been one previous systematic review (Gamez, 2008). The current review differs not only in including the many studies published since 2007, but also in providing a different classification and perspective on past work in this area. Reports of computational studies of consciousness have been increasing rapidly, and are widely scattered throughout the literature of various disciplines (such as neural computation, psychology, neuroscience, philosophy, and AI). This makes a review of the current state-of-the-art both timely and useful in collecting in one place an overview and assessment of the different approaches under study.

The material that follows begins by providing two types of background information that are intended to make this review more widely accessible: alternative philosophical views about the nature of consciousness, and some of the known neural correlates of consciousness. The focus is on summarizing concepts and terminology widely used by researchers in artificial consciousness, and on providing context for discussing some of the neurocomputational models that are based on known neural correlates. After this background information, a systematic review of a broad range of computational models related to conscious mind is provided. Every effort is made in this review to be broad, representative, and unbiased concerning the specific approaches that have been taken. The past models examined are found to fall into five categories based on the fundamental aspect of consciousness that they take to be central to the computational study involved: a global workspace, information integration, internal self-models, higher-level representations, or attention mechanisms. For each of these five approaches, three types of information are presented: (1) an overview of the approach reviewing past work that has been done in the area, (2) an example system described in more detail to illustrate the approach, and (3) an assessment of the methodology's contributions and limitations. This review concludes by assessing the current state-of-the-art and by examining the prospects for ultimately creating a conscious machine.

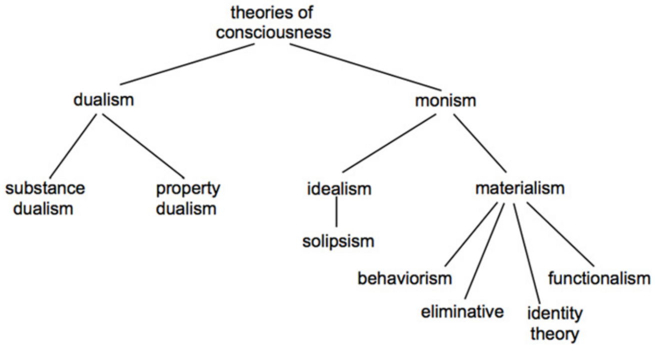

Fig. 1. A sketch of some theoretical positions concerning the nature of conscious mind. Individuals developing computational models of consciousness predominantly approach the issue from a functionalist viewpoint (bottom right). See text for details.

2. The nature of consciousness

How should one view consciousness when examining its simulation or instantiation in machines? The difficulty experienced in even defining consciousness is underscored by the broad range of theories that have been put forward about its fundamental nature and its neurobiological basis (Atkinson, Thomas, & Cleeremans, 2000; Churchland, 1984; Haikonen, 2012; Kriegel, 2007; Searle, 2004). While we cannot resolve these conflicting perspectives here, we can at least be clear about the assumptions and terminology being used by most past work on artificial consciousness. We start by considering how work on artificial consciousness relates to past philosophical views, and consider some neurobiological correlates of consciousness that have inspired the computational models discussed below.

2.1. Philosophical perspectives

In the most general terms, different philosophical perspectives concerning consciousness are often viewed as falling into two broad categories, dualism and monism, as illustrated in Fig. 1.

Dualism posits that there are two distinct realms of existence, the objective material universe that we perceive all around us and a subjective non-physical universe that encompasses the conscious mind. While dualism may actually be the most widely held philosophy of mind in the population at large (Churchland, 1984; Searle, 2004), it is often dismissed by scientific investigators because of its non-physical aspects, the absence of testable predictions, and its inability to explain how fundamentally different physical and non-physical universes/substances could interact (not all scientists reject dualism; for example, see Eccles (1994)). While variants such as property dualism (asserts that there are two types of properties, mental and physical) address some of these issues, dualism has rarely guided past work related to machine consciousness and so is not considered further here.

In contrast to dualism, *monism* assumes that there is only one realm of existence. This leads to two possibilities. The first of these, *idealism*, takes the position that there is only a single, subjective non-physical universe, and that the physical universe is only imagined (Goswami, Reed, & Goswami, 1993; Hutto, 2009). Again, most artificial consciousness investigators dismiss this perspective because it leads to solipsism (the belief that only one's own mind is certain to exist), and because viewing the physical world as essentially ''a dream'' makes the scientific study of idealism very problematic.

We are left then with *materialism* (or *physicalism*), a viewpoint which asserts that conscious mind is a phenomenon that is part of the material/physical world. Further, and in spite of our current ignorance concerning how consciousness arises and its apparently mysterious nature, this perspective assumes that consciousness has a scientific explanation that can ultimately be discovered. Such an explanation would demystify consciousness in the same way that scientific knowledge has, at least in part, greatly demystified the nature of previously confusing concepts such as life and intelligence.

As shown in Fig. 1, there are a variety of materialist perspectives. For example, philosophical *behaviorism* replaces discussion of inner mental states with analysis of behavioral dispositions, effectively avoiding the mind–brain problem all together. *Eliminative materialism* denies the validity of our current common sense psychological concepts (''folk psychology''), and expects that they will ultimately be replaced by a more mature conceptual framework (Churchland, 1984), thus eliminating the need to explain consciousness as it is currently conceived. If one accepts either of these first two viewpoints, then there would be little value in studying machine consciousness. *Identity theory* (or *reductive materialism*) states that mental states can be reduced to, and are effectively the same as, physical states of the brain. The mind is taken to be equivalent to the electrochemical states of biological neuronal networks. This latter viewpoint precludes the possibility of machine consciousness.

In contrast, *functionalism* claims that while mental states correspond to brain states, they are mental states due to their functionality, not due to their physical composition. In other words, conscious mental states are functional states (or computational states) of the brain. The term ''functional state'' means that a state is defined in terms of what it does, not what physical structure underlies that state. For example, a hand (a single physical object) can exist in a variety of functional states that indicate departure (waving), anger (fist), a direction (pointing), approval (thumbs up), and so forth. Functionalism argues that conscious mental states are analogous when it comes to the brain, but on a much larger and more complex scale. Further, mental states are connected via causal relationships to the body, external stimuli, behavior, and other mental states—it is these causal relationships that define the functional states of the brain. Crucially for work on artificial consciousness, and unlike with identity theory, functionalism implies that non-biological machines could possibly be created that are conscious and have a mind (''strong AI'', or ''computationalism''). In such a scenario, computer hardware replaces the brain as a physical substrate, while executing software with similar functionality serves as an artificially conscious mind (Searle, 2004).

It is probably the case that the vast majority of individuals investigating the philosophical and scientific basis of consciousness today, including those developing computer models of consciousness, are functionalists (Churchland, 1984; Eccles, 1994; Searle, 2004). After all, calling a machine a ''computer'' is already implying a functional definition: a computer is a device that carries out a mechanical procedure or algorithm, regardless of its size or the materials from which it is made (electronics, wheels and gears, or biological components) (Hillis, 1998). Metaphorically, the functional states of computer hardware that is executing a program can already be viewed as analogous to mental states of the brain, so if one is a functionalist, it is not such a great leap to argue that computational models of consciousness are possible and worth exploring.

There is a very important distinction that is often made in the consciousness literature; this distinction arises due to the ambiguity of the word ''consciousness''. Specifically, there is a crucial difference between what are referred to as the *easy problem* and the *hard problem* of consciousness (Chalmers, 1996, 2007). The easy problem refers to understanding the information processing aspects of conscious mind: cognitive control of behavior, the ability to integrate information, focusing of attention, and the ability of a system to access and/or report its internal states. Calling these problems ''easy'' does not mean that solving them will be easy, but that we believe that doing so will ultimately be possible in terms of computational and neurobiological mechanisms within the framework of functionalism. In contrast, the hard problem of consciousness refers to the *subjective experience* associated with being conscious. The term *qualia* (felt qualities) is often used for our subjective experiences, such as the sensation of redness we experience during a sunset, the pain of a toothache, or the smell of a rose. The mystery here (the ''hard'' nature of the problem) is why it feels like anything at all to be conscious (Nagel, 1974). Even if science ultimately explains the information processing mechanisms for all of the ''easy'' problems of consciousness, this will not explain *why* these mechanisms or functions do not occur without subjective awareness, but are instead accompanied by subjective experiences or qualia. In other words, there is an ''explanatory gap'' between a successful functional/computational account of consciousness and the subjective experiences that accompany it (Levine, 1983). The term *phenomenal consciousness* is often used to emphasize that one is referring specifically to the phenomena of subjective experience (qualia). This can be contrasted with the term *access consciousness* which refers to the availability of information for conscious processing, a decidedly functionalist concept.

To summarize, there are two important issues that should be understood in considering the past work on artificial consciousness reviewed below. First, there is an important distinction that is often made between the information processing aspects of consciousness (relates to functional theories, the easy problem of consciousness, and access consciousness) and the associated subjective experiences that accompany them (relates to phenomenal consciousness, the hard problem, and qualia). Second, the vast majority of past work in artificial consciousness has been done from the perspective of materialism (assumes there is a physical explanation of consciousness) and specifically of functionalism (assumes conscious mental states are functional states). This latter functionalist perspective, taken a priori, leaves issues related to phenomenal machine consciousness unresolved, a point that we will consider further in the Discussion.

2.2. Neural correlates of consciousness

How does consciousness relate to the brain? There is a large and growing literature on this topic in neuroscience today (Atkinson et al., 2000; Block, 2005; Crick & Koch, 2003; Koch, 2004; Metzinger, 2000b; Rees, 2009; Rees, Kreiman, & Koch, 2002; Ward, 2011). This empirical evidence concerning the mind–brain relationship is derived from a variety of sources, including physiological studies in animals, EEG and functional imaging (fMRI, PET, etc.) in people, and studies of the effects of focal brain damage. In particular, much of the evidence that we consider is based upon *contrastive analysis* in which one compares the difference in brain activity patterns during two similar events, one of which is conscious and the other of which is very similar but unconscious. For example, it is possible under laboratory conditions to control whether or not human subjects become conscious of a written word that is seen very briefly (Dehaene et al., 2001). In such situations, patterns of brain activity seen only when the subject consciously perceives the word are an example of a *neural correlate of consciousness* (a brain state corresponding to a state of consciousness). The word ''correlate'' is chosen carefully here rather than ''causes'' because it can be difficult to establish causal relationships between brain activity patterns and consciousness. In the following, we briefly consider just a few of the known neurobiological observations relating to consciousness, restricting our attention to those most relevant to research on machine consciousness.

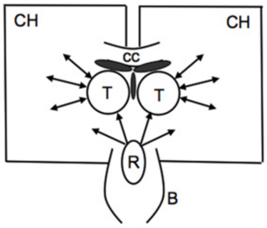

Fig. 2. Idealized sketch of a vertical cross-section of the human brain showing the two cerebral hemispheres (CH), brainstem (B), thalamus (T), ascending arousal system (R), and corpus callosum (cc).

Anatomically, the regions of the brain most closely associated with consciousness, in the sense of being awake versus asleep/comatose, include the highly interconnected *thalamocortical system* (Llinas Ribary, Contreras, & Pedroarena, 1998; Min, 2010; Newman, 1997; Ward, 2011) and the *ascending arousal system* (Posner, Saper, Schiff, & Plum, 2007; Steriade, 1996), which span a large portion of the brain. Fig. 2 summarizes the neuroanatomical organization of these systems in a simplified fashion. The ascending arousal system (located in the midline upper brainstem) and several nuclei in the thalamus (located near the midline just above the brainstem) provide diffuse activation of the cerebral hemispheres. This activation is a critical factor in conscious awareness. The pattern of activity of the cerebral hemispheres is closely related to the content of consciousness, but there is no specific region of cerebral cortex that, in isolation, is responsible for consciousness. Consistent with this picture, the types of brain damage that are correlated with loss of consciousness are localized damage in the ascending arousal system, bilateral thalamic damage, or extensive bilateral impairment of the cerebral hemispheres (Posner et al., 2007). In contrast, many other types of brain damage are not associated with loss of consciousness, including loss of sensory input, motor output, sensory-motor coordination (cerebellum), cortical language areas, pre-frontal attention mechanisms, or hippocampal regions serving memory (Koch & Tononi, 2008; Posner et al., 2007).

Contrastive analysis has demonstrated that several types of brain states correlate with consciousness. Functional imaging tests (PET, fMRI) have shown substantially lower metabolic activity throughout cortical regions, especially the pre-frontal and parietal cortex, in coma and during general anesthesia, relative to during consciousness (Baars, Ramsey, & Laureys, 2003). Further, more globally-distributed brain metabolic activity and increased communication between cerebral cortex regions is associated with effortful, conscious learning of new tasks, when contrasted with the more localized activity seen once a task is learned and essentially automatic (Baars, 2002; Haier et al., 1992).

Different states of the brain's electrical activity have also been found to correlate with the presence/absence of consciousness. For example, electroencephalographic (EEG) recordings from the scalp demonstrate widespread 10–100 Hz activity when someone is awake, but are often dominated by lower frequency activity ($\leq$10 Hz) during sleep or coma (Niedermeyer & Silva, 2005). Further, synchronization in gamma frequency activity (40–100 Hz), which can occur locally in an unconscious subject, can be much more widespread throughout the cerebral cortex during wakefulness, and various properties of this activity have been hypothesized to be a key neural correlate of consciousness and the unified subjective experience that it entails (Crick & Koch, 1990; Llinas et al., 1998; Min, 2010; Ward, 2011). In addition, electromagnetic stimulation of premotor cortical regions of the human brain in awake subjects is transmitted to other cortical regions, while the same stimulation during dreamless sleep is not, consistent with more global information processing during wakefulness (Massimini, Ferrarelli, & Huber, 2005). All of these functional imaging and electrophysiological correlates could be leveraged to sharpen current definitions of consciousness, at least in the sense that consciousness is considered in the clinical settings addressed by neurologists and anesthesiologists.

2.3. The physical basis of consciousness

At present, a typical functionalist view of the neurobiological basis of consciousness goes something like the following. The neural networks of the brain form an incredibly complex electrochemical system whose activity supports conscious memory and information processing. Certain functional states of this complex system, especially those involving global activity, support consciousness as an emergent property arising from the massively parallel computations that are taking place. The neurocomputational models reviewed in the remainder of this paper are largely cast within this framework.

However, while most past work on artificial consciousness is compatible with this view, it is important to recognize that there are alternative views. Many other theories have been developed about how physical processes in the brain may produce a conscious mind. For example, it has been hypothesized that the electromagnetic fields produced by the brain may be associated with consciousness rather than being just a byproduct of the underlying electrochemical activity of neurons (John, 2002; McFadden, 2000; Pockett, 2000), although at present there is no compelling evidence for this and some evidence that appears inconsistent with such an explanation (Pockett, 2002). More widely discussed is that quantum physical processes may be at play (Stapp, 1993). For example, it has been hypothesized that quantum effects might relate to conscious mind by influencing the release of synaptic vesicles (Eccles, 1994) or via effects involving microtubules internal to neurons (Hameroff & Penrose, 1996). While these latter hypotheses are intriguing, it remains unclear today whether or not quantum physics plays any special role at all in explaining conscious mind.

3. Past work on artificial consciousness

Current uncertainties concerning the nature of consciousness and its neurobiological basis form a major barrier to creating an artificial consciousness. For example, how will we even know (should we succeed) that a machine is conscious? This is not a problem specific to machines; it is a special case of the familiar *other minds problem* (Searle, 2004). This problem is that there is no objective way in which one can determine whether or not another person is conscious and has a mind. Rather than accept solipsism, most of us make the assumption that others are conscious based on analogy or a kind of parsimony principle: this is the simplest explanation for what we observe. This issue seems less obvious though when we ask whether animals are conscious, and if so, is there a way to characterize which living beings are conscious and which are not. This same difficulty occurs in studying machine consciousness (Prinz, 2003). While consciousness is sometimes characterized in terms of various defining features (qualitativeness, subjectivity, sense of self, unity, situatedness, etc. (Searle, 2004)) or even measurable human behaviors such as the Glasgow and other coma scales (Posner et al., 2007), trying to apply such criteria to machines in practice is problematic.

Accordingly, there have naturally been a number of efforts to specify criteria for the presence/absence of consciousness that can be applied to a machine. These criteria generally differ in spirit from the classic Turing Test for AI in that they explicitly involve "looking under the hood" at a system's internal mechanisms, rather than being based solely on behavioral criteria. For example, it has been suggested that there are five tests (''axioms'') that can be used to determine the presence of minimal consciousness, roughly involving testing for the existence of perceptual, emotional and imagined internal states, and the existence of attention and planning mechanisms (Aleksander & Dunmall, 2003). These top-down criteria have occasionally been used as a benchmark to assess past designs for machine consciousness (for example, Aleksander and Morton (2007) and Starzyk and Prasad (2011)), but it has been argued that they are really targets for conscious systems rather than true axioms (Clowes & Seth, 2008), and it has also been questioned whether they ''have anything to do with consciousness at all'' (Haikonen, 2007a, p. 196). Another ''consciousness test'' inspects a machine's architecture and functionality to determine that it is conscious if it has inner imagery and grounded inner speech, and if the machine can report that it has these phenomena, describe their contents, and recognize them as its own product, all without being pre-programmed to do so (Haikonen, 2007a). More recent proposals include a multi-level scale for comparing cognitive agents (Arrabales, Ledezma, & Sanchis, 2010), or assessing the similarity of spatiotemporal patterns of a machine's physical states relative to those of the normal adult brain (Gamez, 2012), in the context of machine consciousness research. While these and other criteria that have been proposed are both interesting and thought provoking, from an operational point of view, none are as yet generally accepted as objective tests that can be used in practice to determine the presence/absence of machine consciousness. Many of these criteria have been subjected to substantial criticisms (see, for example, Seth (2009)).

In characterizing past work on artificial consciousness, one can distinguish between two possible objectives of specific studies: simulation versus instantiation of consciousness. Such a distinction parallels the distinction made in Section 2.1 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 phenomena in meteorology, models of non-laminar flow in fluid dynamics, etc.). There is nothing particularly mysterious about such work; just as we would not expect to open a computer used to simulate a tropical rain storm and find it to be wet inside, we should not expect to open a computer used to model some aspect of conscious information processing and discover that ''it is 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 (Charkaoui, 2005; McCauley, 2007; Sanz, Hernandez, & Sanchez-Escribano, 2012). 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 that has led to considerable debate in the philosophical literature (Bishop, 2009; Molyneux, 2012; O'Regan, 2012; Schlagel, 1999). The dichotomy between simulated and instantiated consciousness is reminiscent of the distinction between weak AI (behavioral criteria) and strong AI (artificial mind), and has, by analogy, sometimes been referred to as distinguishing between weak artificial consciousness (simulated) and strong artificial consciousness (instantiated) (Seth, 2009).

Upon first examining past work that explicitly relates to machine consciousness, one is confronted with research that is based on a wide range of philosophical perspectives, that is driven by different objectives, and that makes use of differing computational methods. However, there are recurring themes that provide a rationale for organizing these past studies into five categories based on the fundamental issue that each adopts as most central to consciousness:

1. a global workspace,
2. information integration,
3. an internal self-model,
4. higher-level representations,
5. attention mechanisms.

In this review all past work in artificial consciousness has been classified into one of these five categories. While this classification is not perfect, it does group related work together naturally, makes apparent what has (has not) been done, and seems to naturally encompass almost all past studies related to artificial consciousness. For each of these categories, three topics are covered: an overview of past work involving this type of model, a more detailed example that illustrates the key ideas of the approach, and a commentary that assesses the contributions and limitations of the approach.

3.1. Global workspace models

3.1.1. Overview

One prominent approach to modeling neural correlates of consciousness has focused on the viewpoint that consciousness provides for global information processing in the brain. As summarized in Section 2.2, an increase in globally-distributed brain activity and inter-communication between regions of the cerebral cortex is well documented during conscious mental effort, consistent with this approach (Baars, 2002; Baars et al., 2003; Massimini et al., 2005).

Work in this area has been greatly influenced by Baars' *global workspace theory* (Baars, 1988, 2002). This theory views the human brain as organized into a network of specialized automatic processors that provide for sensation, motor control, language, reasoning, and so forth. Much of the processing in these modules is localized to specific brain regions and taken to be unconscious. However, in addition, there is a global workspace that is widely distributed throughout the brain, especially the cerebral cortex, and whose contents are available to (''broadcast to'') the specialized processors. The specialized processors compete to gain access to this interconnected global workspace, with such access allowing them to send/receive globally available information. In this context, conscious experience is hypothesized to emerge through the collective interactions between the specialized processors via the global workspace. More specifically, information in memory reaches consciousness when the amount of activity representing that information crosses a threshold.

Global workspace theory directly inspired an early computational model of consciousness known as IDA (Intelligent Distributed Agent) (Baars & Franklin, 2007; Franklin, 2003; Franklin & Graesser, 1999). IDA is not a neural network model. It is a multi-agent system consisting of ''codelets'' (Java processes) executing in parallel; these agents represent the specialized processing modules defined by global workspace theory. For example, in an application of IDA to assigning naval personnel to new tasks via a natural language dialog (Franklin, 2003), agents serve to recognize specific portions of text, categorize them, contribute information to the global workspace, and perform conditional actions. The fact that the agents communicate with each other via a global workspace implies (according to Franklin et al.) that IDA is functionally conscious (Franklin, 2003; Franklin, Strain, Snaider, McCall, & Faghihi, 2012). Recent work within this framework has focused on creating an extended, more general version of the model named LIDA (Learning IDA), on analyzing its relationships to neurobiological phenomena (Baars & Franklin, 2009; Franklin, Ramamurthy, & D'mello, 2007; Franklin et al., 2012; Raizer, Paraense, & Gudwin, 2012) and on giving it a self-model (Ramamurthy, Franklin, & Agrawal, 2012).

In contrast to IDA, most work in this area has focused on creating and studying *neural global workspace models* that are implemented as neural networks. An early model like this was developed for solving a Stroop task (a color word naming task) (Dehaene, Kerszberg, & Changeux, 1998), and we will consider this model in more detail below. Subsequent models often incorporated more biologically-realistic features. For example, more recent versions of the approach used in the Stoop task model incorporate explicit oscillatory behavior and spiking neurons. They and other neuronal global workspace models have been used successfully to simulate the "attentional blink" in which the second of two rapidly presented stimuli does not reach conscious awareness (Dehaene, Sergent, & Changeux, 2003; Raffone & Pantani, 2010). Other related models have emphasized the central role of the thalamus and thalamocortical interactions in global workspace theory (Harth, 1995; Newman, Baars, & Cho, 1997), or the cycles of workspace information broadcasts that alternate with competitions between specialized unconscious processors for access to the workspace (Shanahan, 2006). More recent models have incorporated spiking neurons, and systematically explored the range of models/parameters that could support processing in a global workspace (Connor & Shanahan, 2010), or the use of that processing for robotic control (Shanahan, 2008). Collectively these modeling efforts suggest that the global workspace model is a robust approach to modeling conscious information processing and its correlation with empirical data.

3.1.2. Example

Based on the concept that global brain activation is correlated with conscious mental states, Dehaene and colleagues formulated a neural global workspace model (Dehaene et al., 1998; Dehaene & Naccache, 2001; Dehaene et al., 2003). This model was used to contrast neural activity patterns when performing "routine" tasks that in humans are fairly automatic and do not require conscious effort, versus when performing effortful, conscious processing tasks. The architecture of this model is based on a network of cerebral cortex regions, each region representing either a specialized processor or the global workspace. For example, an early version of this model was devised to simulate human performance on Stroop tasks, a widely used test in cognitive psychology. With a word-color Stroop task, a subject is typically shown a temporal sequence of words that name colors ("red", "green", …), with the words appearing in varying ink colors. The subject must read each name as it appears. With congruous word names and colors paired together (such as the word "red" displayed in red ink), human subjects are fast and accurate, but with incongruous words ("red" displayed in green ink) they become slower and more prone to making errors.

Fig. 3 illustrates the architecture of this neural global workspace model for simulating human performance on word-color Stroop tasks (Dehaene et al., 1998). A color name ("red" in Fig. 3) and/or its ink color ("blue") are presented as inputs to the two different cortical regions that initially process words and colors. These regions, which use a local representation of names and colors, connect in a one-to-one fashion with output response units. Specifically, each name/color unit connects directly to just the corresponding correct response unit. Specialized processor neurons send and receive connections from workspace neurons, the latter of which have widespread excitatory and inhibitory connections between each other. Such lateral connectivity between workspace neurons is configured so that only a single "workspace representation" (a meaningful set of active neurons) can be present at any given time. Reinforcement learning occurs based on reward-guided Hebbian weight changes, where reinforcement signals of ±1 depend on whether the model's response is correct/incorrect during training.

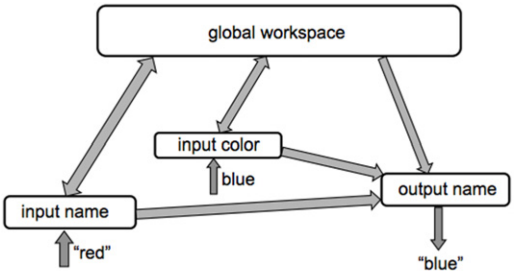

Fig. 3. Sketch of the architecture of a neural global workspace model for a word-color Stroop task (Dehaene et al., 1998). Each rounded box represents a cortical region, and consists of a set or layer of excitatory and inhibitory neurons having sigmoidal activation functions. Localized processing occurs in the special purpose input and output layers, while global processing occurs throughout the workspace layer. Some details of the model, such as a "vigilance node", are omitted for clarity.

This model was trained first on naming input colors presented alone (an easy task to learn), then on naming input words when incongruous input colors are presented simultaneously (relatively easy to learn, since word-to-name connections were stronger than color-to-name connections), and finally on naming input colors when incongruous input names are presented (taken to be a difficult, effortful task). During the first two easy tasks, there was relatively little activity in the workspace layer. In contrast, during the effortful task of naming input colors when incongruous input names are occurring, there was substantial but selective activation of workspace neurons as they learned to suppress activity in the special processor region that handles input words. Increased workspace activity persists until the task becomes "routine" with few errors being made, at which point workspace activity diminishes. The authors interpret the model as matching what happens with human subjects who have correlated widespread cerebral cortex activity during conscious effortful tasks, but not during easy/routine tasks that are largely carried out by unconscious brain processing.

3.1.3. Comments

Most past work on these models has focused on simulated rather than instantiated consciousness. In other words, these models have mostly examined the global workspace as a correlate of human consciousness rather than as an approach for creating machine consciousness. There is a lot to like about the basic hypothesis that a global workspace, distributed primarily throughout cortical regions, is a key correlate of conscious mind. It fits well with contemporary views in neuropsychology concerning the distributed nature of cognition over a network of structurally and functionally interconnected cerebral cortex regions (Shanahan, 2010; Sporns, 2011). It is also consistent with the substantial scientific evidence summarized above that widespread cortical activation correlates with conscious brain states. In addition, global workspace models provide at least a partial intuitive account for the unity of consciousness, and a clear statement about what distinguishes conscious from unconscious brain activity (the former is globally distributed, the latter is localized). The notion of a global workspace also fits well with current views of working memory in cognitive psychology, providing at least a basic account of the very limited capacity of working memory (Cowan, Elliott, Saults, & Morey, 2005) as competition for global representation, something that is otherwise surprising given the size and complexity of the human brain. It can even make predictions about abnormalities that might be expected with conscious machines. For example, mathematical analysis using the theorems of information theory suggests that machine consciousness based on a global workspace model will, after sufficient time, possibly lead to failure modes reminiscent of human inattentional blindness due to the absence of the corrective mechanisms that exist in biological systems (Wallace, 2005, 2006).

However, neural global workspace models of consciousness are not without limitations. While they capture and allow one to study an important aspect of consciousness, they have not yet shed significant light on *why* global information processing is a neural correlate of consciousness. There are many other multi-layer neural networks with architectures similar to global workspace models, such as neural blackboard architectures (van der Velde & de Kamps, 2006), and also AI blackboard models such as the HEARSAY system for speech interpretation (Erman, Hayes-Roth, Lessor, & Reddy, 1980). Some of these systems deal with low-level information processing tasks that occur largely at a subconscious level. Such models have generally *not* been advocated as relating to simulated consciousness, making it unclear what it is about the very similar neuronal global workspace models that justifies claiming that they represent simulated consciousness. Explaining more clearly what the global workspace specifically contributes to consciousness thus remains an important research direction for work in this area. Information integration provides one potential explanation.

3.2. *Information integration*

3.2.1. *Overview*

Another approach to studying machine consciousness has focused on the viewpoint that information processing and integration is central in explaining consciousness (Agrawala, 2012; John, 2002; Tononi, Sporns, & Edelman, 1994). Usually the term "information" in this context refers to classic Shannon information theory (Shannon, 1948), although this is not always the case (Agrawala, 2012; Sloman & Chrisley, 2003). The basic idea is that the shared or mutual information among brain regions that is above and beyond their information content as individual regions enables them to interact in a constructive fashion. This information integration is viewed as a neural correlate of consciousness. Accordingly, information theory has occasionally been used to interpret neurobiological data relevant to human consciousness (for a recent review, see Gamez (2010)).

Computational work in this area has been greatly influenced by *information integration theory* developed by Tononi and his colleagues (Balduzzi & Tononi, 2008; Tononi, 2004, 2008; Tononi & Sporns, 2003). This theory hypothesizes that consciousness is equivalent to the capacity of a system to integrate information into a unified experience and to differentiate among a large collection of different conscious states. The theory provides a quantitative measure ϕ of a specific network's ability to integrate information based on the calculated mutual information of all possible bi-partitions of the network. A network's ϕ value differs from earlier neural complexity measures in assessing causal interactions between network components, not just statistical dependencies (Tononi & Sporns, 2003).

Having a quantitative measure ϕ of a system's ability to integrate information is a useful concept. It (hypothetically) provides an objective measure of the extent to which a system has conscious experience, and it accounts for some important properties of consciousness: the availability of a very large number of conscious experiences, the unity of each experience, and several neurobiological observations concerning consciousness (Tononi, 2004). However, integrated information theory is limited in practical value due to the computational expense of measuring ϕ in any but the very smallest networks. Accepting this measure implies accepting that consciousness is not an all or none phenomenon, but instead is a graded quantity. Roughly, the intuition is that the larger the value of ϕ is for a network, the more the components of the network causally influence one another, and thus the more conscious it is. The concept that consciousness *is* integrated information is claimed to explain a number of experimental observations (for example, Massimini et al. (2005)), including the close relationship of conscious mind with the human thalamo-cortical system rather than other parts of the brain. Information integration theory makes other stronger claims, including that a system's subjective experience is equivalent to its capacity to integrate information, that the "informational relations" of a system's components define a space of qualia, and that a limited form of panpsychism holds (Tononi, 2004, 2008).

Given the practical barriers to measuring ϕ on any but the smallest networks, it is not surprising that relatively little computational work has been done yet within this framework, and that which has been done is fairly recent. One study has used attractor neural networks as associative memories that store images, based on weightless neurons, to examine how different connection patterns and architectures influence a network's estimated information integration (Aleksander & Gamez, 2009). Rather than computing ϕ on a state-by-state basis, this work attempts to monitor information integration over time via computational experiments. Effective information integration was found to be maximized by strong distributed connectivity in a network, an observation that is consistent with the ideas of global workspace theory discussed in the previous section. In a related subsequent study based on a more recent extension to integrated information theory (Balduzzi & Tononi, 2008), ϕ was systematically calculated for simple four neuron networks and compared to network *liveness*, another measure of network interactions introduced by the authors (Aleksander & Gamez, 2011). This latter work showed that network liveness is strongly correlated with ϕ in these small networks.

3.2.2. *Example*

Information integration theory claims that those portions of a neural network associated with high ϕ values are phenomenally conscious (Tononi, 2004, 2008). Starting with this hypothesis, Gamez developed a neurocontroller for a robotic vision system (Gamez, 2010). The goal of this project was for the neurocontroller to learn to look preferentially towards "positive" stimuli in the form of red blocks and away from "negative" stimuli in the form of blue blocks. Once the controller was trained, the distribution of information integration throughout the regions of the network was examined to assess the extent to which individual parts of the system could be considered to be conscious.

The architecture of the neurocontroller in this study is illustrated in Fig. 4. In this system, visual input from a movable robotic eye flows through a simulated visual system. The direction of the robotic eye's gaze is determined by a sustained activity pattern over the Motor Cortex region that acts through a multi-layer motor control system. The system learns to associate activity patterns in the Motor Integration region with blue/red objects in the environment by competitive Hebbian learning on the connections between the Motor Integration region and the Visual Association region. Once trained, built in mechanisms (not shown in Fig. 4) direct the system to move its gaze direction away from blue blocks and towards red blocks. Of special interest here are the Emotion and Inhibition regions that gate activity in other regions.[1] The Inhibition region, when active, effectively shuts off

[1] The name "Emotion region" was motivated by analogy with the neuromodulatory aspects of emotions and a priori theoretical considerations (Aleksander & Dun-

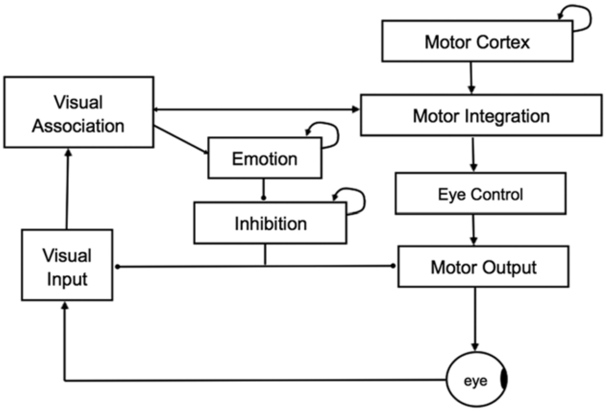

Fig. 4. Slightly simplified top-level architecture of a robotic vision system neurocontroller consisting of approximately 17 thousand spiking neurons and 700 thousand connections (Gamez, 2010). Boxes indicate layers/regions of simulated neurons, and arrows indicate connection pathways (those with round endings —● indicate inhibition). Images generated by the eye pass through the visual system (on the left), and the motor control system (right) directs eye gaze direction. The Motor Cortex, Emotion and Inhibition regions have recurrent, self-excitatory connections, enabling them to sustain activity in the absence of input from other layers. Neurons in the Motor Cortex and Inhibition regions are injected with noise (random activity) at each time step so that they maintain continuous activity.

the Visual Input and Motor Output regions via strong inhibitory connections, isolating the controller from the physical robot. When the controller is in this offline mode, the random activity patterns appearing in the Motor Cortex region generate potential eye gaze directions; this continues until the eye's intended viewing position would correspond to a remembered red block in the environment. At this point, the Emotion region is activated, and it in turn shuts off the Inhibition region, bringing the controller back online. This reconnects the controller to the environment and robot eye, causing the robot's gaze direction to shift towards the remembered red block.

Once the neurocontroller was successfully trained, its information integration properties were analyzed to determine "the distribution of consciousness" throughout its neural networks using the algorithm in Tononi and Sporns (2003). Direct use of this algorithm is computationally intractable for a network of this size (estimated to require 10^{9000} years on the available computing resources (Gamez, 2010, p. 30)). Accordingly, Gamez developed and validated an approximate algorithm for computing ϕ that is more computationally efficient and also removes some normalizing steps of the original algorithm that were inappropriate in this situation. According to integrated information theory, the subnetwork or "main complex" with the largest ϕ value is the conscious part of the system (Tononi & Sporns, 2003). In this specific neurocontroller, analysis unexpectedly identified the subnetwork with maximal ϕ as involving 91 neurons that included the entire Inhibition region, most of the Emotion region, and a few neurons each in several other regions. Further analysis calculating the "predicted consciousness per neuron" led to the conclusion that only the Emotion and Inhibition regions of the model would be significantly correlated with consciousness in this specific network.

3.2.3. Comments

The basic ideas of integrated information theory have been discussed for roughly a decade, and some of its concepts can be traced back even further to work on neural complexity measures and selectionist views of neural processing (Edelman, 1989; Tononi & Edelman, 1998; Tononi et al., 1994). However, development of neural models of consciousness specifically inspired by integrated information theory and related approaches is fairly recent. This theory is attractive in that it provides a specific, well-defined theoretical framework within which to study and quantitatively measure the extent to which a system, natural or artificial, is conscious. It seems probable that work within this context will expand in coming years. In particular, integrated information theory provides an objective, quantitative tool for assessing *other* theories about the nature of machine consciousness, and may become adopted for that purpose. For example, a reasonable question is how the theory's predictions match up with global workspace theory. While it seems intuitively plausible that a global workspace would contribute to information integration in a system beyond what its associated specialized processors alone provide, this issue is just beginning to be analyzed. On the other hand, the eye movement neurocontroller described above was unexpectedly found to have a subnetwork with relatively high ϕ that was interpreted (within information integration theory) as the most conscious component of the system, and this subnetwork is arguably *not* a global workspace. Interestingly, the high-ϕ components of this system were effectively serving as gating circuits. Gating in neural architectures has recently received increased recognition as important to cognitive control mechanisms for working memory, something that is closely associated with conscious mind and that has been the focus or recent computational studies (O'Reilly & Frank, 2006; Sylvester, Reggia, & Weems, 2011; Sylvester, Reggia, Weems, & Bunting, 2013). Such "high-level" gating, which regulates the flow of information between "low-level" regions of a model, can also be related to the higher order theories of consciousness discussed below (Section 3.4).

On the other hand, integrated information theory faces substantial philosophical and practical barriers as a guide to machine consciousness. From a philosophical perspective, and in spite of explicit claims that subjective experience is equivalent to the capacity to integrate information (Tononi, 2004), at the present time it remains an open question as to whether the measure ϕ has anything to do with the subjective experience that we associate with conscious mind rather than with the information processing properties of a system (conscious or not) (Manzotti, 2012). In significant ways, the ϕ score seems more directly related to a system's potential to exhibit intelligent information processing rather than subjective experience. For example, it has been shown that, with a suitable selection of synaptic weights, one can construct simple, fully connected neural networks having arbitrarily high ϕ values (Seth, Izhikevich, Reeke, & Edelman, 2006). According to integrated information theory, at some (unspecified) ϕ value such networks would therefore exhibit instantiated/phenomenal consciousness, a conclusion that is difficult for some to accept. Even if one does accept this conclusion, it is unclear that it sheds any light on the nature of consciousness (Seth, 2009). It has also been argued that this theory does not have some basic properties that any functional theory of consciousness needs to satisfy in order to be well formed, such as structural coherence (a correspondence between conscious experience and awareness) and organizational invariance (identical experiences across systems with the same functional organization) (Cerullo, 2011).

On the practical side, a major barrier to advancement in this area is that one cannot apply ϕ directly to neural networks of appreciable size. Further, a ϕ value alone does not really provide a meaningful indication of whether consciousness is present or not: without a golden standard such as the ϕ value for a typical

mall, 2003). The terms gate and gating used here and throughout this paper refer to neural elements that regulate the flow of activity into, through, or out of other brain regions, much as spigots can regulate the flow of water through a system of pipes.

human brain to use as a comparison, it is extremely difficult to interpret what any given value of ϕ actually means (Gamez, 2010). Such a measurement for real brains is currently out of reach, not only because of the size of the networks involved but also because our inability to precisely characterize the detailed neural circuitry of the brain. A critical direction for future research in information integration theory is the development of computationally efficient algorithms that can be used to approximate ϕ values for large neural networks, a topic on which some initial work has already been done (Aleksander & Gamez, 2011; Gamez, 2010). Such algorithms might then be applied to one or more of the large-scale brain models currently under development (reviewed in Cattell and Parker (2012) and Garis, Shuo, Goertzel, and Ruiting (2010)) to provide a point of comparison for ϕ values measured on other networks. In addition, integrated information theory itself clearly needs further theoretical development. Since it provides precise, objective definitions and quantitative methods, it would be particularly helpful to examine whether any useful theorems can be derived from these foundations, and whether they could lead to testable predictions that could be assessed, at least through computational models. Finally, given the emergence of quantum information theory in recent years, an important issue would be how integrated information theory can carry over into this framework, and in particular, how it would relate to current theories about the quantum nature of conscious mind (see Section 2.2).

3.3. Internal self-models

3.3.1. Overview

Self-awareness has long been viewed as a central, if somewhat poorly understood, aspect of conscious mind (Samsonovich & Nadel, 2005; Searle, 2004). Metzinger's philosophical analysis of the first-person perspective provides a useful framework in which to consider computational efforts in this area (Metzinger, 2000a). The key idea is that our mind includes a *self-model* that is supported by neural activation patterns in the brain. This self-model is based on an internal representation of the spatial properties of the body referred to as a *body image*, a concept that can be related to the topographic and feature maps found in sensorimotor cortex (Goldenberg, 2003; Silver & Kastner, 2009). The body image is a *virtual model* that, while in part innate, is highly adaptable. This is consistent, for example, with the observation that cortical sensorimotor maps can change dramatically depending on the sensory patterns being received or following cortical damage (Reggia, Goodall, Revett, & Ruppin, 2000). The body image's virtual nature is illustrated by astronauts in weightless situations who lose their subjective body axis (i.e., their spatial frame of reference), and phantom limbs where a person who has lost an arm still subjectively experiences the presence of the arm. Metzinger's argument is that the subjective, conscious experience of a "self" arises because the body image is always there due to the proprioceptive, somatic and other sensory input that the brain constantly receives, and because of the inability of our cognitive processes to recognize that the body image is a virtual representation of the physical body. In short, within this philosophical framework, self-modeling of the body as an agent that causes one's behavior leads to phenomenal subjectivity and conscious experience.

Building on these and similar philosophical arguments, there is a long history of suggestions by workers in AI concerning the importance of a self-model to machine intelligence (for example, Minsky (1968)). The key idea is that an intelligent agent has an internal model encompassing not just the external environment, but also including a model of itself, as illustrated in Fig. 5(a). In recent years, AI and cognitive science researchers have made a number of proposals for how subjective conscious experience could emerge in *embodied agents* from such self-models. One suggested possibility is that self-models could lead to "strong self-reference" whereby an artificial agent that reasons about itself can reason about its own reasoning about itself, leading to a "strongly self-referring loop" (Perlis, 1997). Another possibility is that phenomenal consciousness could emerge from introspective reasoning mechanisms about perception (McDermott, 2007). It has alternatively been suggested that some key properties of a self-concept (existence, continuity over time, supervenience on a physical substrate, etc.) can form the basis for self-modeling in intelligent agents, regardless of whether or not they are embodied (Samsonovich & Ascoli, 2005). Building on this latter idea, it has been proposed that introducing a self-concept during learning is the key event in creating "computational consciousness" in suitable cognitive architectures (Samsonovich, 2007; Samsonovich & DeJong, 2005).

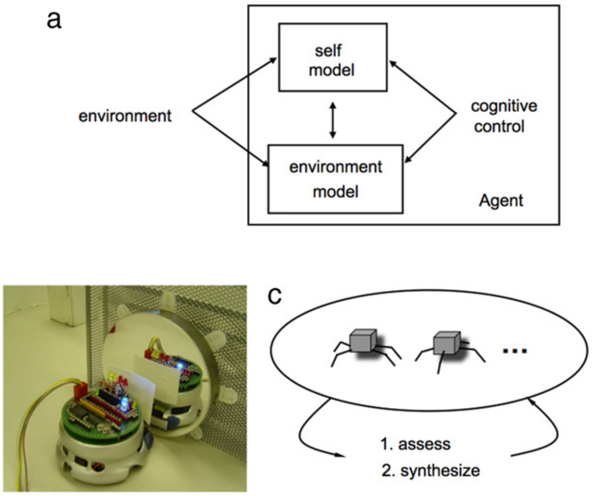

Fig. 5. (a) Architecture of an agent that has not only a model of the external environment, but also a self-model about which it can reason. (b) A small robot that is imitating the movements made by its own image in a mirror. Photo courtesy of Prof. Junichi Takeno. (c) An agent maintains a set of candidate self-models or body images (pictured inside oval). These models are repeatedly refined via cycles of assessing their correctness and then constructing revised models.

One specific type of internal model that is claimed to capture various features of consciousness is the *virtual machine architecture* (Sloman & Chrisley, 2003). In computer science/engineering, a virtual machine is a simulation of one machine, implemented in software, that runs on a different physical machine (hardware). This approach views a conscious human mind as a virtual machine (software) that is being executed by the brain (hardware), where, like the rest of the body, the brain is considered to be a machine. The "mental" states of the virtual machine are generally but not necessarily causally connected to one another, consistent with a functional view of consciousness. Self-modeling occurs in a suitable virtual machine when the machine develops concepts for categorizing and labeling its own states as sensed by internal monitors. In this context, qualia are defined to be what virtual machines "refer to when referring to the objects of internal self observation" (Sloman & Chrisley, 2003). In other words, qualia are viewed as being a side-effect of having an executive or meta-management component as part of the virtual machine that allows it to examine its own internal functioning and representations. These investigators recognize that many people would object to this characterization of qualia because it fails to address the hard problem of consciousness. Their response to this is "the fact that many people do think like this is part of what needs to be explained by any theory of consciousness" (Sloman & Chrisley, 2003).

Given the rich history of using internal models in robotics, it is perhaps not surprising that a number of other studies concerning self-modeling as the basis for artificial consciousness have been done in the context of physical or simulated robots. For example, the humanoid robot CRONOS has been studied in this fashion (Holland, 2007; Holland & Goodman, 2003). CRONOS' controller includes a self-model, implemented as a physics-based rigid body simulator, that interacts with a separate internal model of the robot's external environment. This self-model is intended to process sensory information and to move in a fashion identical to the physical robot, based on the state of the external environment model. The claim is that this process qualitatively reflects the cognitive contents of human consciousness, from a functional perspective, and that human phenomenal consciousness is rooted in a roughly equivalent internal self-model or body image in the brain (Holland, 2007).

Another robotics study has focused more explicitly on self-awareness (Katayama & Takeno, 2011; Takeno, 2011, 2013). This work, done over a period of several years, has studied "conscious robots" that are controlled by recurrent neural networks that undergo supervised learning. A central focus of this work is handling temporal sequences of stimuli, predicting the event that will occur at the next time step from past and current events. The predicted next state and the actual next state of the world are compared at each time step, and when these two states are the same a robot is declared to have instantiated consciousness ("consistency of cognition and behavior generates consciousness" (Takeno, 2011)). In support of this definition of consciousness, these expectation-driven robots have been shown to be capable of self-recognition of themselves in a mirror (Takeno, 2008). Such a demonstration was motivated by past studies involving a *mirror test* of self-awareness in which scientists examined whether animals recognize that the image they see in a mirror is themselves and not another animal of the same species (Gallup, 1970). In this case it was possible to have a small commercial robot (Khepera II) imitate the movements of its own image as reflected in a mirror (Fig. 5(b)). Because this was done more successfully than imitating the movements of other robots, the investigators claim that it represents self-recognition (Takeno, 2008). Other recent work has also demonstrated robot mirror self-recognition as well as self-recognition of a robot's own arm when observed visually, but this has been motivated by practical concerns rather than by issues related to artificial consciousness (Gold & Scassellati, 2007).

3.3.2. *Example*

Why might the brain have evolved to support self-models? While the usefulness of an internal model of the environment seems evident, it is less clear what advantages would accrue to an embodied agent from having a self-model. Efforts to examine this issue and to create robust robots include Bongard and colleagues' study of "resilient machines" that incorporate self-modeling (Bongard, Zykov, & Lipson, 2006). This work used a four-legged physical robot, where each multi-jointed leg has both touch and position sensors and is moved by multiple motors. Initially, the robot has no information about how its legs are connected to its body or about their lengths, and it does not know how to move its legs in a coordinated fashion to propel itself forward. It must learn a body image via a trial-and-error process of attempted movements.

The robot used in this study goes through repeated hypothesize-and-test cycles to construct its body image (Fig. 5(c)). Starting with a randomly generated action and the sensory feedback resulting from that action, a set of candidate self-models (body images) is generated—this is the hypothesis phase of the cycle. Each self-model is encoded as 16 parameters that indicate where its body parts are attached to one another. An action is then selected that is judged most likely to help the robot discriminate among the candidate self-models. Action selection is done via a search process during which the robot "imagines" (explicitly simulates) the expected sensory results that would occur with different actions, and then selects the action that is expected to produce the most disagreement among predictions made by the different candidate models. The physical robot then performs this selected action and receives sensory feedback—the test phase of the cycle. Based on comparisons between predicted and actual sensory results, the robot constructs a new set of candidate self-models by making random changes to the existing ones and keeping such modifications if they improve the model's prediction (a form of hill climbing). After a number of hypothesize-and-test cycles like this, the robot's ability to move effectively is assessed. This is done by generating a locomotion sequence that is executed by the physical robot and then measuring how far it actually moves within a given time period. Relative to baseline algorithms using random data acquisition, robot movement based on self-model-driven acquisition of sensory data was significantly more likely to produce successful robot behaviors. Similarly, following damage to the physical robot, such as reduction in a leg's length, the self-model-driven robot was better able to accurately model its altered form and to change its motor control mechanisms so that it could continue to move effectively.

Bongard and colleagues argued that the learning of predictive self-models in their robot could inform future investigation of higher levels of machine cognition (Bongard et al., 2006). Other discussion of this work has gone further in suggesting that this kind of self-modeling is an important step towards achieving self-awareness and consciousness (Holland, 2007). The periods of offline "imagination" of the consequences of actions based on the robot's self-modeling has even been interpreted as "robotic dreams" and related to the role of dream sleep in people (Adami, 2006; Conduit, 2007). The sense is that, as more neurobiologically realistic formulations of self-modeling emerge, it should be possible to use them to better understand the role of self-modeling in human conscious cognitive processes and the neural correlates of these processes.

3.3.3. *Comments*

The artificial systems described in this section have focused on awareness of self as a critical aspect of consciousness. These systems either explicitly incorporate a self-model, or they have properties that allow them to distinguish self from "other" and in that sense implicitly incorporate a self-model. In either case, a key aspect of a self-model is that it has predictive value, allowing an agent to set up expectations for what is about to occur, an ability that has previously been argued to be associated with consciousness (Ascoli, 2005). The concept that an internal self-model is a crucial aspect of consciousness is thus quite appealing. It is widely recognized in engineering that models of the external environment can be useful aspects of robotic control, such as in effectively navigating the environment or in recovering from unexpected changes. Work on self-models extends this recognition of the usefulness of models to include self-models (Hart & Scassellati, 2011). As we just saw, it has been possible to demonstrate, for example, that a self-model can enable a robotic device to restore functionality following sudden, unexpected damage. This is a significant observation because it establishes one potential way in which self-modeling can contribute to an embodied agent's fitness, and thus may help to explain in part why self-modeling (and by extension, consciousness) may have evolved. Such an idea is consistent with an earlier proposal by the evolutionary biologist Dawkins that evolution of the ability of the brain to simulate may have culminated in consciousness when such simulation of the world became sufficiently complex that it could incorporate a self-model (Dawkins, 1976, p. 63).

In contrast to the past work on global workspace models and integrated information that was described in previous sections, the work reviewed here has focused much more on producing more effective artificial agents than on explaining or understanding the neural correlates of consciousness. A major limitation of this work is that most of the self-models developed to date are quite simple and deal only with low level sensory and motor processing. Predictions have focused on what sensory input to anticipate next, or the outcome to be expected from taking a certain action. These types of predictions do not come close to capturing the richness of human meta-knowledge about other aspects of the self such as one's own cognitive processes. Important directions for future research thus include extending self-modeling to more complex cognitive processes, relating it to self-referential mathematical formalisms (Goertzel, 2011), understanding how self-modeling relates to the brain's ability to simulate behavior and perception (Hesslow, 2002; Hesslow & Jierenhed, 2007; Revonsuo, 2006), and in general to better relate work in this area to neurobiological correlates of consciousness.

3.4. *Higher-level representations*

3.4.1. *Overview*

In this section we consider an alternative view that conscious mental states are distinguished from unconscious mental states by their level of representation. The term *representation* is used here to mean any pattern of information or encoding that denotes something. This term is widely used in reference to computational work in cognitive science and related fields. For example, in symbolic AI, it is common to refer to the knowledge representation problem, which is concerned with which data structures (rules, causal associations, etc.) are best suited for expressing human expert knowledge in a form that can be processed by a machine. In neural networks, it is often said that a pattern of activity over a set of neurons provides a representation of some concept, and that the trained synaptic weights on the connections of a network provide a representation of what has been learned. Relating the opaque, distributed representations often learned by neural networks to symbolic representations that are more human-understandable is an active and challenging research problem (Huynh & Reggia, 2011, 2012; Monner & Reggia, submitted for publication).

Theoretical arguments and some early recurrent neural network models of conscious brain activity hypothesized that attractor states, certain specific neural activity patterns ("bubbles"), or even the explicit encoding of information in general, are associated with, or characterize, consciousness (Aleksander, 1996; O'Brien & Opie, 1999; Taylor, 1997). The difficulty with this hypothesis is that there are many physical systems that are generally not viewed as conscious that have such attractor states/activity patterns, leading to the suggestion that just representing consciousness in such a fashion is not helpful: it may relate to the content of consciousness, but it is defective in that it involves no "internal experience" (Taylor, 2003a). It has been suggested that some of this difficulty may be ameliorated by taking trajectories, rather than instantaneous patterns, in the state space of possible activation patterns as representational primitives for qualia and subjective experience (Fekete & Edelman, 2011). In contrast to these past viewpoints, the following models generally postulate that conscious mental activity uses a *higher level* of representation than unconscious mental activity. The distinction between higher and lower levels of representation can take different forms. We consider two particular versions of this functionalist approach to studying consciousness computationally: models based on symbolic versus neural levels of representation, and models based on the concept of higher-order thoughts.

The CLARION system illustrates the use of a representational difference approach to understanding consciousness (Sun, 1997, 1999, 2002; Sun & Franklin, 2007). CLARION is a complex cognitive architecture incorporating both declarative and procedural knowledge, a motivational subsystem representing goals and drives, and cognitive control mechanisms. The key point of interest here is that this system is explicitly intended to explain consciousness based on a distinction between local/symbolic versus distributed/neural representations. In other words, the components of CLARION are organized in terms of two levels of information processing, conscious and unconscious, that interact with one another. Conscious information processing is based on a *local* or high-level symbolic representation of information: each concept is represented by localized activation of an individual node in a network, and production rules are used to capture procedural knowledge. In contrast, unconscious processing is based on a *distributed* or low-level representation of information: a concept is represented by a sub-symbolic pattern of activity across a neural network, with many nodes participating in representing a single concept, and the activity patterns representing different concepts overlapping with one another. This representational distinction matches the fact that conscious aspects of cognition are directly accessible—they are transparent in the sense that they can be reported using language. In contrast, unconscious aspects of cognition are inaccessible in this same sense—they are opaque/implicit and cannot be reported through language. CLARION has been impressively successful in accounting for a variety of psychological data in different contexts (reaction time studies, grammar learning, solving Tower of Hanoi problems, and others) (Sun, 2002; Sun & Franklin, 2007). It provides a principled, explicit separation between conscious and unconscious information processing that distinguishes it from many other general cognitive architectures (reviewed in Sun and Franklin (2007)).

Other studies examining the prospects for creating self-consciousness in physical robots have also emphasized the importance of levels of representation. For example, Kitamura and colleagues studied a five-level "consciousness-based architecture" robotic controller where the lowest level is a reactive, reflex-based module, while the highest, most conscious level is a symbolic rule-based module that directs overall movement strategies (Kitamura, Tahara, & Asami, 2000). This approach was successfully implemented in a pair of Khepera robots and used to control their behavior in a pursue-and-capture situation. The claim is that a conscious self emerges at the highest levels of this robot when tasks become effortful. In other words, when automatic, mostly reflexive actions on the lower levels are blocked, the higher levels are brought into action to simulate conscious directing of effortful tasks. More recently, a comparable four-level cognitive architecture named CERA-CRANIUM was proposed and implemented using both simulated and physical Pioneer 3DX robots (Arrabales, Ledezma, & Sanchis, 2009). Although originally inspired by global workspace theory, these authors suggest that qualia and conscious experience arise just in the highest cognitive level of this architecture because, instead of directly accessing sensory data, it indirectly processes percepts and treats them as if they are spatially located in the outside world.

Chella and colleagues have also emphasized multiple levels of representation as they relate to perceptual information processing in physical robots (Chella, 2007; Chella, Frixione, & Gaglio, 2008). This work involves an operational mobile robot that serves as a museum tour guide. Three levels of representation are used: a "sub-conceptual area" concerned with low-level processing of sensor data, an intermediate "conceptual area" that organizes the lower level sensory data into conceptual categories and structured information, and a higher-level "linguistic area" based on first-order predicate calculus. The conceptual area serves as a bridge to ground[2] the symbols used in the linguistic area; it is pre-symbolic and organized as a metric space in which clusters of concepts can form. In contrast, the high-level linguistic space is a semantic network of symbols and their relationships related to the robot's perceptions and actions, and it is here that predictive logical inferences occur, setting up expectations for subsequent events in the conceptual area. These investigators claim that the source of self-consciousness in cognitive architectures like this are higher-order representations of a robot's low-level first-order perceptions of the external world (Chella et al., 2008). Higher-order representations at the linguistic level are taken to be meta-predicates that describe the robot perceiving its situation and actions.

The emphasis on higher-order representations of low level perceptions links the above studies to a rich history of philosophical discussions concerning the nature of mind. *Higher-order thought (HOT) theory* (Rosenthal, 1996), and other higher-order theories in philosophy (Carruthers, 2005; Hellie, 2009; Lycan, 2009), provide well-known functionalist arguments for casting the problem of separating conscious from unconscious mental states as being a problem of representational differences. Rather than viewing consciousness as being an intrinsic property of a mental state, HOT theory views it as a relational property: it depends on the relations between different mental states. Specifically, HOT theory postulates that a mental state *S* is conscious to the extent that there is some other mental state (a "higher-order thought") that directly refers to *S* or is about *S*. This idea is hierarchical, or transitive, in the sense that a higher-order mental state might itself also be conscious if there is some other mental state that directly refers to it. It has been suggested that artificial systems with higher-order representations consistent with HOT theory could be viewed as conscious, even to the extent of having qualia, but only if they use grounded symbols (Rolls, 1997, 2007).

Recent work has begun to investigate such possibilities using neurocomputational models. More specifically, HOT theory has served as a framework for developing *metacognitive networks* in which some components of a neural architecture, referred to as higher-order networks, are able to represent, access, and/or manipulate information about other parts (lower-order networks) (Cleeremans, Timmermans, & Pasquali, 2007). In practice, this has involved architectures having a first-order network that performs a primary task such as pattern classification, and a second-order network that observes states of the first-order network as the basis for making decisions on a secondary task. It is the second-order network's learned reference to the first-order network's internal representations that makes the latter functionally conscious, according to HOT theory. A simple example of this kind of model is provided by an error backpropagation network that has the primary task of classifying input images of simple numeric digits (Cleeremans et al., 2007). The architecture of this model is shown in Fig. 6(a). The second-order network receives inputs solely from the hidden layer of the first-order network, and learns to output whether the first-order network's answer for the current input pattern is correct or not. In effect, the extent to which the second-order network's output is accurate is interpreted to be the extent to which the system is "consciously aware" of the first-order network's internal representation. The second-order network learns simultaneously with the first-order network, initially adopting a strategy of always deciding that the first-order network is incorrect. As the first-order network increasingly learns the primary task, the second-order network's initial strategy performs poorly until it ultimately learns features of the first-order network's hidden layer representations that discriminate correct from incorrect digit classifications. These investigators claim that "conscious experience occurs if and only if an information processing system has learned about its own representation of the world" (Cleeremans et al., 2007).

3.4.2. Example

Metacognitive nets have been applied to several tasks, of which we consider one here to illustrate the potential effectiveness of this approach. Our example models experimental data related to blindsight (Pasquali, Timmermans, & Cleeremans, 2010). The oxymoron *blindsight* refers to a syndrome in which individuals with damage to their primary visual cortex are able to respond correctly to visual stimuli to which they are blind, i.e., to which they cannot consciously see Weiskrantz (2009). Such individuals will be unable to verbally report properties of stimuli, such as location or movement direction, in their blind visual field, but when forced to guess can be remarkably accurate. It is hypothesized that blindsight occurs due to subcortical pathways that bypass primary visual cortex and transmit visual information to other downstream cortical visual areas, weakly activating these regions relative to when the primary visual cortex is intact.

In a recent behavioral study, a subject with blindsight made wagers about the correctness of his "guesses" concerning the presence/absence of visual stimuli (Persaud, Mcleod, & Cowey, 2007). This post-decision wagering was intended to be an objective measure of the subject's conscious awareness of his own performance. As might be expected, the subject's wagers were usually optimal[3] for suprathreshold stimuli in intact portions of his visual field, and usually arbitrary for subthreshold stimuli (i.e., optimal wagers occurred at roughly a chance level in the latter situation even though the subject was frequently correct about the occurrence or not of stimuli). Pasquali and colleagues used a metacognitive architecture structured as shown in Fig. 6(b) to simulate such post-decision wagering (Pasquali et al., 2010). The hidden layer in this model's second-order network (comparator layer in Fig. 6(b)) was atypical in that it consisted of nodes that received connections from corresponding pairs of input and output nodes in the first order network. The comparator layer's incoming connections were hardwired and non-adaptive. Activity patterns over the comparator layer were interpreted as meta-representations and argued to play a crucial role in the emergence of conscious percepts. The first-order and second-order networks of the intact model were trained simultaneously, the former to correctly localize input stimuli and the latter to produce optimal wagers (decisions about decisions). To subsequently simulate a blindsight subject, subthreshold stimuli were modeled using stimuli that were degraded by the introduction of noise. The model's performance reasonably matched that of the human blindsight subject: wagers were usually optimal for suprathreshold stimuli, and usually not for subthreshold stimuli (i.e., optimal wagers occurred at chance levels). While the investigators do not claim their model has instantiated conscious in any sense, they do conclude that the higher-order representations formed in this and their other metacognitive models capture the core mechanisms of HOT theories of consciousness (Pasquali et al., 2010).

3.4.3. Comments

Distinguishing between conscious and unconscious mental states based on level of representation relates to a number of past

[2] To ground a symbol means to associate it with the object/event that it refers to in the external world (Harnad, 1990).

[3] Optimal here means that the subject wagered high when correct on the primary task, and low when incorrect on the primary task.

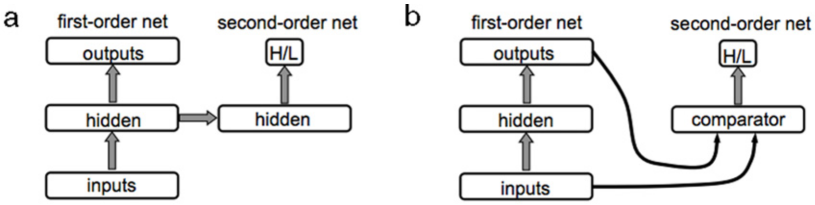

Fig. 6. Two examples of some basic metacognitive neural architectures. Boxes indicate network layers. Broad gray arrows indicate all-to-all connections between layers whose weights change during learning, while thin black arrows indicate localized non-adaptive connections. (a) The first-order network on the left is a feedforward error backpropagation network, often used for pattern classification tasks. The second-order network on the right has no causal influence on the first-order network's performance; instead, it monitors the first-order network's hidden layer representation and makes a high (H) or low (L) "wager" as to the correctness of the first-order network's output for each input pattern that it receives. The correctness of this wager over time reflects the extent to which the second-order network has learned, also via error backpropagation, a meta-representation of the first-order network's representation. (b) A metacognitive architecture used to model experimental findings obtained from a human subject who had the clinical syndrome known as blindsight.

theories in philosophy about the nature of consciousness (Conrad & Sun, 2007), to widely discussed issues surrounding symbolic information processing in both natural and artificial neural networks, and to formal models of representational relations in theories of consciousness (Bosse, Jonker, & Treur, 2008). For example, there are well known opinions that consciousness is based on language (Jaynes, 1976), and is thus a very recent event in evolutionary history that, by implication, is not present in animals. The high-level syntactic manipulation of symbols in language processing has also been argued to be closely related to HOT theories (Rolls, 1997, 2007). However, given the strong left hemisphere specialization for language in most people, the hypothesis that language is a fundamental basis of consciousness is difficult to sustain: there is substantial evidence that the right hemisphere as well as the left supports consciousness (reviewed in Keenan, Rubio, Racioppi, Johnson, and Barnacz (2005)), and sudden severe damage to the left hemisphere with accompanying aphasia is no more likely to be associated with loss of consciousness than is comparable right hemisphere damage (Posner et al., 2007, pp. 151–152,). While a number of computational models of hemispheric specialization have been studied (Reggia & Levitan, 2003), only one has explicitly examined issues related to consciousness (Cook, 1999). This latter work suggested that, while aspects of conscious information processing are represented bilaterally in the brain, left hemisphere activity can be viewed as the "nucleus" of consciousness, while right hemisphere activity is the "fringe" of consciousness (Galin, 1996).

Linking the study of artificial consciousness to HOT theories and to other related higher-order theories (Carruthers, 2005; Hellie, 2009) ties this work to a rich history in philosophy of mind that has attempted to deflate the mysteriousness of consciousness. Metacognitive networks based on post-decision wagering provide a well-defined, concrete framework within which to study the implications of HOT theories, regardless of the extent to which such wagering is or is not a direct measure of awareness (Persaud, Mcleod, & Cowey, 2008; Seth, 2008). Further, HOT theories provide a context in which self-modeling as the basis for consciousness can also be examined. Hints of such a unification of HOT and self-modeling approaches to consciousness can in fact already be found in some of the work discussed earlier in this review (Chella et al., 2008; Cox, 2007; Gordon, Hobbs, & Cox, 2011; Samsonovich & Ascoli, 2005). Focusing on level of representation also links the models discussed in this section to a large AI literature on metareasoning/metacognition, although researchers in this latter area have been motivated more by trying to improve machine reasoning abilities rather than by explicitly trying to understand consciousness (Cox, 2011; Hart & Scassellati, 2011).

Models of conscious information processing based on level of representation also face some important challenges. For example, contemporary computers are generally believed by many people not to be conscious, and yet symbolic processing is widely done on computers, suggesting (according to some of the theories described in this section) that they are conscious. A similar situation occurs with HOT Theory: modern computers have monitoring systems that keep track of their internal states, indicating that they are conscious according to HOT theory (Rey, 1983). Does HOT theory really solve the hard problem of consciousness, as is sometimes claimed (Rolls, 1997, 2007), being able to account for the existence of qualia? At present this remains controversial. A recent review of higher-order representation theories of consciousness has discussed several objections to such theories (Lycan, 2009). A particularly important direction for future research is to make testable, falsifiable predictions based on the kinds of models discussed in this section, predictions that can address such past critiques. It is also important to get a better understanding about how different levels of representation can interact effectively. This latter issue is closely coupled to active research topics in cognitive science and AI concerning symbol grounding and how compositional symbol processing can be mapped onto neural networks that use a distributed representation of information (Chella, 2007; Chella et al., 2008; Monner & Reggia, 2011, submitted for publication).

3.5. Attention mechanisms

3.5.1. Overview

At each moment in time, a person is only consciously aware of a fraction of the ongoing stream of visual, auditory, touch and other sensory information being received. The *attention mechanisms* that actively select what we attend to have the ability to focus on behaviorally important aspects of sensory information, selecting, for example, which specific seen object to attend to based on its location in space, what type of object it is, and/or its general properties. A very large amount of research in cognitive psychology has demonstrated that attention mechanisms are controlled by both bottom-up (exogenous) and top-down (endogenous) processes in the brain that operate over many levels, from subcortical nuclei through a network of interconnected cortical regions (Buschman & Miller, 2007; Corbetta & Shulman, 2002; Crick, 1984; Kastner, 2009; Prinz, 2003; Shipp, 2004). Experimental studies, using techniques that range from single-neuron recording to functional imaging, have consistently shown that these attention mechanisms modulate incoming sensory information, attenuating irrelevant neural activity and accentuating activity representing external objects that are the focus of attention (Kanwisher & Wojciulik, 2000; Reynolds & Chelazzi, 2004).

Attention and conscious awareness are not equivalent processes, but normally they are closely linked and highly correlated (Koch & Tsuchiya, 2006; Lamme, 2003; Treisman, 2009). It is therefore not surprising that several studies related to artificial consciousness have concentrated on modeling attention mechanisms as the basis for developing a better understanding of consciousness. Such work has been done from two perspectives. First, there are models that treat attention mechanisms, in and of themselves, as representing conscious information processing, effectively equating a system's attending to a topic and the system being conscious of that topic at a functional level. Second, and in contrast, there are other models that identify a specific aspect or component of attention mechanisms as being responsible for any associated conscious awareness.

We first consider several models that effectively take attention mechanisms to be the functional basis of consciousness. For example, an approach like this was used by Tinsley (2008) who describes a computational model in which an attention network's output is simply taken to be a conscious representation of a stimulus. The model consists of a multi-layer network of topographic brain regions composed of simulated spiking neurons. Different sensory modalities are processed in parallel, culminating in a "selector super-model region" that provides the model's output. Sensory modality and/or stimulus location can be used to select what part of the input activity pattern at any moment reaches the model's output and thus enters consciousness. While the implementation of this model is fairly abstract and limited in size, it illustrates how attentional gating mechanisms can act to select the content of "conscious representations" in topographically-structured networks.

Haikonen (2003, 2007a, 2007b) offers a very different design for attention mechanisms and their relationship with consciousness. This approach proposes a conscious machine based on a number of properties, including inner speech and the machine's ability to report its inner states. The machine is composed of a collection of inter-communicating modules that are similar to the unconscious specialized processors in global workspace theory, but in this case there is no separate global workspace. Instead, the specialized modules communicate with one another, operating in the same fashion regardless of whether the overall system is considered to be conscious or unconscious. The distinction between a topic being conscious or not is based on whether the machine as a whole attends to that topic, as follows. At each moment of time, each specialized module of the machine can broadcast information about the topic/problem it is addressing to all of the other modules. Each broadcasting module competes to recruit other modules to attend to the same topic that it is processing. In effect, the overall machine's attention mechanism is not captured in a separate localized module as is often the case in other attention models, but instead is based on this distributed competitive determination of what the specialized modules collectively work on. The subject of this unified collective focus of attention is claimed to enter consciousness. Specifically, "The machine becomes 'conscious' about a topic when the various modules cooperate in this way in unison and focus their attention on the percepts about this topic" (Haikonen, 2007a, p. 187). Recently implemented in a robot (Haikonen, 2012), this design suggests that a machine can only be conscious of one topic at a time, and that it can be self-aware if its collective of modules are all working on a topic that involves the machine itself. Like global workspace theory, this model is consistent with neurobiological evidence that conscious brain states are associated with global communication between cerebral cortex regions (Massimini et al., 2005).

Recently a similar model of consciousness based on attention mechanisms has been proposed (Starzyk & Prasad, 2011). This model addresses embodied intelligent agents such as those typically modeled in AI as biologically-inspired cognitive architectures. The claim in this case is that an embodied agent must have a central executive system that monitors all aspects of a machine's behavior, including unconscious aspects, in order for machine consciousness to occur. Crucially, the appearance of consciousness is claimed to depend on the executive's attention mechanisms; the control of attention switching between topics is the "core mechanism for conscious selection of the attention spotlight and this drives the machine consciousness" (Starzyk & Prasad, 2011, p. 269). This framework is reminiscent of Haikonen's model in that there is no central location of decision making, and in the use of competing signals/messages between the physically distributed modules that direct the machine's behavior. While this design has also not been implemented, this model is claimed to be in agreement with previously published tests for machine consciousness (Aleksander & Dunmall, 2003).

Another related study has addressed how control systems must deal with extensive information processing and attention control demands, emphasizing that any such system must cope with having limited information processing resources (Coward & Gedeon, 2009). The point is that this issue of "bounded rationality" must be faced by any potentially conscious machine, and that it constrains the kinds of architectures that are appropriate to consider (modular, hierarchical, etc.). A multi-layer *recommendation architecture* modeled after cerebral cortex structure is suggested as a general remedy to this problem, and it is argued that this architecture may contribute to understanding self-awareness and the existence of qualia. In particular, past approaches based on global workspace theory and virtual machines are criticized for not adequately addressing the issue of limited resources, especially given the importance of learning to conscious cognitive processing.

While the models we have just considered effectively equate a system attending to a topic and that topic's access to conscious processing, other attention-based models have instead identified some specific aspect or component of an attention mechanism that is responsible for a topic entering consciousness. An early model like this addressed the attention mechanisms involved in classical conditioning, claiming that the different processing afforded to novel stimuli is the key factor in a stimulus entering conscious awareness (Gray, Buhusi, & Schmajuk, 1997). In contrast to the automatic unconscious processing of familiar stimuli, a novel stimulus activates specific neural circuitry forming a separate *novelty system* that increases the attention system's activity and accelerates learning. This transition from low to high levels of attention is taken to be the transition from unconscious to conscious modes of processing. This model has been mapped onto neuroanatomical structures, and it has been related to latent inhibition occurring during classical conditioning and to the cognitive abnormalities that are characteristic of schizophrenia (Gray et al., 1997).

More recently, Kuipers (2005, 2008) has proposed a model in which conscious processing arises from the *symbol grounding* aspect of attention mechanisms. Any cognitive agent, be it a robotic controller or the human brain, faces a major problem in selecting which portions of the continuous, massive amounts of sensory data being received deserve attention. For an AI agent using symbolic memory storage and inference methods, such selective attention can be implemented via "trackers". A tracker is a symbolic pointer into the overwhelming data stream ("the fire hose of experience") that maintains over time a correspondence between a high-level, symbolically represented concept and its continually changing low-level representation in the data stream. In effect, the part of the attention mechanism that performs symbol grounding, anchoring structured symbolic representations to selected spatio-temporal segments of the sensory data stream, is taken to be responsible for consciousness. The claim is that any system organized in this fashion, having both bottom-up and top-down attention mechanisms that create trackers along with a reasoning system of control laws that makes use of these grounded symbols, is genuinely instantiated conscious, has subjective experiences corresponding to qualia, and has a sense of self-awareness (Kuipers, 2005). While a system like this has not been implemented yet, the plausibility of its basic concepts have been evaluated by assessing their ability to meet the eleven criteria that Searle (2004) has argued any philosophical–scientific theory of consciousness should satisfy (Kuipers, 2008). The conclusion of this evaluation was that the essential features of consciousness can, in principle, be implemented in an embodied machine that has sufficient computational power (Kuipers, 2008). However, it has recently been argued that the dimensionality reduction implied by symbolic trackers is just the opposite of what is needed to create phenomenal machine consciousness (Chella & Gaglio, 2012).

Finally, two separate studies have suggested that the *efference copy* (or *corollary discharge*) associated with top-down control mechanisms is a vital feature of the neural mechanisms underlying consciousness. An efference signal or corollary discharge is a copy of the output of a control system that is fed back to input regions of the control system, as explained further below. Cotterill (1995, 1996, 1997, 1998) has long theorized that efference copies, generated by premotor and supplementary motor cortex and sent to sensory regions of posterior cortex, are a critical aspect of information being attended to and becoming conscious in the brain. This theory is largely based on efference copies of muscle control signals, and was a key feature of the design of CyberChild, a system intended for studying consciousness whose complex neural controller was strongly based upon neuroanatomical structures of the human brain (Cotterill, 2003). Independently, similar ideas derived from engineering control theory rather than neuroanatomical considerations were investigated, and they were extended from motor control mechanisms to attention control mechanisms by Taylor (2003b).

3.5.2. Example

Over the last decade, John Taylor and his colleagues studied a neurocomputational model of how consciousness could arise from attention mechanisms (Taylor, 2003a, 2007, 2012). This model, known as the Corollary Discharge of Attention Movement (CODAM) model, is founded on adopting an engineering control theory perspective of attention. Basic control theory, with its use of feedforward and feedback modules, has been very successful in general at creating effective neurocomputational motor control systems for directing robot movements (Norgaard, Ravn, Poulsen, & Hansen, 2000; Oh, Gentili, Reggia, & Contreras-Vidal, 2012) and also for understanding brain mechanisms of motor control (reviewed in Taylor (2003b)). Based on the close relationship between attention and consciousness, Taylor hypothesized that these same engineering approaches, converted to serve as models for controlling attention rather than physical movements, could generate insight into the neurobiological basis of consciousness.

Fig. 7(a) shows a generic scheme for the brain's arm movement control as it is viewed from an engineering perspective. An executive system (pre-frontal cortex) generates a target arm movement that is converted by an "inverse model" into control signals that direct appropriate arm muscle contractions. Proprioceptive and visual feedback to a monitoring system specifies the actual movement that results. In addition, an efference copy of the outgoing control signals, serving as the *corollary discharge*, generates feedback to the monitor specifying the predicted resulting movement. An advantage of the corollary discharge feedback via a forward model in engineering control systems is that it allows the controller to react much faster to errors than were it to wait for the actual sensory feedback which is somewhat delayed. There is substantial evidence that such a forward model for motor control is used in the brain (Taylor, 2003b), and it is this motor corollary discharge that was earlier suggested to be critical to consciousness (Cotterill, 2003).

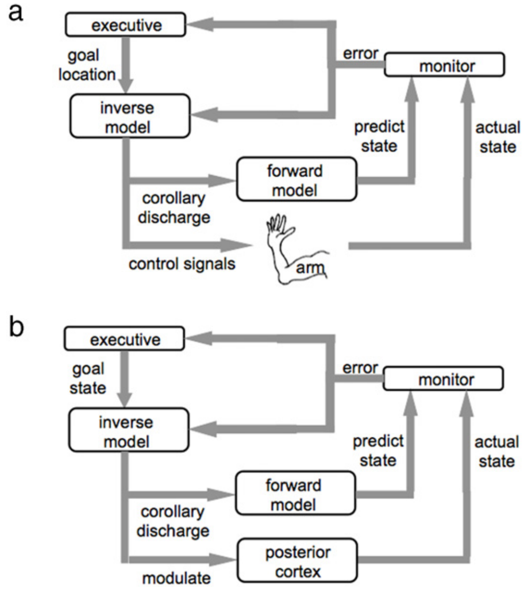

Fig. 7. (a) An engineering view of the brain's arm movement control mechanisms. See text for details. (b) The CODAM model of attention control, slightly simplified and arranged to illustrate how its structure is analogous to the motor control system in (a).

The hypothesized CODAM model is intended to be an analogous model for controlling changes in the focus of attention, or "attention movements", rather than changes in motor control (Taylor, 2003b, 2007). Fig. 7(b) shows a slightly simplified version of the CODAM model of the brain's top-down attention control mechanisms, where the components are arranged to emphasize the analogy with the brain's more established motor control system. The executive system (dorsolateral pre-frontal cortex) sends top-down control signals via an inverse model (parietal cortex) that selectively modulate posterior sensory cortex input, amplifying inputs deserving of attention and reducing other inputs that are to be filtered out of the sensory stream. Taylor hypothesized that, just as with physical movement control, the inverse model produces a corollary discharge that quickly tells a monitoring system (cingulate gyrus and parietal lobe) the predicted changes in the focus of attention. The corollary discharge is postulated to redirect the focus of attention as needed, and to contribute to error correction.

Implementations of the CODAM model have been applied to a number of tasks used by cognitive psychologists to study human attention mechanisms. In several cases the results produced, such as when modeling the attentional blink or change blindness, have been qualitatively similar to those seen with human subjects, providing support for the basic assumptions of the model (Taylor & Fragopanagos, 2007). Most recently, an analysis of fMRI, MEG and EEG activity during various attention-related tasks concluded that there was substantial support for the existence of a corollary discharge signal in the human brain (Taylor, 2012).

The CODAM model is a well-developed theory of the cognitive control mechanisms for attention, but how does it relate to consciousness? The basic argument is that consciousness is specifically due to the corollary discharge signal (Taylor, 2003b, 2007). The central claim is that this signal, buffered in the monitor, is directly responsible for the conscious experience of ownership of actions that control one's attention to various topics. Such ownership can be viewed as "the most primitive form of self-knowledge" (Taylor, 2003b). It provides a sense of agency—a sense that the self is responsible for the actions being taken by one's own cognitive control mechanisms. In this way the CODAM model provides a bridge to the other models of consciousness discussed earlier that are based on self-models, suggesting a mechanism by which the self's perception of ownership and agency might arise.

3.5.3. Comments

The computational models of attention mechanisms described in this section differ from a number of other past neural models of attention in being related explicitly to conscious information processing by their developers. To date these studies have been fairly evenly divided between those that are directed towards improving our understanding of natural consciousness and those that are focused on developing methods to produce artificial consciousness in machines. Given the central role of neurobiological mechanisms for attention in determining the contents of human consciousness, these models are intuitively plausible: they attempt to leverage our improving scientific understanding of attention control to get a handle on the fundamental nature of consciousness, and may provide a way to relate consciousness to learning (Bryson, 2012). This strategy has been reasonably effective and has served as the basis for some significant results as outlined above. Arguments have been developed concerning the role of symbol grounding implemented as part of attention mechanisms in bridging lower and higher levels of representation and in accounting for intentionality and self-awareness. Further, the idea that copies of top-down attention control signals, or corollary discharges, are a neural correlate of consciousness has been proposed as an explanation for the sense of ownership that we experience concerning our own thoughts and actions.

As with the other approaches to artificial consciousness that we have considered, work in this area has some substantial limitations. Some of the designs presented, while intriguing and creative, have not been implemented or tested against any sort of empirical data. Theories that essentially equate attention to consciousness do not yet provide a convincing explanation for why selecting information for representation as an activity pattern over a neural network makes that information conscious. More specific models that identify symbol grounding or corollary discharges as explanations for subjective experience fare better in this regard. However, while such features may correlate with intentionality, sense of ownership, and self-recognition, it remains an open question as to whether they actually have any causal role in creating conscious experience. Future work is needed to more clearly distinguish the selective information processing brought about by attention mechanisms, something that can plausibly be viewed as unconscious, from the properties of attention that give specific information access to conscious awareness.

4. Discussion

The pioneering work surveyed in this review has been done in spite of ongoing controversy about the nature of consciousness and our inability to even define consciousness adequately at present. While similar definitional issues exist with "intelligence" and "life" in the fields of AI and artificial life, this has not prevented substantial advances in those fields. However, consciousness remains much more mysterious today than these other concepts, and we do not yet have clear operational criteria for identifying the presence of machine consciousness (such as an analog of the Turing Test for machine intelligence). It even remains difficult to discuss consciousness, with great care needed in providing a precise and objective language that can handle subjective aspects of the topic. For example, arguments have been made that terms such as "phenomenal consciousness" are semantically flawed and thus unsuitable as a target of scientific research (Sloman, 2010). Those arguments build on suggestions that, regardless of terminology, the objective methods of science will never be able to unravel the basis of consciousness because of its subjective nature (McGinn, 2004).

In spite of these difficulties, work on artificial consciousness during the last two decades has been quite substantial. This review found that most of this work, regardless of whether it concerns simulated or instantiated consciousness, adopted one of five approaches/hypotheses as the key concept to use in creating computational models of consciousness. These approaches are the use of a global workspace to coordinate information processing by specialized modules, maximizing the capacity of a system to integrate information, development of a self-model that can be referenced by an agent's cognitive system, use of a high-level representation for information, and adopting or extending attention mechanisms. Each of these approaches corresponds to a neural, cognitive or behavioral correlate of consciousness, and each represents a theoretical position about the fundamental importance of that correlate. Interestingly, these approaches are not mutually exclusive possibilities, but instead can be viewed as complementary hypotheses about how to get a handle on producing machine consciousness.

4.1. Conclusions

Based on this survey of past work involving artificial consciousness, three main conclusions seem appropriate.

First, *during the last two decades computational modeling has become an effective and accepted methodology for scientifically studying consciousness*, complementing more traditional approaches involving cognitive science and neuroscience. This is a major achievement, given the existing barriers to the study of consciousness in general that were outlined earlier in this paper. Like computer modeling in other fields, modeling the correlates of consciousness forces one to be very explicit in formulating a theory for implementation as a formal machine algorithm, it can reveal unexpected implications of research hypotheses, and it allows one to inspect the details of the underlying information processing that produces observable behaviors in ways that are not practical in cognitive science and neuroscience. While the field of artificial consciousness is still in its infancy, it is becoming respectable, with growing interest in its results.

Second, *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. Scientifically this is extremely important. For example, it is providing a way to test whether theories about key neural correlates of consciousness, when implemented as computer models, can produce results in agreement with experimental data. In addition to this scientific significance, the development of simulated consciousness represents important progress towards producing machines that can exhibit external behaviors that are associated with human consciousness. For example, advances that have increased our capabilities to produce simulated consciousness in machines include new methods for automated reasoning about self-models, attention switching based on outcomes that do not match expectations, the development of robotic self-recognition, and recognition of the importance of symbol grounding as a bridge between low level neural computations and higher level symbolic reasoning. One does not have to believe that machines are "really conscious" to recognize that such advances may lead to future artificial agents that can reason more effectively and interact with people in more natural ways. Machines that are functionally self-aware or that can simulate a theory of mind could ultimately produce much more human like intelligent behavior than those that currently exist. It thus seems likely that simulated consciousness will play a significant role in future work on creating an artificial general intelligence.

Third, *at the present time no existing approach to artificial consciousness has presented a compelling demonstration of instantiated* *consciousness in a machine, or even clear evidence that instantiated machine consciousness will eventually be possible.* While some investigators have claimed 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. The author of this review believes that none of the past studies examined, even when claimed otherwise, has yet provided a convincing argument for how the approach being studied would eventually lead to instantiated artificial consciousness. On the other hand, and more positively, no evidence has yet been presented (including by the work surveyed in this review) that instantiated machine consciousness could not one day ultimately be possible, a view that has been expressed by others (Koch, 2001). In substantial ways, we are at the same point today with artificial consciousness as Alan Turing was in 1950 regarding AI: there are many objections to the possibility of instantiated machine consciousness, but none are as yet compelling.

4.2. *The future*

Given the barriers to investigating artificial consciousness, it is not surprising that past work has been limited in a number of ways. The five different approaches that have dominated work in artificial consciousness so far bring to mind nothing so much as the parable of the blind men and the elephant. While the work done to date has been substantial, it has not adequately separated out the information processing (functional) aspects of consciousness from the phenomenal aspects that are associated with the "hard problem" of consciousness. Further, while each of the five core properties underlying past computational models does seem to correlate with consciousness, each also seems to err on the side of being overly encompassing. For example, global workspace theory appears to encompass past "blackboard models" in AI that are not intended to model consciousness (Erman et al., 1980; van der Velde & de Kamps, 2006), and integrated information theory encompasses certain attractor neural networks specifically constructed to have arbitrarily large ϕ coefficients but that seem unlikely to exhibit consciousness (Seth et al., 2006). As a result, at present it is not clear which, if any, of the five correlates of consciousness that past models have been founded on are potentially fundamental causative factors, and which are instead secondary aspects that will be readily explained once we have a better scientific understanding of the nature of consciousness. Further, several of the implementations that have been reviewed here involve fairly small systems that have undergone quite limited, if any, critical evaluation. This is especially problematic because some investigators view theories that apply to small networks (just as well as they do to large networks) as being inadequate models of consciousness (Herzog, Esfeld, & Gerstner, 2007).

Nonetheless, and in spite of these limitations and continuing philosophical discussion of the implausibility of creating an artificial mind (Swiatczak, 2011), the fundamental scientific importance of understanding consciousness provides a compelling reason for continued and even expanded research in this area. As long as the creation of machine consciousness remains a possibility, it seems very desirable to pursue a broad research program in this area. What might be promising issues for further research? One possibility is to examine the inter-relationships between the five approaches that have been studied so far. For example, how does global workspace theory relate to integrated information, and what is the relationship between the corollary discharge associated with control of attention and self-models? Work in this area would also benefit a great deal from closer ties with empirical investigations in cognitive psychology and neuroscience that could better relate the results of computational models to experimental data. This would be especially true if models can make novel testable predictions rather than just confirming compliance with experimental data that already exists. Some progress in this direction has already been made (for example, see Dehaene et al., 2003; Massimini et al., 2005). Finally, given that most past work on machine consciousness has focused on the five approaches considered in this review, another important research direction is to explore additional approaches that may prove useful.

What would be the ethical implications if we were to be successful in producing an artifact possessing instantiated conscious, having subjective experiences and a mind in the same sense that a person does? The increasing plausibility that machine consciousness may eventually be achieved has led to a number of discussions of the rights such machines might merit and the ethics associated with intelligent machines in general (Lin, Abney, & Bekey, 2011; Tonkens, 2009; White, 2012). However, historically the most prominent concern has been the existential threat that such an event might pose to humanity – the so-called "Terminator Scenario" – that continues to be a recurring theme in contemporary science fiction. Of course, science fiction is, first and foremost, fiction. Nonetheless, the science fiction literature considered as a whole also reflects our shared, collective imagination about the future of science and technology. When it comes to the future of machine consciousness, most people are primarily familiar with the darker aspects of the science fiction literature, such as the *Terminator* movie series in which Skynet, upon becoming conscious, destroys civilization, or the recent novel *Robopocalypse* with a similar theme, or Philip Dick's 1968 story *Do Androids Dream of Electric Sheep* that inspired the movie *Blade Runner*. It is therefore surprising to many people that there is probably an equal amount of less sensational science fiction literature about the benefits of conscious machines to humanity, such as Robert Heinlein's *The Moon is a Harsh Mistress*, Isaac Asimov's *Bicentennial Man*, and Robert Sawyer's recent *www Trilogy*. It seems that our collective imagination as represented in this literature is actually fairly evenly split concerning whether phenomenally conscious machines would destroy us or would be an enormous benefit to humanity.

In recent years, speculation on the implications of conscious machines has been receiving increasing attention from scientific investigators and others who are decidedly not concerned with fiction. Here again, one finds remarkably different views. On the one hand, a number of organizations have emerged that are attempting to assess the long term dangers/benefits of AI to humanity, including machines with super-human intelligence and artificial consciousness (Singularity Institute, Lifeboat Foundation, etc.), and to ameliorate any perceived risks involved. Given predictions of a forthcoming Technological Singularity that may be associated with super-intelligent and conscious machines that pose an existential threat to humanity (Chalmers, 2010; Kurzweil, 2005; Vinge, 1993), this seems like a prudent step.

On the other hand, a technological breakthrough producing machine consciousness may simply be a natural aspect of evolution that could lead to substantial benefits to humanity. For example, it has been speculated that developing an understanding of machine consciousness could enable us to copy a conscious human mind into a machine, thereby enormously expanding the duration of a person's mental life (Goertzel & Ikle, 2012). At the least it could lead to more effective and natural human–machine interactions, and to a deeper understanding of natural consciousness, perhaps even shedding light on various neurological and psychiatric disorders.

The current author's opinion is that instantiated/phenomenal machine consciousness will almost certainly be achieved, perhaps as soon as in the next two or three decades.[4] A critical prerequisite for this will be a deeper understanding of consciousness as it occurs in people and (presumably) animals. Until we understand the fundamental biological and physical basis for how the human brain relates to our own conscious awareness, it will remain very difficult to create artifacts that truly model or support analogous artificial conscious states. Research in cognitive psychology and neuroscience that is attempting to provide a better understanding of natural consciousness has been accelerating in recent years. Once an improved understanding occurs, advances in machine consciousness will probably be very rapid, and as Samuel Butler suggested over 140 years ago, possibly inevitable.

References

Adami, C. (2006). What do robots dream of? *Science*, *314*, 1093–1094.

Agrawala, A. (2012). A framework for consciousness and its interactions with the physical world. In *Toward a science of consciousness*. Arizona: Tucson.

Aleksander, I. (1996). *Impossible minds*. Imperial College Press.

Aleksander, I. (2007). Machine consciousness. In M. Velmans, & S. Schneider (Eds.), *The Blackwell companion to consciousness* (pp. 87–98). Blackwell.

Aleksander, I., & Dunmall, B. (2003). Axioms and tests for the presence of minimal consciousness in agents. *Journal of Consciousness Studies*, *10*, 7–18.

Aleksander, I., & Gamez, D. (2009). Iconic training and effective information. In *Proceedings of the AAAI fall symposium BICA II* (pp. 2–10). AAAI.

Aleksander, I., & Gamez, D. (2011). Informational theories of consciousness. In C. Hernandez, et al. (Eds.), *From brains to systems* (pp. 139–147). Springer.

Aleksander, I., & Morton, H. (2007). Phenomenology and digital neural architectures. *Neural Networks*, *20*, 932–937.

Arrabales, R., Ledezma, A., & Sanchis, A. (2009). CERA–CRANIUM: a test bed for machine consciousness research. In *Proceedings of the international workshop on machine consciousness* (pp. 105–124). Hong Kong.

Arrabales, R., Ledezma, A., & Sanchis, A. (2010). The cognitive development of machine consciousness implementations. *International Journal of Machine Consciousness*, *2*, 213–235.

Ascoli, G. (2005). Brain and mind at the crossroads of time. *Cortex*, *41*, 619–620.

Atkinson, A., Thomas, M., & Cleeremans, A. (2000). Consciousness: mapping the theoretical landscape. *Trends in Cognitive Sciences*, *4*, 372–382.

Baars, B. (1988). *A cognitive theory of consciousness*. Cambridge University Press.

Baars, B. (2002). The conscious access hypothesis. *Trends in Cognitive Sciences*, *6*, 47–52.

Baars, B., & Franklin, S. (2007). An architectural model of conscious and unconscious brain function. *Neural Networks*, *20*, 955–961.

Baars, B., & Franklin, S. (2009). Consciousness is computational: the LIDA model of global workspace theory. *International Journal of Machine Consciousness*, *1*, 23–32.

Baars, B., Ramsey, T., & Laureys, S. (2003). Brain, conscious experience, and the observing self. *Trends in Neuroscience*, *26*, 671–675.

Balduzzi, D., & Tononi, G. (2008). Integrated information in discrete dynamical systems. *PLoS Computational Biology*, *4*, 1–18.

Bishop, M. (2009). Why computers can't feel pain. *Minds and Machines*, *19*, 507–516.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*, 227–247.

Block, N. (2005). Two neural correlates of consciousness. *Trends in Cognitive Sciences*, *9*, 46–52.

Bongard, J., Zykov, V., & Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science*, *314*, 1118–1121.

Bosse, T., Jonker, C., & Treur, J. (2008). Formalization of Damasio's theory of emotion, feeling and core consciousness. *Consciousness and Cognition*, *17*, 94–113.

Bryson, J. (2012). A role for consciousness in action selection. *International Journal of Machine Consciousness*, *4*, 471–482.

Buschman, T., & Miller, E. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science*, *315*, 1860–1862.

Buttazzo, G. (2008). Artificial consciousness: hazardous questions (and answers). *Artificial Intelligence in Medicine*, *44*, 139–146.

Carruthers, P. (2005). *Consciousness—essays from a higher-order perspective*. Oxford University Press.

Cattell, R., & Parker, A. (2012). Challenges for brain emulation. In *Natural intelligence, vol. 1* (pp. 17–31).

Cerullo, M. (2011). Integrated information theory. *Journal of Consciousness Studies*, *18*, 45–58.

Chalmers, D. (1996). *The conscious mind*. Oxford University Press.

Chalmers, D. (2007). The hard problem of consciousness. In M. Velmans, & S. Schneider (Eds.), *The blackwell companion to consciousness* (pp. 225–235). Blackwell.

Chalmers, D. (2010). The singularity. *Journal of Consciousness Studies*, *17*, 7–65.

Charkaoui, N. (2005). A computational model of minimal consciousness functions. *Proceedings of the World Academy of Science, Engineering and Technology*, *9*, 78–85.

Chella, A. (2007). Towards robot conscious perception. In A. Chella, & R. Manzotti (Eds.), *Artificial consciousness* (pp. 124–140). Imprint Academic.

Chella, A., Frixione, M., & Gaglio, S. (2008). A cognitive architecture for robot self-consciousness. *Artificial Intelligence in Medicine*, *44*, 147–154.

Chella, A., & Gaglio, S. (2012). Synthetic phenomenology and high-dimensional buffer hypothesis. *International Journal of Machine Consciousness*, *4*, 353–365.

Churchland, P. (1984). *Matter and consciousness*. MIT Press.

Cleeremans, A., Timmermans, B., & Pasquali, A. (2007). Consciousness and metarepresentation: a computational sketch. *Neural Networks*, *20*, 1032–1039.

Clowes, R., & Seth, A. (2008). Axioms, properties and criteria: roles for synthesis in the science of consciousness. *Artificial Intelligence in Medicine*, *44*, 91–104.

Conduit, R. (2007). To sleep, perchance to dream. *Science*, *315*, 1219–1220.

Connor, D., & Shanahan, M. (2010). A computational model of a global neuronal workspace with stochastic connections. *Neural Networks*, *23*, 1139–1154.

Conrad, L., & Sun, R. (2007). Hierarchical approaches to understanding consciousness. *Neural Networks*, *20*, 947–954.

Cook, N. (1999). Simulating consciousness in a bilateral neural network. *Consciousness and Cognition*, *8*, 62–93.

Corbetta, M., & Shulman, G. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*, 201–215.

Cotterill, R. (1995). On the unity of conscious experience. *Journal of Consciousness Studies*, *2*, 290–312.

Cotterill, R. (1996). Prediction and internal feedback in conscious perception. *Journal of Consciousness Studies*, *3*, 245–266.

Cotterill, R. (1997). On the mechanism of consciousness. *Journal of Consciousness Studies*, *4*, 231–247.

Cotterill, R. (1998). *Enchanted looms*. Cambridge University Press.

Cotterill, R. (2003). CyberChild—a simulation test-bed for consciousness studies. *Journal of Consciousness Studies*, *10*, 31–45.

Cowan, N., Elliott, E., Saults, J., Morey, C., et al. (2005). On the capacity of attention: its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, *51*, 42–100.

Coward, L., & Gedeon, R. (2009). Implications of resource limitations for a conscious machine. *Neurocomputing*, *72*, 767–788.

Cox, M. (2007). Perpetual self-aware cognitive agents. In *AI magazine* (pp. 32–45). [spring issue].

Cox, M. (2011). Metareasoning, monitoring, and self-explanation. In M. Cox, & A. Raja (Eds.), *Metareasoning* (pp. 131–149). MIT Press.

Crick, F. (1984). Function of the thalamic reticular complex: the searchlight hypothesis. *Proceedings of the National Academy of Sciences*, *81*, 4586–4590.

Crick, F. (1994). *The astonishing hypothesis*. Charles Scribner's Sons.

Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences*, 263–275.

Crick, F., & Koch, C. (2003). A framework for consciousness. *Nature Neuroscience*, *6*, 2003.

Culbertson, J. (1982). Consciousness. In *Natural and artificial*. Libra.

Dawkins, R. (1976). *The selfish gene*. Oxford University Press.

Dehaene, S., Kerszberg, M., & Changeux, J. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings National Academy of Sciences*, *95*, 14529–14534.

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness. *Cognition*, *79*, 1–37.

Dehaene, S., Naccache, L., Cohen, L., et al. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. *Nature Neuroscience*, *4*, 752–758.

Dehaene, S., Sergent, C., & Changeux, J. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings National Academy of Sciences*, *100*, 8520–8525.

Eccles, J. (1994). *How the self controls its brain*. Springer Verlag.

Edelman, G. (1989). *The remembered present*. Basic Books.

Erman, L., Hayes-Roth, F., Lessor, V., & Reddy, D. (1980). The Hearsay-II speech understanding system. *Computing Surveys*, *12*, 213–253.

Fekete, T., & Edelman, S. (2011). Towards a computational theory of experience. *Consciousness and Cognition*, *20*, 807–827.

Franklin, S. (2003). IDA—a conscious artifact? *Journal of Consciousness Studies*, *10*, 47–66.

Franklin, S., & Graesser, A. (1999). A software agent model of consciousness. *Consciousness and Cognition*, *8*, 285–305.

Franklin, S., Ramamurthy, U., & D'mello, S. et al. (2007). LIDA: a computational model of global workspace theory and developmental learning. In *Proceedings of the AAAI fall symposium on artificial intelligence and consciousness*.

Franklin, S., Strain, S., Snaider, J., McCall, R., & Faghihi, U. (2012). Global workspace theory, its LIDA model, and the underlying neuroscience. *Biologically Inspired Cognitive Architectures*, *1*, 32–43.

Galin, D. (1996). The structure of subjective experience. In S. Hameroff, A. Kaszniak, & A. Scott (Eds.), *Toward a science of consciousness* (pp. 121–140). MIT Press.

Gallup, G. (1970). Chimpanzees: self-recognition. *Science*, *167*, 86–87.

Gamez, D. (2008). Progress in machine consciousness. *Consciousness and Cognition*, *17*, 887–910.

Gamez, D. (2010). Information integration based predictions about the conscious states of a spiking neural network. *Consciousness and Cognition*, *19*, 294–310.

Gamez, D. (2012). Empirically grounded claims about consciousness in computers. *International Journal of Machine Consciousness*, *4*, 421–438.

[4] Based on projections of increasing computer memory size, it has been argued that machines will have sufficient memory to become self-aware by 2029 (Buttazzo, 2008). Coincidently, this is the year that the fictional Skynet is to be activated.

Garis, H., Shuo, C., Goertzel, B., & Ruiting, L. (2010). A world survey of artificial brain projects. *Neurocomputing*, *74*, 3–29.

Goertzel, B. (2011). Hyperset models of self, will and reflective consciousness. *International Journal of Machine Consciousness*, *3*, 19–53.

Goertzel, B., & Ikle, M. (2012). Mind uploading (introduction to a special issue on this topic). *International Journal of Machine Consciousness*, *4*, 1–3.

Gold, K., & Scassellati, B. (2007). A bayesian robot that distinguishes 'self' from 'other'. In *Proceedings of the twenty-ninth annual meeting of the cognitive science society* (pp. 384–392). Erlbaum.

Goldenberg, G. (2003). Disorders of body perception and representation. In T. Feinberg, & M. Farah (Eds.), *Behavioral neurology and neuropsychology* (pp. 285–294).

Gordon, A., Hobbs, J., & Cox, M. (2011). Anthropomorphic self-models for metareasoning agents. In M. Cox, & A. Raja (Eds.), *Metareasoning* (pp. 296–305). MIT Press.

Goswami, A., Reed, R., & Goswami, M. (1993). *The self-aware universe*. Putnam.

Gray, J., Buhusi, C., & Schmajuk, N. (1997). The transition from automatic to controlled processing. *Neural Networks*, *10*, 1257–1268.

Haier, R., et al. (1992). Regional glucose metabolic changes after learning a complex visuospatial/motor task. *Brain Research*, *570*, 134–143.

Haikonen, P. (2003). *The cognitive approach to conscious machines*. Imprint Academic.

Haikonen, P. (2007a). *Robot brains: circuits and systems for conscious machines*. John Wiley & Sons.

Haikonen, P. (2007b). Essential issues of conscious machines. *Journal of Consciousness Studies*, *14*, 72–84.

Haikonen, P. (2012). *Consciousness and robot sentience*. World Scientific.

Hameroff, S., & Penrose, R. (1996). Orchestrated reduction of quantum coherence in brain microtubules. *Journal of Consciousness Studies*, *3*, 36–53.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, *42*, 335–346.

Hart, J., & Scassellati, B. (2011). Robotic models of self. In M. Cox, & A. Raja (Eds.), *Metareasoning* (pp. 283–293). MIT Press.

Harth, E. (1995). The sketchpad model. *Consciousness and Cognition*, *4*, 346–368.

Hellie, B. (2009). In T. Bayne, A. Cleeremans, & P. Wilken (Eds.), *The Oxford companion to consciousness*, *Representationalism* (pp. 563–567). Oxford University Press.

Herzog, M., Esfeld, M., & Gerstner, W. (2007). Consciousness and the small network argument. *Neural Networks*, *20*, 1054–1056.

Hesslow, G. (2002). Conscious thought as simulation of behavior and perception. *Trends in Cognitive Sciences*, *6*, 242–247.

Hesslow, G., & Jierenhed, D. (2007). The inner world of a simple robot. *Journal of Consciousness Studies*, *14*, 85–96.

Hillis, W. (1998). *The pattern on the stone*. Basic Books.

Holland, O. (2007). A strongly embodied approach to machine consciousness. *Journal of Consciousness Studies*, *14*, 97–110.

Holland, O. (2009). In T. Bayne, A. Cleeremans, & P. Wilken (Eds.), *The Oxford companion to consciousness*, *Machine consciousness* (pp. 415–417). Oxford University Press.

Holland, O., & Goodman, R. (2003). Robots with internal models. *Journal of Consciousness Studies*, *10*, 77–109.

Hutto, D. (2009). In T. Bayne, A. Cleeremans, & P. Wilken (Eds.), *The Oxford companion to consciousness*, *Idealism* (pp. 357–359). Oxford University Press.

Huynh, T., & Reggia, J. (2011). Guiding hidden layer representations for improved rule extraction from neural networks. *IEEE Transactions on Neural Networks*, *22*, 264–275.

Huynh, T., & Reggia, J. (2012). Symbolic representation of recurrent neural network dynamics. *IEEE Transactions on Neural Networks and Learning Systems*, *23*, 1649–1658.

Jaynes, J. (1976). *The origin of consciousness in the breakdown of the bicameral mind*. Houghton Mifflin.

John, E. (2002). The neurophysics of consciousness. *Brain Research Reviews*, *39*, 1–28.

Kanwisher, N., & Wojciulik, E. (2000). Visual attention: insights from brain imaging. *Nature Reviews Neuroscience*, *1*, 91–100.

Kastner, S. (2009). Attention—neural basis. In T. Bayne, A. Cleeremans, & P. Wilken (Eds.), *The Oxford companion to consciousness* (pp. 72–78). Oxford University Press.

Katayama, K., & Takeno, J. (2011). Conscious expectation system. In *Proceedings biologically inspired cognitive architectures* (pp. 180–187). IOS Press.

Keenan, J., Rubio, J., Racioppi, C., Johnson, A., & Barnacz, B. (2005). The right hemisphere and the dark side of consciousness. *Cortex*, *41*, 695–704.

Kitamura, T., Tahara, T., & Asami, K. (2000). How can a robot have consciousness? *Advanced Robotics*, *14*, 263–275.

Koch, C. (2001). *Final report of the workshop "can a machine be conscieous?"*. The Swartz Foundation, http://www.theswartzfoundation.org/abstracts/2001_summary.asp.

Koch, C. (2004). *The quest for consciousness*. Roberts & Co.

Koch, C., & Tononi, G. (2008). Can machines be conscious? *IEEE Spectrum*, (June), 55–59.

Koch, C., & Tsuchiya, N. (2006). Attention and consciousness: two distinct brain processes. *Trends in Cognitive Sciences*, *11*, 16–22.

Kriegel, U. (2007). Philosophical theories of consciousness. In P. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The cambridge handbook of consciousness* (pp. 35–66). Cambridge University Press.

Kuipers, B. (2005). Consciousness: drinking from the firehose of experience. In *Proceedings 20th national conference on artificial intelligence* (pp. 1298–1305). AI Press.

Kuipers, B. (2008). Drinking from the firehose of experience. *Artificial Intelligence in Medicine*, *44*, 155–170.

Kurzweil, R. (2005). *The singularity is near*. New York: Viking.

Lamme, V. (2003). Why visual attention and awareness are different. *Trends in Cognitive Sciences*, *7*, 12–18.

Levine, J. (1983). Materialism and qualia: the explanatory gap. *Pacific Philosophical Quarterly*, *64*, 354–361.

Lin, P., Abney, K., & Bekey, G. (2011). *Robot ethics*. MIT Press.

Llinas Ribary, U., Contreras, D., & Pedroarena, C. (1998). *Philosophical Transactions of the Royal Society London B*, *353*, 1841–1849.

Lycan, W. (2009). Higher-order representation theories of consciousness. In T. Bayne, A. Cleeremans, & P. Wilken (Eds.), *The Oxford companion to consciousness* (pp. 346–350). Oxford University Press.

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.

McCauley, L. (2007). Demonstrating the benefit of computational consciousness. In *Proceedings of the AAAI fall 2007 symposia* (pp. 108–114).

McDermott, D. (2007). Artificial intelligence and consciousness. In P. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *Cambridge handbook of consciousness* (pp. 117–150). Cambridge University Press.

McFadden, J. (2000). *Quantum evolution*. Norton.

McGinn, C. (2004). *Consciousness and its origins*. Oxford University Press.

Metzinger, T. (2000a). The subjectivity of subjective experience. In T. Metzinger (Ed.), *Neural correlates of consciousness* (pp. 285–306). MIT Press.

Metzinger, T. (2000b). *Neural correlates of consciousness*. MIT Press.

Min, B. (2010). A thalamic reticular networking model of consciousness. *Theoretical Biology and Medical Modelling*, *7*(10), 1–18.

Minsky, M. (1968). Matter, mind and models. In M. Minsky (Ed.), *Semantic information processing* (pp. 425–532). MIT Press.

Molyneux, B. (2012). How the problem of consciousness could emerge in robots. *Minds and Machines*, *22*, 277–297.

Monner, D., & Reggia, J. (2011). Systematically grounding language through vision in a deep recurrent neural network. In J. Schmidhuber, K. Thorision, & M. Looks (Eds.), *Proceedings fourth international conference on artificial general intelligence* (pp. 112–121). Springer Verlag.

Monner, D., & Reggia, J. (2013). Emergent latent symbol systems in recurrent neural networks (submitted for publication).

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, *4*, 435–450.

Newman, J. (1997). Putting the puzzle together. *Journal of Consciousness Studies*, *4*, 47–66.

Newman, J., Baars, B., & Cho, S. (1997). A neural global workspace model for conscious attention. *Neural Networks*, *10*, 1195–1206.

Niedermeyer, E., & Silva, F. (2005). *Electroencephalography*. Lippincott Williams & Wilkins.

Norgaard, M., Ravn, O., Poulsen, N., & Hansen, L. (2000). *Neural networks for modelling and control of dynamic systems*. Springer.

O'Brien, G., & Opie, J. (1999). A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences*, *22*, 127–196.

Oh., H., Gentili, R., Reggia, J., & Contreras-Vidal, J. (2012). Modeling of visuospatial perspectives processing and modulation of fronto-parietal network activity during action imitation. In *Proceedings 34th annual international conference of the ieee engineering in medicine and biology society*.

O'Regan, J. (2012). How to build a robot that is conscious and feels. *Minds and Machines*, *22*, 117–136.

O'Reilly, R., & Frank, M. (2006). Making working memory work. *Neural Computation*, *18*, 283–328.

Pasquali, A., Timmermans, B., & Cleeremans, A. (2010). Know thyself: metacognitive networks and measures of consciousness. *Cognition*, *117*, 182–190.

Persaud, N., Mcleod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience*, *10*, 257–261.

Persaud, N., Mcleod, P., & Cowey, A. (2008). Experiments show what post-decision wagering measures. *Consciousness and Cognition*, *17*, 984–985.

Perlis, D. (1997). Consciousness as self-function. *Journal of Consciousness Studies*, *4*, 509–525.

Pockett, S. (2000). *The nature of consciousness: a hypothesis*. Writers Club Press.

Pockett, S. (2002). Difficulties with the electromagnetic field theory of consciousness. *Journal of Consciousness Studies*, *9*, 51–56.

Posner, J., Saper, C., Schiff, N., & Plum, F. (2007). *Plum and Posner's diagnosis of stupor and coma*. Oxford University Press.

Prinz, J. (2003). Level-headed mysterianism and artificial experience. *Journal of Consciousness Studies*, *10*, 111–132.

Raffone, A., & Pantani, M. (2010). A global workspace model for phenomenal and access consciousness. *Consciousness and Cognition*, *19*, 580–596.

Raizer, K., Paraense, A., & Gudwin, R. (2012). A cognitive architecture with incremental levels of machine consciousness inspired by cognitive neuroscience. *International Journal of Machine Consciousness*, *4*, 335–352.

Ramamurthy, U., Franklin, S., & Agrawal, P. (2012). Self-system in a model of cognition. *International Journal of Machine Consciousness*, *4*, 325–333.

Rees, G. (2009). Correlates of consciousness. In T. Bayne, A. Cleeremans, & P. Wilken (Eds.), *The Oxford companion to consciousness* (pp. 207–211). Oxford University Press.

Rees, G., Kreiman, G., & Koch, C. (2002). Neural correlates of consciousness in humans. *Nature Reviews Neuroscience*, *3*, 261–270.

Reggia, J., Goodall, S., Revett, K., & Ruppin, E. (2000). Computational modeling of the cortical response to focal damage. In H. Levin, & J. Grafman (Eds.), *Cerebral reorganization of function after brain damage* (pp. 331–353). Oxford University Press.

Reggia, J., & Levitan, S. (2003). Hemisphere interactions and specialization. In M. Arbib (Ed.), *The handbook of brain theory and neural networks* (pp. 525–528). MIT Press.

Revonsuo, A. (2006). *Inner presence: consciousness as a biological phenomenon*. MIT Press.

Rey, G. (1983). A reason for doubting the existence of consciousness. In R. Davidson, G. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation, vol. 3*. Springer.

Reynolds, J., & Chelazzi, L. (2004). Attentional modulation of visual processing. *Annual Review of Neuroscience*, *27*, 611–647.

Rolls, E. (1997). Consciousness in neural networks? *Neural Networks*, *10*, 1227–1240.

Rolls, E. (2007). A computational neuroscience approach to consciousness. *Neural Networks*, *20*, 962–968.

Rosenthal, D. (1996). A theory of consciousness. In N. Block, O. Flanagan, & G. Guzeldere (Eds.), *The nature of consciousness* (pp. 729–753). MIT Press.

Samsonovich, A. (2007). *Universal learner as an embryo of computational consciousness. In A. Chella, R. Manzotti, (Eds.), In AI and consciousness: theoretical foundations and current approaches, AAAI fall symposium, AAAI Technical Report FS-07-01*. (pp. 129–134).

Samsonovich, A., & Ascoli, G. (2005). The conscious self: ontology, epistemology and the mirror quest. *Cortex*, *41*, 621–636.

Samsonovich, A., & DeJong, K. (2005). A general-purpose computational model of the conscious mind. In M. Lovett, et al. (Eds.), *ICCM-2004, Proceedings of the sixth international conference on cognitive modeling* (pp. 382–383).

Samsonovich, A., & Nadel, L. (2005). Fundamental principles and mechanisms of the conscious self. *Cortex*, *41*, 669–689.

Sanz, R., Hernandez, C., & Sanchez-Escribano, M. (2012). Consciousness, action selection, meaning and phenomenic anticipation. *International Journal of Machine Consciousness*, *4*, 383–399.

Schlagel, R. (1999). Why not artificial consciousness or thought? *Minds and Machines*, *9*, 3–28.

Searle, J. (2004). *Mind*. Oxford University Press.

Seth, A. (2008). Post-decision wagering measures metacognitive content, not sensory consciousness. *Consciousness and Cognition*, *17*, 981–983.

Seth, A. (2009). The strength of weak artificial consciousness. *International Journal of Machine Consciousness*, *1*, 71–82.

Seth, A., Izhikevich, E., Reeke, G., & Edelman, G. (2006). Theories and measures of consciousness. *Proceedings of the National Academy of Sciences*, *103*, 10799–10804.

Shanahan, M. (2006). A cognitive architecture that combines internal simulation with a global workspace. *Consciousness and Cognition*, *15*, 433–449.

Shanahan, M. (2008). A spiking neuron model of cortical broadcast and competition. *Consciousness and Cognition*, *17*, 288–303.

Shanahan, M. (2010). *Embodiment and the inner life*. Oxford University Press.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423. 623–656.

Shipp, S. (2004). The brain circuitry of attention. *Trends in Cognitive Sciences*, *8*, 223–230.

Silver, M., & Kastner, S. (2009). Topographic maps in human frontal and parietal cortex. *Trends in Cognitive Sciences*, *13*, 488–495.

Sloman, A. (2010). Phenomenal and access consciousness and the 'hard' problem. *International Journal of Machine Consciousness*, *2*, 117–169.

Sloman, A., & Chrisley, R. (2003). Virtual machines and consciousness. *Journal of Consciousness Studies*, *10*, 133–172.

Sporns, O. (2011). *Networks of the brain*. MIT Press.

Stapp, H. (1993). *Mind, matter and quantum mechanics*. Springer Verlag.

Starzyk, J., & Prasad, D. (2011). A computational model of machine consciousness. *International Journal of Machine Consciousness*, *3*, 255–281.

Steriade, M. (1996). Arousal: revisiting the reticular activating system. *Science*, *272*, 225–226.

Sun, R. (1997). Learning, action and consciousness. *Neural Networks*, *10*, 1317–1331.

Sun, R. (1999). Accounting for the computational basis of consciousness. *Consciousness and Cognition*, *8*, 529–565.

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

Sun, R., & Franklin, S. (2007). Computational models of consciousness. In P. Zelazo, & M. Moscovitch (Eds.), *Cambridge handbook of consciousness* (pp. 151–174). Cambridge University Press.

Swiatczak, B. (2011). Conscious representations: an intractable problem for the computational theory of mind. *Minds and Machines*, *21*, 19–32.

Sylvester, J., Reggia, J., & Weems, S. (2011). Cognitive control as a gated cortical network. In A. Samsonovich, & K. Johannsdottir (Eds.), *Proceedings second international conference on biologically-inspired cognitive architectures* (pp. 371–376). IOP Press.

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.

Takeno, J. (2011). MoNAD structure and self-awareness. In *Proceedings biologically inspired cognitive architectures* (pp. 377–382). IOP Press.

Takeno, J. (2013). *Creation of a conscious robot*. Pan Stanford.

Taylor, J. (1997). Neural networks for consciousness. *Neural Networks*, *10*, 1207–1225.

Taylor, J. (2003a). Neural models of consciousness. In M. Arbib (Ed.), *Handbook of brain theory and neural networks* (pp. 263–267). MIT Press.

Taylor, J. (2003b). Paying attention to consciousness. *Progress in Neurobiology*, *71*, 305–335.

Taylor, J. (2007). CODAM: a neural network model of consciousness. *Neural Networks*, *20*, 983–992.

Taylor, J. (2012). Does the corollary discharge of attention exist? *Consciousness and Cognition*, *21*, 325–339.

Taylor, J., & Fragopanagos, N. (2007). Resolving some confusions over attention and consciousness. *Neural Networks*, *20*, 993–1003.

Tinsley, C. (2008). Using topographic networks to build a representation of consciousness. *BioSystems*, *92*, 29–41.

Tonkens, R. (2009). A challenge for machine ethics. *Minds and Machines*, *19*, 421–438.

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.

Tononi, G., & Edelman, G. (1998). Consciousness and complexity. *Science*, *282*, 1846–1851.

Tononi, G., & Sporns, O. (2003). Measuring information integration. *BMC Neuroscience*, *4*, 31.

Tononi, G., Sporns, O., & Edelman, G. (1994). A measure for brain complexity. *Proceedings of the National Academy of Sciences*, *91*, 5033–5037.

Treisman, A. (2009). Attention: theoretical and psychological perspectives. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 189–204). MIT Press.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *59*, 433–460.

van der Velde, F., & de Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. *Behavioral and Brain Sciences*, *29*, 37–108.

Vinge, V. (1993). The coming technological singularity: how to survive in the post-human era. In *Whole earth review*. Winter.

Wallace, R. (2005). *Consciousness: a mathematical treatment of the global neuronal workspace model*. Springer.

Wallace, R. (2006). Pitfalls in biological computing: canonical and idiosyncratic dysfunction of conscious machines. *Mind and Matter*, *4*, 91–113.

Ward, L. (2011). The thalamic dynamic core theory of conscious experience. *Consciousness and Cognition*, *20*, 464–486.

Weiskrantz, L. (2009). *Blindsight*. Oxford University Press.

White, J. (2012). *Natural intelligence (summer issue), Autonomy rebuilt: rethinking traditional ethics towards a comprehensive account of autonomous moral agency* (pp. 32–39). International Neural Network Society.

Wilks, Y. (1984). Machines and consciousness. In C. Hookway (Ed.), *Minds, machines and evolution* (pp. 105–128). Cambridge University Press.

Zeman, A. (2001). Consciousness. *Brain*, *124*, 1263–1289.