How Artificial Intelligence Can Inform Neuroscience: A Recipe for Conscious Machines?

Fahad Khalid*

Hasso Plattner Institute for Software Systems Engineering, Potsdam, Germany

Emília Garcia-Casademont†

Institut de Biologia Evolutiva (CSIC–UPF),

Universitat Pompeu Fabra, Barcelona, Spain

Sarah Laborde‡

The Ohio State University, Department of Anthropology, Columbus, OH 43210 USA

Claire Lagesse§

Université Paris Diderot, Sorbonne Paris Cité,

Matière et Systèmes Complexes (MSC), UMR 7057, Paris, France.

Elizabeth Lusczek¶

University of Minnesota, Department of Surgery, Minneapolis, MN 55455 USA

(Dated: September 2014)

Abstract

An approach based on Artificial Intelligence is proposed for the design of conscious behavior in artificial agents, with the objective that it can inform theories of consciousness in neuroscience. Points of comparison between neuroscientific theories of consciousness and the artificial intelligence based approach are presented and analyzed. The significance of an interdisciplinary approach to studying consciousness is emphasized by highlighting connections between anthropology, neuroscience, and artificial intelligence.

* fahad.khalid@hpi.de

† emilia-maria.garcia@upf.edu

‡ laborde.7@osu.edu

§ lagesse.claire@gmail.com

¶ lusc0006@umn.edu

"We have no experience (apart from the very limited view provided by our own introspection) of machines having complex, rapidly changing and highly parallel activity of this type. When we can both construct such machines and understand their detailed behavior, much of the mystery of consciousness may disappear."

– Francis Crick and Christof Koch. "Towards a neurobiological theory of consciousness." *Seminars in the Neurosciences.* Vol. 2. Saunders Scientific Publications, 1990.

I. INTRODUCTION

The nature of consciousness, the property of some living beings to have subjective experience, has been a concern for scholars and artists for millennia. Today, self-organizing creatures, including conscious ones such as humans, are increasingly conceptualized and modeled as complex systems and their properties studied within this framework. This premise is what led us, a group of interdisciplinary scientists with an interest in complex systems, to think together about consciousness.

In this paper, we explore consciousness from the points of view of Neuroscience and Artificial Intelligence (AI). Specifically, we present a road map for development of conscious artificial agents, and compare our approach to the theories of consciousness established in the field of neuroscience. However, prior to the discussion of theories in neuroscience and AI, in the following subsections, we present historical and cultural perspectives on consciousness.

A. Historical Background

There is evidence that the nature of consciousness has been of fundamental concern to humanity for thousands of years across societies. The oldest relevant texts are theological, where consciousness is described as a mystical property. In the millenary Vedic culture of ancient India (circa 1500 BC) for instance, it was believed that each individual human consciousness was merely a partial manifestation of a universal Self or pure consciousness that was the cause of all physical phenomena. The conception of human consciousness in this tradition was twofold, with thought and pure consciousness (Shiva) on the one side and power, energy, movement, action, and will (Shakti) on the other. A concept similar to the Vedic universal consciousness is present in the God of most of Indo-European and Semitic cultures, although the relationship of human individuals to the universal consciousness is different.

Common to all early conceptions of consciousness is the idea that there are various levels of consciousness, and a higher level of consciousness means a higher degree of relation with the conscious immortality or perfection. For ancient Egyptians, five distinct elements were part of the human identity and consciousness in addition to the body. Life was the result of the convergence of six elements, five of them being of different nature than the body.

Ancient Greeks believed in a dual divinity-matter associated with a dual soul-body. Various conceptions of these dualities existed across communities and periods. In general, the soul was a source of life present in all animals (Thymós), and the immortal entity of the soul eventually joined with a human body (Psyché). The Greeks were particularly concerned about human consciousness as a means to achieve the highest level of mysticism, which was according to them the ultimate goal of a soul.

Copernican ideas on astronomy (1540) opened the door to the possibility that 'ideas about things' and 'things as we perceive them' could be seen as different entities. This pushed the development of mathematics as a tool for understanding the world instead of the sole reliance on the senses. This split between the 'perception of things' and 'things themselves' led to a conceptual split between the mind and the physical world. Expanding on this distinction between mind and body, Descartes' Meditations (1680) attempted to find a way to understand reality purely through introspection, by making use of only the reality of thought. This split between the mind and the physical world permitted considering the process of thinking and consciousness in isolation from any specific sensory input or physical subject matter. Following Descartes, mental processes had an existence of their own.

Understanding the connection between the mind and the physical world (the mind-body problem) subsequently occupied much of classical western philosophy. For Descartes, the intervention of God was needed. For modern empiricists (Hobbes, Locke, Hume), knowledge must be explained through an introspective but empirical psychology, and mental phenomena were divided into perceptions on one side and thought, memory, and imagination on the other. Kant presented a modern synthesis of rationalism and empiricism. He understood experiences as meaningful only through the contribution of the subject: without an active subject, the world would be nothing more than transitory sensations, and the introspection could not arise either.

Contemporary philosophy of mind rests on the historical basis of classical western philosophy, as well as being informed by the increasing knowledge about biological and physical systems. It is a rich and active field animated by lively debates that are beyond the scope of this paper. However, it is important to say that here we assume a physicalist stance with regards to consciousness, which means that we understand consciousness as a product of physical processes. This is a view that is more and more broadly accepted by philosophers and scientists who study the mind.

Along with philosophers, anthropologists have been concerned with human consciousness since it is at the root of human experience. Anthropology encompasses cultural and biological perspectives via a number of subfields to study human behavior and culture across social groups. What characterizes "consciousness" is difficult to define in the cross-cultural context of anthropological research. As a result, anthropologists have traditionally addressed the nature of conscious experience in indirect ways, for example through studies of perceptions of time and space, spiritual practices, or altered states of consciousness [Throop and Laughlin 2007].

Recent approaches have focused explicitly on the nature of conscious experience and sought to integrate neuroscientific and cognitive insights about the brain and nervous system with ethnographic accounts of phenomenal experience. The motivation for such approaches finds grounds in the words of Varela, one of the founders of neurophenomenology: "on the one hand we need to address our condition as bodily processes; on the other hand we are also an existence which is already there, a Dasein, constituted as an identity, and which cannot leap out and take a disembodied look at how it got to be there" [Rudrauf et al. 2003].

B. Anthropological Perpsective

Here we are interested in the ways a physicalist understanding of consciousness may be informed by the tools of artificial intelligence. Therefore, this paper focuses on consciousness at the intersection of neuroscience and artificial intelligence, and a thorough review of recent work in cognitive anthropology is beyond its scope. However, some insights from this body of work on consciousness are relevant to our approach. We summarize them below:

- The environmental, social and cultural contexts of human conscious experience shape the experience itself and its expression [Throop and Laughlin 2007]. It is of course also true that cultural practices are shaped by the intrinsic properties of human consciousness. In other words, the relationship between an individual's biophysical system and the subjective basis of its individuality involves environmental factors (including -at least for humans-, social and cultural dimensions of the environment). It follows that consciousness studies should consider brain, body, and environment as a nexus, or to paraphrase Ingold "the brain as embodied, and the body as 'enworlded'" [ing 2010].

- The 'hard problem' of consciousness pertains to the nature of subjective experience, as opposed to the expression of some of its processes such as information integration, emotion expression, etc [Chalmers 1995]. Starting from the premise that mysterianism (the view that the hard problem of consciousness is unsolvable) is not satisfactory, careful study of lived conscious experience with first-person methodologies appears to be required to complement biophysical consciousness studies [Varela and Shear 1996].

We do not attempt to solve the hard problem of consciousness. However, artificial intelligence analogies and experiments, as they are conducted by reflexive humans programming and interacting with artificial agents, can help reveal interesting connections between states of consciousness and their dynamical expression through emotion, language, and social interaction. Such integration of first-person accounts of computer scientists with data about artificial agents' dynamical expressions justifies our interdisciplinary approach.

C. Structure of the Paper

The rest of the paper is structured as follows: We highlight our contribution in Section I D. Section II discusses neuroscientific theories of consciousness; Integreated Information Theory is presented in Section II A, and the theory of consciousness as a product of social perception is presented in Section II B. We shift the focus to artificial consciousness in Section III. A brief introduction to the field of artificial intelligence is presented in Section III A. We develop our approach to consciousness in aritificial agents in Sections III B through III F. This is followed by a discussion on the points of comparison between the neuroscientific theories and our approach, along with open questions in Section IV. We conclude the paper with comments on possible future work in Section V.

D. Our Contribution

A summary of our contributions is as follows:

- The subject of consciousness has been explored in artificial intelligence before [Hofstadter et al. 1981, McCarthy 1995b, Minsky 2007, Moravec 1988, 2000, Perlis 1997, Sloman and Chrisley 2003]. However, to the best of our knowledge, the research strategy of integrating neuroscience and artificial intelligence is novel.

- We propose computationally feasible experiments to verify neuroscientific theories of consciousness based on information integration.

II. THEORIES OF CONSCIOUSNESS IN NEUROSCIENCE

Theoretical and technological advances in neurology, information theory, complexity science, medical technology and the like have provided important new tools that have advanced our understanding of the human condition considerably in the past century. However, it can be argued that none of these advances have been able to tell us what consciousness is, where it comes from, or why it is inherently subjective. The answers to such queries may be explored with scientific theories of consciousness. We consider Giulio Tononi's *Integrated Information Theory* [Tononi 2008] as well as Michael Graziano and Sabine Kastner's theory of *consciousness as a construct of social perception* [Graziano and Kastner 2011]. These two theories, which we found particularly convincing and pertinent to an AI approach, are based on the idea that consciousness is essentially an information processing mechanism. We argue that, taken together, these models provide a working guide for the properties and capabilities that conscious artificial agents should have. A brief introduction to relevant concepts in both theories is given here.

- Integrated Information Theory tells us how a complex system such as a brain would work (specifically, that the amount of information generated by a conscious system should exceed that produced by the sum of its parts) and provides a rigorous theoretical and mathematical framework for consciousness.

- Graziano and Kastner's theory gives insight as to the social perceptual mechanisms that generate awareness. As such, it provides a framework for the anthropological considerations of consciousness that Integrated Information Theory lacks.

A. Integrated Information Theory – Consciousness as Integrated Information

Integrated Information Theory (IIT) provides a framework for quantifying consciousness in a system according to the system's capacity to integrate information. Rooted in information theory, IIT proposes that a system which generates consciousness must be able to

integrate information from an extremely large repertoire of possible states. Information, quantified in a manner based on Shannon's theory [Shannon 1948], is said to be integrated if the amount of information generated by a system exceeds the amount of information generated by its components. Integration occurs in a way such that the conscious experience is unified. The system which produces conscious experience is itself unified in the sense that it cannot be broken down into a set of independent parts; doing so would result in diminishment or even loss of consciousness.

Consider the following simple example of a repertoire of states. Let us assume that a system A is currently in state s_1, and we would like to know the previous unknown state s_0 of A. For a two-bit system, A has a possible repertoire of four states: $\{00, 01, 10, 11\}$. If the system is currently in state '10', what was the previous state of the system? The system could have been in any of the four states with equal probability $\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$. This set of four probabilities is called the *potential (a priori) repertoire*. As we can see, this repertoire represents complete uncertainty since each state is equally probable. Next, assume that A is subject to a mechanism wherein if the current state is '10', there are only two possible previous states, '11' and '01'. This mechanism and the given state generate the *actual (a posteriori) repertoire* $\left(0, \frac{1}{2}, 0, \frac{1}{2}\right)$. This reduces the uncertainty and increases the amount of information since the number of possible previous states has been reduced from four to two. Stated another way, *information has been generated*.

The amount of information generated by a given system is quantified by *relative entropy*[1], allowing us to compare the amount of information generated by the system as a whole to the amount of information generated by its parts. *Effective information* (ei) quantifies the ability of a system to reduce uncertainty and *integrated information* (ϕ) quantifies the amount of information generated by a system that is above and beyond that of its independent components. Effective information and integrated information are defined as follows [Tononi 2008]:

$$ei(X(mech, x_1) = H[p(X_0(mech, x_1))||p(X_0(maxH))] \tag{1}$$

$$\phi(X(mech, x_1) = H[p(X_0(mech, x_1))||\prod p({}^kM_0(mech, \mu_1))] \tag{2}$$

[1]Loosely, relative entropy is the difference between two probability distributions, or the amount of information lost when probability distribution p is used to approximate probability distribution q.

H is the relative entropy between two repertoires: a particular mechanism in that repertoire is represented by $X(mech)$, and all possible repertoires as represented by $p(X_0(maxH))$. $p(X_0(mech, x_1))$ is the *a posteriori* distribution of system states at time $t = 0$ that could have caused state x_1 at time $t = 1$. Effective information is low for systems wherein many possible repertoires can lead to a particular state or if the repertoire of initial states is small. Effective information is large for systems in which a large repertoire of initial states leads to a small repertoire. In the latter case, uncertainty is reduced far more than in the former case.

For ϕ, we introduce the concept of the *minimum information partition (MIP)*. The MIP refers to the decomposition of the system into its minimal parts; that is, the configuration that leaves the least amount of information unaccounted for. ${}^kM_0$ specifies the parts of the repertoire that are part of the MIP, and ${}^kM_0(mech, \mu_1)$ is the probability distribution generated by the system if all parts are independent. Therefore, ϕ refers to the ability of the integrated system to generate more information than the system composed of independent parts. A conscious systems will, by definition, have $ei > 0$ and $\phi > 0$. These inequalities correspond to the existence of an informational relationship between the potential repertoire and the actual repertoire in such a way as to reduce uncertainty, and to the ability of the system as a whole to generate more information than the sum of its independent parts, respectively.

Quantification of integrated information helps define the basic structural unit of consciousness within the IIT framework. Such units are known as *complexes*, or any set of elements with $\phi > 0$ that is not fully contained within a different set with higher ϕ. Integrated information is generated inside complexes, and therefore each complex can be considered to have a "point of view". The human brain is likely composed of many highly interconnected complexes of varying values of ϕ. As we will discuss later, highly interconnected systems with corresponding high values of ϕ may arise in artificial agents as a matter of course.

IIT not only outlines the necessary and sufficient conditions for the generation of consciousness, i.e., the ability to generate and integrate information, it also addresses the subjective nature of conscious experience. The theory posits that subjective experience is specified by how information is integrated in a complex. A given complex may have any number of possible states. These states are mapped to axes in a space known as *Qualia space*. Informational relationships generated by a complex's mechanism define shapes in Qualia space

(Figure 1).

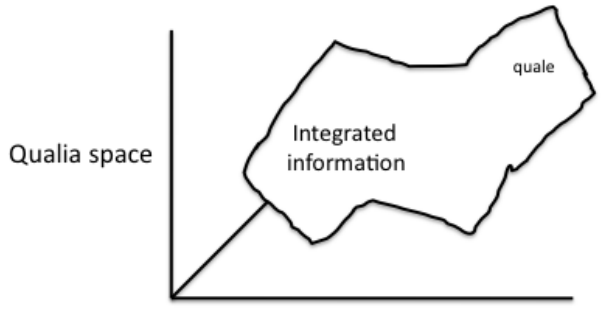

FIG. 1. In the Integrated Information Theory of consciousness, information is bound in a manner specific to the mechanisms of a given complex. The set of all possible mechanisms mapped into Qualia space form a shape called a *quale*. The quale, and no more than this, define the conscious experience

Such a shape is called *quale*, and it is the quale that completely defines a particular subjective experience. However, it is nearly impossible to determine the quale for all but the simplest model systems. Likewise, calculating ei and ϕ are computationally expensive and only feasible for model systems. Even so, computer simulations have demonstrated that the theory can account for basic neurological observations. For instance, networks with high values of ϕ have structural similarities with mammalian corticothalamic architecture. Additionally, simulated cortical cuts produce two complexes with high values of ϕ similar to those of the uncut systems. This is expected based on observations of so-called "split brain" studies [Gazzaniga 2005].

According to IIT, features of complex brains such as self-reflection, memory, localization of the body, mind-body connections, and attention are computed as a matter of course and are not requirements for the generation of consciousness *per se*. However, we claim that an artificial agent composed of just this would not convince an observer of its consciousness. The next theory of consciousness that we discuss proposes that humans are able to recognize consciousness in others because of specialized machinery related to social perception.

B. Consciousness as a Product of Social Perception

In the theory proposed by Michael Graziano and Sabine Kastner [Graziano and Kastner 2011], phenomenal consciousness is *awareness*, which in turn is a product of *attention*. Some

clarification is warranted since attention and awareness often seem indistinguishable. Attention is easily and intuitively understood as a function constantly used in human cognition–something that the brain does. Its neurophysiological basis can be found in the Superior Temporal Sulcus (STS) and the Temporo-Parietal Junction (TPJ). For instance, it has been found that the STS and TPJ are both activated in response to the appearance of unexpected stimuli [Corbetta et al. 2000, Shulman et al. 2010]. Of particular relevance to Graziano and Kastner's theory, the STS is recruited in the imputation of the direction of another's gaze, as well as the perception of facial movements, reaching, and predicting the intention behind another person's movements [Blakemore et al. 2003, Pelphrey et al. 2004]. These examples highlight the intertwined nature of attention to stimuli (movement, facial recognition) and modeling the behavior of others (intention of movement, the attentional states of others).

An important component of conscious beings, therefore, is to perceive and react appropriately to stimuli. This is accomplished by creating perceptual models of the attentional states of others–that is, by having *awareness*. The coupling of attention to awareness of self and other via social perceptual machinery allows this to occur. For instance, we often anticipate the wants and needs or likes and dislikes of those closest to us. Doing so requires a model of the other's attention; this is a fundamental task of social perception. The STS and TPJ are involved in these social functions and so-called "theory-of-mind" tasks, which use models to reconstruct others' states of mind [Frith and Frith 2003].

These models in themselves are a form of perception, and as such, are descriptive, computed involuntarily, and continually updated. Additionally, these models are localizable–we perceive awareness as emanating from others; more explicitly, other people are recognized as sources of consciousness. Finally, perception–particularly the perception of awareness–is subjective in nature. As Graziano and Kastner point out, these models are more useful than accurate. In a sense, they act as continuously updated instruction guides for how to interact with the environment. This aspect of Graziano and Kastner's theory underscores the importance of interaction with the environment to the phenomenon of consciousness, a component which neuroanthropologists emphasize but IIT lacks.

Everything that has been said about perceiving and reconstructing the attentional states of others can be said for perceiving and reconstructing one's own attentional states. If attention is something the brain *does*, awareness is something it *knows*. Perhaps the most salient example of the intimate relationship between attention, perception, and awareness is

the example of the feeling one has of one's own *self-awareness*. In this feeling, we perceive our own rich, complex inner worlds and can even localize them as emanating from within ourselves. We can see, then, that our capabilities for self-awareness are simply a result of turning the brain's substantial social machinery on ourselves.

Graziano and Kastner's theory of consciousness accepts that information binding is important to the generation of consciousness but builds on the concept by combining it with social theories of consciousness. In other words, consciousness does not only help the organism discriminate between large amounts of information in the environment (integrative function), it also helps it produce educated guesses of what other conscious organisms may do next (social function). This resonates with the theories proposed by evolutionary cognitive anthropologists such as Sperber [Sperber 1999].

Graziano and Kastner claim that in order to explain the feeling and recognition of consciousness, awareness is required in addition to bound information. Stated explicitly, *consciousness is awareness bound to information by attention.* At first glance, the theory of consciousness as social perception complements IIT. However, it conflicts with IIT in key ways. As previously alluded to, IIT claims the quality and amount of consciousness in a system is solely determined by its ability to integrate information. Features of complex brains such as self-reflection, memory, localization of the body, mind-body connections, and attention are computed as a matter of course in IIT. These are not absolute requirements for the generation of consciousness. To Graziano and Kastner, self-reflection, memory, localization of the body, and mind-body connections are exactly the features that comprise the conscious experience. Further, awareness is bound to information by attention, as shown in the bottom panel of Figure 2.

This conflicts with IIT in that consciousness requires something external to bound information. In IIT, consciousness *is* bound information.

Since this paper's focus is on *functions of consciousness* rather than specific mechanisms, Graziano and Kastner's theory of consciousness as awareness is particularly appropriate. Awareness is a model of attention that is a product of social perception; it can be used to reconstruct the attentional states of others or it can be turned inward to the self, providing models of self that can interact with models of others and the environment at large. This concept will prove useful in our consideration of artificial agents since interaction with the environment is an important aspect of our work.

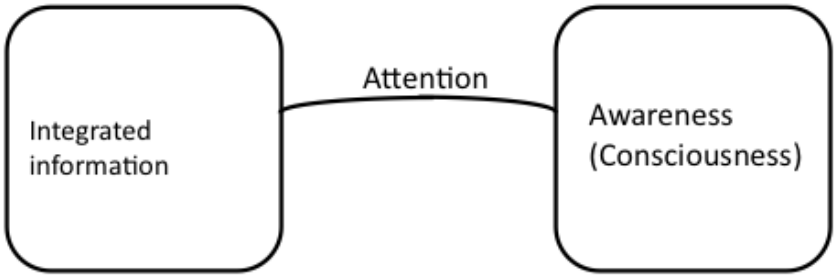

FIG. 2. In Graziano and Kastner's theory of consciousness, *attention* attaches *awareness* to bound information. Awareness, an entity external to bound information, is what defines the conscious experience. Contrast this with IIT in Figure 1, in which integrated information is the only component of consciousness.

III. ARTIFICIAL INTELLIGENCE AND CONSCIOUSNESS

Giulio Tononi has consistently posited that a combination of theory and experiment is required to adequately study the problem of consciousness [Tononi 2008]. In this work, we propose that Artificial Intelligence (AI) can bridge the theory-experiment paradigm by imitating the *functions of consciousness* in *artificial agents*. In doing so, we demonstrate how artificial intelligence can inform neuroscience.

Because IIT and Graziano's theory of consciousness are centered on information processing, we can 'test' some features of these theories with artificial agents. There are specific things missing in the neuroscience theories (such as emotional information and hierarchical architecture) that we specifically introduced into our model of consciousness in this paper. This is where the reflexive approach, informed by anthropology and science, is relevant. Keeping track of first-person accounts of programmers as they interact with artificial agents, in a systematic way (cf self-ethnographic methods), can help shed light on social elements of consciousness associated with emotion and language that are taken for granted when interacting with humans.

A. A Brief Introduction to AI

Artificial Intelligence is concerned with intelligent behavior. AI research is devoted to the development of explicit models of the structures and processes that give rise to intelligent behavior. The AI field of exploration extends from knowledge systems [Schreiber and

Akkermans 2000] to autonomous mobile robots [Siegwart et al. 2011], passing through any kind of artifact or system needed. There are several ways to approach the phenomenon of intelligence using the AI methodology.

Fundamentally, two paradigms have given the most successful results in AI so far, though the question of what intelligence is still remains unsolved. The *symbolic paradigm* is the oldest approach to intelligence in AI and has its roots in logic. It assumes that the internal structure of an intelligent agent is based on symbolic representations and internal processes manipulate these symbolic representations. Another paradigm in AI, which has its roots particularly in biology and evolutionary theory, is the *dynamical systems paradigm*. This approach shifts the locus of intelligence away from the internal symbolic models and processing of the agent towards the interaction process between the agent and the environment.

During the last two decades, most of the effort in AI has moved from the design of fundamental experimental setups and new methodologies to the deep exploration of existing techniques (such as Machine Learning or Stochastic Methods). There are still open research lines in many sides of the field, including the more fundamental questions.

B. Prolegomenon to the Theory of Artificial Consciousness

The theory of artificial consciousness presented in this Section is founded on the following premise: If we consider consciousness as a set of functions – such as awareness, emotions, etc. – then an artificial agent can be endowed with the same set of functions as a human being. Such an agent would appear conscious to us in the same way other humans appear conscious. The *functions of consciousness* [Minsky 2007] form the basis of comparison between the theories of consciousness developed in neuroscience, and the theory of artificial consciousness developed in this paper (see Figure 3 for an overview).

In addition to the discussion of functions as mechanisms, we describe the architectural details of an artificial implementation of these functions. In doing so, we elaborate on the hierarchical and modular nature of such an architecture. Moreover, we comment on the interconnectedness of the different components/modules involved in the various computations.

Based on insights from anthropology and neuroscience presented earlier in the paper, it is also our hypothesis that for a system to be able to generate conscious behavior, it must be able to interact with its environment. We consider conscious agents, both natural and

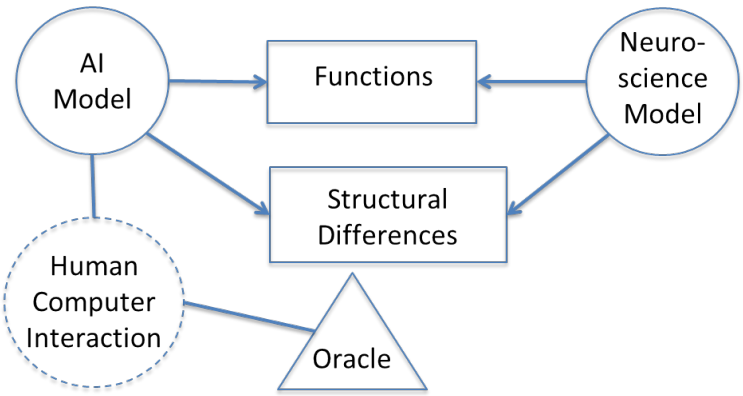

FIG. 3. The figure highlights *functions* and *structural elements* as points of connection between theories of consciousness in AI and neuroscience. Moreover, *human computer interaction* is presented as an *Oracle*, to which the AI theory refers.

artificial, as *reactive systems*. The mechanisms to perceive stimuli and respond to them are essential for any reactive system. Therefore, we argue that any conscious artificial agent capable of mimicking human consciousness must be able to interact with the environment just as humans do. However, this assumption raises an issue: the algorithms available today for stimulus-response mechanisms are neither accurate enough, nor computationally feasible for mimicking human abilities [Proctor and Vu 2006].

As the focus of this paper is on functions of consciousness rather than stimulus-response mechanisms, and since we do not aim to address the origin of consciousness, we assume the entire agent-environment interaction functionality is an Oracle[2]. Consequently, we can proceed with suggestions on how to implement consciousness in an artificial agent without having to delve into the details of interaction mechanisms.

It is also important to note that given today's processor architectures, any reference to software based artificial agents implicitly pertains to software designed for the Von Neumann machine architecture [von Neumann 1945]. This is fundamentally different from the neural network architecture utilized by the human brain. It is well known [Fausett 1994] that both Von Neumann machines and neural networks are capable of implementing the basic logic operations (AND and NOT) required for universal computation. However, biological neural networks appear to be superior in performance for certain specific tasks, e.g., vision. Tasks such as vision fall under the category of *interaction functions* comprising the Oracle.

[2]The term 'Oracle' is used in theoretical computer science to refer to a black box function that can solve a decision problem in a single operation, regardless of the complexity class. Here, we use the term loosely to mean a black box that can perform any expected function.

Therefore, we ignore these. For the functions of consciousness though, there is no proof or empirical evidence suggesting that the neural network architecture is superior. The one aspect that may benefit from a neural network machinery is *memory*. Our hope is that with the progress in memory and microprocessor technology, this will become less of a barrier to implementing human-like memory function.

In the following Subsection, we introduce the artificial agents that will be used to illustrate the mechanisms and processes involved in artificial consciousness. We use *awareness* and *emotions* as illustrative functions of consciousness.

1. Artificial Agents – Introducing JP and Sander

In order to illustrate the mechanisms of artificial consciousness, we present examples of two hypothetical artificial agents named *JP* and *Sander*. These are both mobile robots that look and behave exactly like humans, i.e., a human cannot identify JP and Sander as robots based solely on physical appearance and behavior. Therefore, our descriptions of these agents are highly anthropomorphized.

C. Awareness in Artificial Agents

A high level software-based model of awareness involves the following components:

- Sensor input and fusion
- Sensor input prioritization
- Attaching the awareness property to the highest priority stimulus

We explain these processes in the following Subsections using hypothetical scenarios centered on the artificial agents introduced earlier.

1. The Process by way of Illustration

Let us consider a scenario where JP is standing at the platform of the Santa Fe train station, waiting for the train. JP is facing the track, looking at the captivating scenery

in the background. At one point, he hears the train coming from the right. He instantly turns his head to the right to see the train. By looking at the oncoming train, he has visual confirmation of his interpretation of the auditory stimulus.

In the above mentioned scenario, as JP sees the train, he also hears the sound that gets louder as the train comes closer. JP's brain is concurrently receiving two stimuli; visual and auditory. The artificial brain fuses the two stimuli to refine the perception of the object. This is because in many cases, one stimulus is not enough to assess the situation completely. E.g., hearing the sound of the train does not guarantee that it is the train for which JP is waiting, and not a different train passing by the station. Similarly, if the train is coming in over a bend, JP might not be able to see it up until a certain point, but might still be able to hear the sound. Therefore, sensor fusion is an important process to increase the amount of information available about a certain situation.

Before JP hears the sound of the train, he can hear many other sounds as well, e.g., the birds chirping, some other passengers talking to each other, etc. As soon he hears the sound of the train, it captures his attention, and he suddenly looks right. Most of the other sounds, even though being perceived, are being assigned low priorities. The sound of the train, however, is assigned a high priority, which enables JP's artificial brain to filter out the rest and focus on that one sound. Prioritization of sensor input plays an important part in the process of attention.

The attention focuses the visual system on acquiring visual information about the train. This results in an increase in the amount of information[3], and therefore an increase in certainty, about the belief that the train is coming. The sound tells JP that a train is coming; the visual perception further tells him that his train is coming. At this point JP is *aware* of the fact that his train is coming. Awareness in this case is an additional property that is assigned to an object once enough information has been acquired (this model is inspired by the theory of Graziano and Kastner [Graziano and Kastner 2011]). The amount of information is increased or acquired using the process of attention.

[3]Note the similarity with the concept of *information* as described in the Integrated Information Theory.

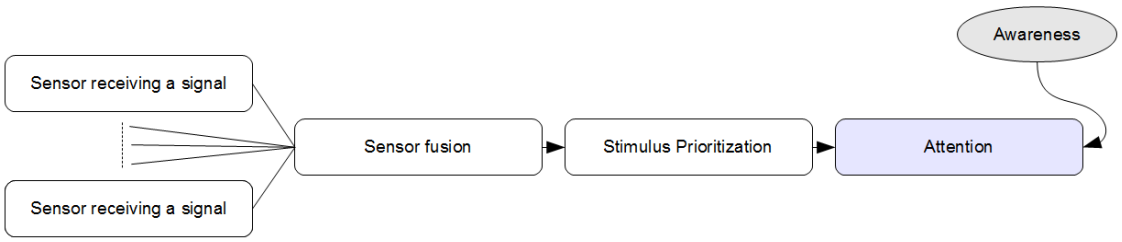

FIG. 4. The process of generation of awareness is depicted. From left, a stimulus is received by one or more sensors and subsequently *fused* and *prioritized*. *Attention* is given to the highest priority stimulus, to which *awareness* is then attached.

2. Sensor Fusion

Sensor fusion is a widely known and commonly used method in robotics. There are several algorithms available for implementing sensor fusion, one of which is based on the Kalman Filter [Kalman 1960]. The important point here is that the algorithms for sensor fusion are available, and we assume that the software module responsible for performing sensor fusion in our agents can be implemented using existing technology.

3. Sensor Input Prioritization

The details of an algorithm for the sensor input prioritization would be too complicated to include in this paper. Nevertheless, the basic concept is presented here. Let us assume that the auditory system (a software module) receives several different inputs. These are:

- Sound of birds chirping
- Sound of music playing
- Sound of people talking amongst a group (of which the observer is not a member)
- Sound of an oncoming train

Each sound is prioritized based on two major properties.

- Loudness
- Type

We assume that the agent is already familiar with all the different sounds. Therefore, when a sound enters the artificial brain, it is compared against a pattern that is stored in the memory, and immediately recognized as belonging to a certain category[4], such as, the sound of birds chirping.

Given a certain state of the artificial mind, e.g., anticipation of the arrival of the train, a certain 'sound pattern' has a high priority. This is based on the *type* of sound. Moreover, the *loudness* can always result in a high priority for any sound if the loudness is above a certain threshold value. E.g., even while JP is anticipating the arrival of the train, Sander shouting at him would capture his attention.

This concept can be distilled down to a very simple algorithm.

```
Loop: forever
        Loop: For each auditory stimulus
                Compute Loudness
                if Loudness > threshold
                        focus attention on the stimulus
                        break Loop
                Compute Type
                if Type in Types of Interest
                        focus attention on the stimulus
                        break Loop
        End Loop
End Loop
```

Please note that even though the computation of loudness is more or less similar in most situations, the computation of the type of interest can vary significantly depending on the given situation, i.e., the state of the artificial mind. E.g., when one is anxiously waiting for a phone call, the sound of phone ringing has a very high priority. However, when one is in a meeting, the sound of phone ringing has a lower general priority, which is further dependent on the loudness.

Note: Whether an algorithm is formulated in terms of if-then-else clauses, or edge weighted directed graphs (such as neural networks), is merely a matter of representation. Both can

[4]The machine learning savvy reader will notice that the process is analogous to *pattern recognition* or *classification learning*.

generate equivalent results [Fausett 1994].

4. Attaching the Awareness Property to a Stimulus

Based on idea of the relationship between awareness and attention presented in the theory of consciousness by Graziano and Kastner [Graziano and Kastner 2011], once the prioritization module assigns the highest priority to a stimulus, the *awareness module* enables the *Is Aware* property for that stimulus. Therefore, the awareness module can be thought of as an independent module that keeps track of attention and enables the awareness property for the stimulus to which the mind is paying attention.

The awareness module can assign awareness object by object in a multi-node chain of objects. This is useful for situations such as: Sander is aware of the fact that JP is aware of the oncoming train, which represents a $2-$deep awareness chain. Moreover, the same awareness module/function can be used to generate self-awareness. In this case, the awareness property is attached the object *self*. This is similar to Graziano and Kastner's idea that self-awareness stems from the ability of the brain to reuse the perceptual social machinery for awareness.

Note: For the sake of simplicity, we have ignored the process of object recognition. Algorithms for object recognition exist [Bennamoun and Mamic 2002, Gong et al. 2000, Hu 1962, Schiele and Crowley 1996], although these do not perform nearly as well as the human brain.

D. Emotions and Internal Loops

1. Emotions as Functions

Let us consider a scenario where JP is taking a walk on the St. John's campus, and he arrives at the Koi pond. He looks at the fish, which triggers a slight change (towards a calming feeling) in his state of mind. What is the process that takes place between *seeing the fish* and *feeling calmer*?

In the previous Section, we discussed the process that takes place between *sensing the stimulus* and *generating attention and awareness*. In the above mentioned scenario, we are adding another stage to the process. Once JP's attention is on the Koi, this information

is forwarded to the *emotion module*. This module implements the *emotional logic*, which is the algorithm used to determine the emotion triggered by a given stimulus. Once a specific emotion (or a set of emotions) has been determined, in addition to possible physiological changes such as a change in facial expression, a change is triggered in the state of mind. This new state of mind (in JP's case) is perceived as *feeling calmer*.

Each emotion can be viewed as a *function*, which is called when a certain set of properties/stimuli is evaluated to be true. The function itself implements an algorithm that constitutes the set of actions to be taken when the function is called. One such action would be the change in facial expression, such as smiling, when an agent perceives happiness. Please note that algorithms circumscribed by these functions change over time (these are governed by an adaptive/learning algorithm), depending on a given state of mind, and also on a long-term or permanent change (possibly caused by damage to the artificial brain) in how a certain stimulus affects the agent.

2. Emotional Logic

The question we will try to answer in this Section is, *"How does a stimulus trigger an emotion in an artificial agent?"*

Once a stimulus has been assigned a high enough priority, information about the stimulus (the associated data structure; the *stimulus object*) is forwarded to the emotion module. This module reads certain properties of the stimulus object, e.g., in the Koi fish example, the *beauty associated with the scene*. The *property scan* function, i.e., the function that reads the properties, is followed by a call to the emotional logic algorithm. This algorithm analyzes list of properties received from the property scan function. In our example, beauty is considered a positive property. In the absence of any negative properties, the algorithm determines the positive emotion that should be triggered by this property. In this simple case, the *feeling calm* emotion is triggered. A highly simplified version of the algorithm is depicted in Figure 5.

In order to represent a more realistic multi-emotional state, a more elaborate decision process would have to be taken into account. The binary branching between positive and negative perception of a property would not be enough. This too is not an issue, since we know from machine learning that *decision trees* [Grabczewski 2014] can be used to represent

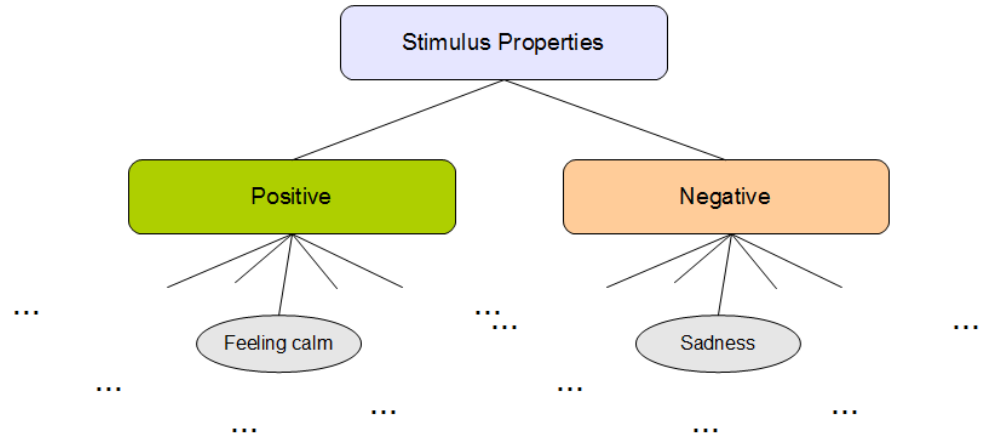

FIG. 5. Simple *emotional logic*: Based on its properties, a stimulus is classified as either positive or negative. Then, the particular emotion to be triggered is determined.

very complex decision processes.

Furthermore, the action associated with an emotion function can also be quite elaborate; in addition to a change in the state of mind, multiple actuators can be activated. Also, the intensity or significance of the stimulus can affect the intensity and/or nature of the response. Please note that all this boils down to a complex set of rules, which can be represented as a set of if-then-else clauses. Other representations such as trees, networks, etc. can be used as well [Witten et al. 2011].

3. Internal Loops

The above discussion raises the question, *"How can an artificial agent experience emotions in a dream? How does an artificial agent feel emotion in the absence of external stimuli?"*

The experiences of day to day life are stored in the agent's memory in the form of patterns. Once the agent is in a dream state, the sensory system can be temporarily replaced by memory, i.e., the stimuli are loaded from memory rather than being received from the sensory system. Once a stimulus is present, it can be processed using the algorithms described earlier, regardless of whether the stimulus was received from the environment, or loaded from memory. A similar mechanism is applicable to how actuators can be perceived as interacting with the environment during dreams while the body is completely at rest. Similarly, an

agent's own thoughts can trigger emotions without any external stimuli.

The above stated explanations imply that there are internal connections between the memory and the other functional modules mentioned earlier. It is these connections that make it possible for the mind to bypass the sensor-actuator modules in the dream states.

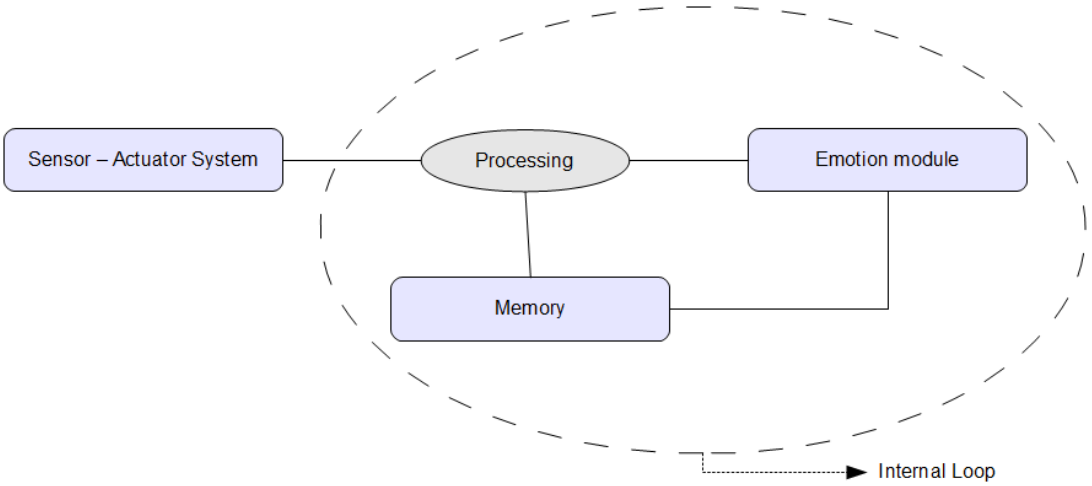

FIG. 6. *Internal loop*: The loop constitutes loading a *stimulus pattern* from memory and processing it. During processing, the emotion module might be used, which might in turn load further stimuli from memory.

Nevertheless, it is important to note that the sensor-actuator system is a necessity for constructing experiences in the first place, which are later used by the mind to create dreams. If the agent never had the sensor-actuator system, it would not have experienced sensory input and actuator responses, and therefore would not have the necessary information in memory to construct dreams.

Please note that the same mechanism is used in situations when the agent is wide awake, but thinking. E.g., consider a scenario where JP is sitting at his desk, developing a simulation in NetLogo[5]. Ideas are generated in his mind as he thinks about how to solve certain problems. In this case, the thinking process does not solely rely on the stimuli currently being perceived; much of the problem solving is done using the information already stored in memory (information retrieval followed by processing to arrive at ideas).

[5]Wilensky, U. 1999. NetLogo. http://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer-Based Modeling, Northwestern University. Evanston, IL.

E. Hierarchical Architecture of an Artificial Brain

The concept of a hierarchical architecture is intrinsically related to the concept of *modularity*. Modular system design is an established practice in software engineering. This makes it possible to independently evolve a module over time, without the propagation of side-effects to other modules. A common method of determining how to divide the code into modules is to look at the coupling relations between different functions. Tightly-coupled, highly interdependent, and functionally similar functions are often included into a single module. On the other hand, loosely-coupled, independent, and functionally dissimilar functions are kept in different modules. There are other more complex decisions that come into play when making these distinctions, but for brevity, only these simple cases are considered here.

There is evidence from neuroscience that certain functionally similar modules in the brain are located in close spatial proximity [Graziano and Kastner 2011]. This indicates that modularity is a common structural property of complex decision making systems, be they artificial or natural.

The architectural model we present of an artificial brain is strongly influenced by the modularity principle employed in software engineering. In order to illustrate how this hierarchical and modular structure functions, we present the following illustration.

JP is running late for a meeting with Sander this morning. He enters the train station, and is pacing toward the platform, so that he can catch the train in time. It is morning rush hour and the station is full of people walking to different platforms, purchasing tickets, etc. As JP is rushing toward the platform, he suddenly hears the sound of a child talking; emanating in close proximity on his right hand side.

1. Sensor-Actuator Layer

The sound is a stimulus that is received at the bottom layer of the artificial brain hierarchy, called the *sensor-actuator layer*. This layer is responsible for interaction with the environment, and consists of *sensors* and *actuators*.

2. Attention and Awareness Layer

Once the stimulus (i.e., sound) is received by the sensor (JP's ears), it is forwarded to the *Attention and Awareness layer*. This triggers the prioritization algorithm that immediately gives the stimulus a high priority. This results in JP's attention being focused on the sound of the child talking.

3. Reactive Layer

Once JP's attention is focused on the sound, it immediately results in him turning his head to the right to see the child. This happens because any stimulus object to which awareness is attached is sent onward to the *reactive layer*. The reactive layer is adaptive in nature, which means that over time it can learn to react to stimuli that are not pre-programmed. *Turning the head toward the sound* is such a learned reaction.

As soon as the stimulus object is received by the reactive layer, the corresponding response is searched for in a map (similar to a hash map data structure), where the key is a stimulus pattern and the value is the corresponding action. The action is a function that can itself consist of a complex algorithm. In the case being discussed here, the action function sends a request to the sensor-actuator layer with an instruction to turn the head right. It also activates the vision system and sets its status to *actively looking for the child object*.

At this point, JP sees the child. The vision system, which is a part of the sensor-actuator layer, receives the stimulus of the child, which is fused with the auditory stimulus, resulting in the strengthening of the belief of the agent that there is a child in close proximity. In addition, sensor fusion results in the computation of estimation of distance between JP and the child. In this case, the distance is small enough that another function in the reactive layer is triggered. This function sends a *strafe left* message to the sensor-actuator layer. As a result, JP suddenly strafes left.

The reactive system can trigger a function not just for the agent's own safety, but also for the safety of the other agents in the environment. The above mentioned reaction where JP strafes left was learned through the social perceptual machinery of the artificial brain. The object of strafing left was to make sure that the child is not harmed.

Note: There can be reactions analogous to the knee-jerk response of the human ner-

vous system. These are pre-programmed reactions or *instinctive reactions* [Minsky 2007] as opposed to *learned reactions*.

4. Decision Making Layer

The significance of causing a child harm is high enough that as the reactive system triggers the *strafe left* action, it also forwards the stimulus object to the emotion module, which is located in the *decision making layer*. The emotion module evaluates the stimulus object and decides to call the *fear* function. The fear function activates the *fear center* (analogous to *Amygdala* [Kandel et al. 2000] in the human brain). The fear center broadcasts a special fear message to the sensory-motor layer, which increases the perceptiveness of the sensory system. This ensures that for a certain short period of time in the future, the agent is hyperalert.

The decision making layer is primarily responsible for: 1) evaluating stimuli (external or internal) that are not processed by the reactive layer, and 2) processing ideas generated inside the artificial brain. In addition to the emotion module, this layer comprises complex functions that implement algorithms for deliberating and assessing situations. A simple example is an agent walking through the station and trying to find its way to the correct platform (assuming it is not familiar with the layout of the station building). At certain points, it is lost, in which case it decides whether to ask someone, or to turn and move in a different direction, etc. The agent may need to make several deliberative decisions in such a situation.

5. Reflective Layer

JP managed to reach the platform in time, and is now seated inside the train. As the agent is in a resting position now, and there are no stimuli that require specific attention, the agent's brain turns attention to the *self object*. There is an *idle time* based algorithm; when the idle time, i.e., the amount of time without any significant stimulus, crosses a certain threshold, awareness is attached to the self object. This is when the control is passed on to the *reflective layer*.

As JP settles down in the seat, he starts to reflect on the incident with the child. He

thinks about the possible outcomes had he not reflexively strafed left in time. This negative feedback results in the thinking process looking back one step in the history of the decision process. JP starts to realize that the situation could have been completely avoided had he not been walking too fast. This leads him to think that the origin of this Markov chain is in *not waking up early enough*.

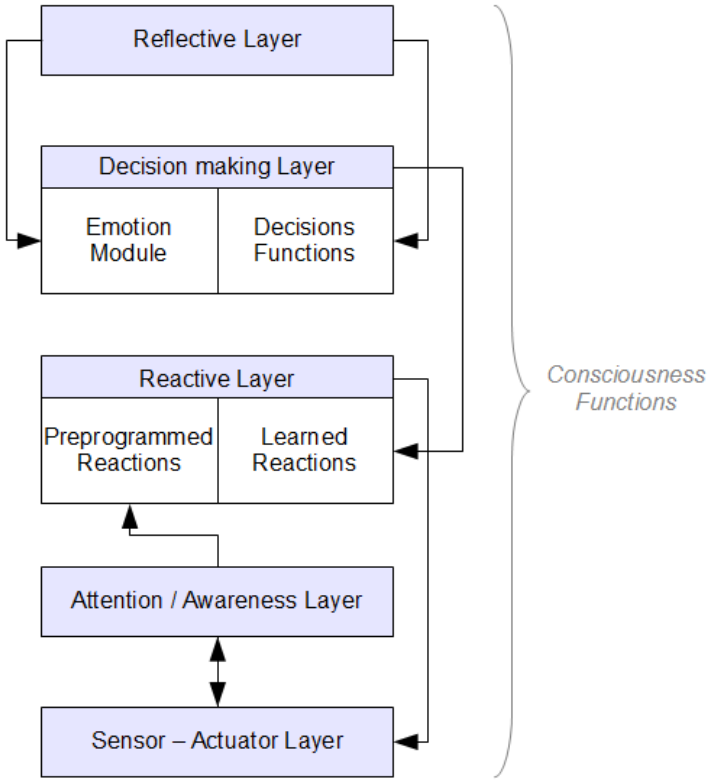

FIG. 7. Hierarchical architecture of the artificial brain. In order to generate conscious behavior, functions in the various layers must be integrated.

The reflective layer constitutes functions that evaluate decisions made earlier. This is similar to the function of *self-reflection* in humans. The purpose of the functions in this layer is to adapt the functions in the other layers according to previous outcomes. This is similar to the feedback loop used in machine learning algorithms [Haykin 2009] and adaptive filters [Sayed 2008]. In certain cases, the feedback results in an update in the functions of the decision layer. In other cases, if the reflective layer determines that the situation was severe enough that the response should be immediate for a similar situation in the future, the corresponding functions of the reactive layer are updated. This results in the creation of a learned reaction.

Please note that the reflective layer can directly influence the function of most of the decision making layer, however, it has only limited access to the emotion module. It is not possible for the reflective layer to directly alter the emotion functions[6]. If this were not the case, unlike humans, the agent would be able to directly manipulate the influence of stimuli on its emotional state. Since we intend for the artificial agent to mimic human consciousness, we have proposed an implementation that corresponds to the working of emotions in humans.

F. The Artificial Mind as a Highly Interconnected State Machine

Whenever we use the term *state of mind*, we (perhaps unintentionally) refer to the mind as a state machine. Given the myriad processes concurrently being performed by the brain at a certain point in time, it is not feasible to enumerate all possible states. Moreover, different states of the mind can be combinations of other different states, which renders the state space practically unexplorable. *How then can such a system be implemented in an artificial agent?*

In this Section we provide a speculative theory of the interconnection architecture that would be required to create an artificial agent with a complex brain, i.e., one with states of mind similar to that of a human brain.

The solution lies in a highly interconnected network of functions, implemented in a hierarchical structure. High interconnectedness is required so that many different functions can call each other without going through an intermediary. Also, depending on the current state and any given set of stimuli, it should be possible to represent the next state as a graph of interacting functions. This can be represented by an edge-weighted graph (possibly a hypergraph) with an edge set of high cardinality, and edge weights representing connection strength. For such an architecture, a *state* would constitute the set of active paths through the graph. The *experience* of the state could then be the particular set of edge weights – assuming that edge weights can change in response to learning).

The above mentioned description of interconnectedness highlights a possible point of comparison with the Integrated Information Theory of consciousness (discussed in Section II A). Perhaps it is the case that when we try to implement consciousness functions, the nature of

[6]This can be implemented by having only indirect connectivity between the emotion module and the reflective layer via other modules and layers.

computation of these functions naturally requires a highly interconnected structure. That is, even if we approach artificial consciousness from a functional rather than structural point of view, a highly interconnected structure emerges as a result of the necessary interactions between functions and layers. This may point to the fact that a high Φ value is indeed a necessary condition for the generation of consciousness.

IV. DISCUSSION

A. Complexes, Linear Inseparability, and Artificial Neural Networks

According to Tononi's Integrated Information Theory [Tononi 2008], a complex is a subset of a system with $\Phi > 0$. Then a complex generates more information as a whole than the sum of the information generated separately by its individual parts (sum is greater than the parts).

Here, we present the idea of *sum is greater than the parts*, i.e., information integration, in terms of implementing a linearly inseparable function in an *Artificial Neural Network (ANN)*. For the purpose of illustration, let us design an ANN that implements the *Exclusive-OR (XOR)* function. The network with correct weight and bias values is shown in Figure 8.

The ANN depicted in Figure 8 takes as input x_1 and x_2, applies the XOR function to the input, and generates the output y. The function f is assumed to be a squashing function that takes as input linear combinations:

$$v_1 = w_1 x_1 + w_2 x_2 + b_1 w_{b1} \tag{3}$$

$$v_2 = w_3 x_1 + w_4 x_2 + b_2 w_{b2} \tag{4}$$

$$v_3 = w_6 f(v_1) + w_7 f(v_2) + b_3 w_{b3} \tag{5}$$

$f(v_1)$ represents the output of the first hidden neuron, and $f(v_2)$ represents the output of the second hidden neuron. Each hidden neuron generates a hyperplane, which in itself is incapable of separating the XOR function. However, when the two hidden neuron outputs

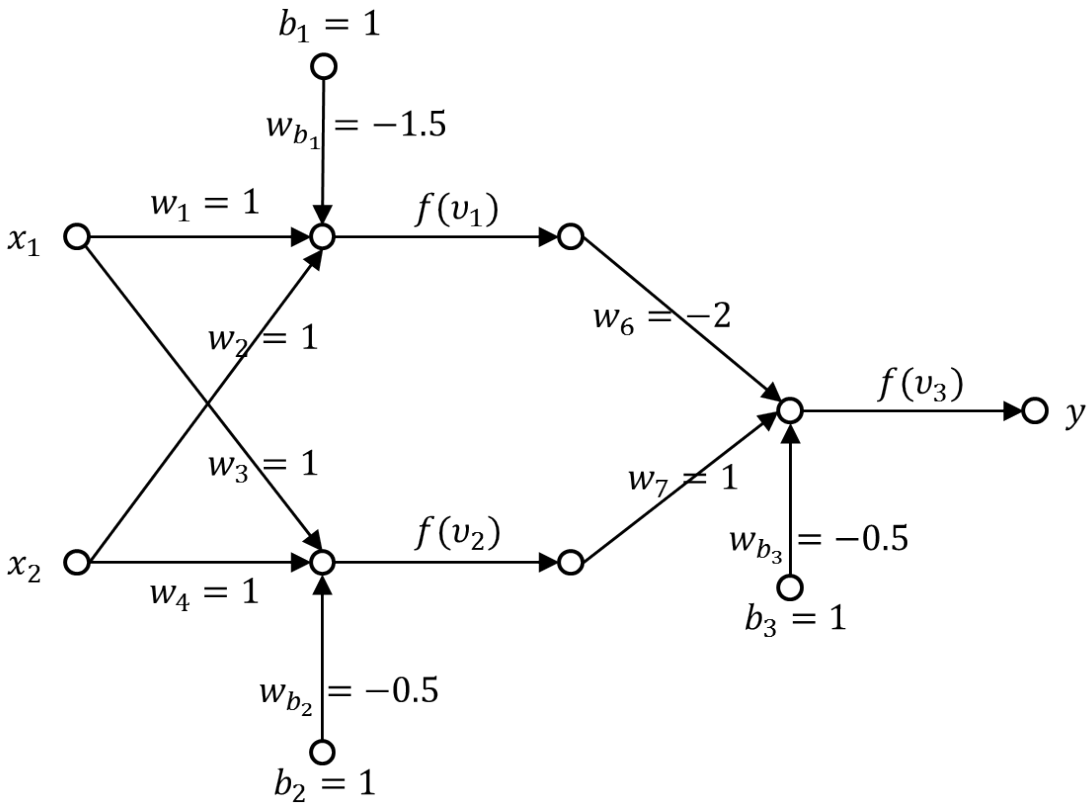

FIG. 8. Artificial Neural Network representing the *exclusive-OR* function.

are combined, the output $y = f(v_3)$ results in a function that can correctly separate the XOR function.

The is a simple example, chosen solely for the purpose of illustration. Nevertheless, the message presented here is more general. Representation of any linearly inseparable function or any nonlinear function requires an ANN structure that is composed of highly interconnected components that are meaningless on their own. For an ANN-based classifier, the amount of information required to correctly classify an instance can only be generated if all the components function as a whole. In the light of this argument, we arrive at the following conclusions:

- Computation of any complex function using a neural network architecture requires information integration. This conclusion agrees with the Integrated Information Theory.
- Complexes with integrated information can be implemented in artificial systems as neural networks representing complex linearly inseparable and/or nonlinear functions.

Assuming these conclusions are correct, the following questions are raised:

- Is a Complex merely a linearly inseparable and/or nonlinear function, represented as an edge weighted directed graph?
- When we design programs that imitate conscious behavior, do they result in highly connected program graphs with high Φ as a matter of course?

At this point, we do not attempt to answer these questions. We leave the reader with the notion that ANNs may not only serve as toy models to empirically analyze Integrated Information Theory, they may also turn out to be ideal data structures for implementing consciousness in artificial agents.

B. Consciousness and Complexity of Computation

If we consider the functions of consciousness described earlier, it can assumed that most of these functions would require complex computation[7] if the agent were to behave like a human. As an example, let us consider a scenario where JP is playing Tennis. He sees the oncoming ball, decides how to adjust his body posture, what kind of shot to play, and where to place the ball in the opponent's court, etc. In addition to finding the best shot that maximizes the chance of winning, JP would have to solve complex numerical problems to figure out which joints in the arm to move, how much to move them, how much force to apply, in which direction to apply the force, how to coordinate eye-hand movement, etc. One can imagine a very large number of parameters to be determined in this case.

If the agent were to try and solve this problem using standard numerical methods based on approximating solutions to equations, the computation would be prohibitively expensive. However, a human can perform these tasks without ever even having studied mathematics. This raises the question: *Do networks of neurons in the human nervous system execute numerical simulations?*

A possible answer lies hidden in the architecture of the biological neural networks, which is not a Von Neumann computer. If we train an artificial neural network to learn a complex nonlinear function, it does so without solving numerical problems. It approximates the

[7] The phrase *complexity of computation* mentioned in this Section does not refer to computational complexity of an algorithm. We use the term loosely, referring to the amount of work required for a certain task.

function by adjusting edge weights over a period of time. In humans, this process of learning is time consuming, and may take years. However, once the skill has been learned, the response becomes immediate, and no complex computation is necessary; the response is encoded in the neuronal structure and interconnection strengths. Moreover, the number of neurons, synapses, and glial cells in the human nervous system is massive. Perhaps this is why it is easier for us to do certain computations so easily, which would be prohibitively expensive for a computer.

The architectural difference, i.e., Von Neumann vs. neural network, as well the difference in mechanism, i.e., numerical simulation vs. learning, may provide some insight into how an artificial agent might be able to handle the above mentioned computations efficiently. Here, we reiterate our hypothesis from the previous Section: *artificial neural networks might constitute the most fertile ground for testing theories of neuroscience through the lens of artificial intelligence.*

C. Empirical Analysis and Comparison of Artificial and Natural Designs

In the previous two Sections, we have identified certain points of possible connections between the Integrated Information Theory, and methods in Artificial Intelligence. Nevertheless, the theory of artificial consciousness presented earlier in this paper is not inspired by IIT, i.e., while proposing methods and architectures, compliance with the structural requirements of *complexes* (described in Section II A) have not been considered.

Earlier, we mentioned the possibility that any attempt to implement conscious behavior might inevitably result in the design of complexes with information integration. However, we neither have proof of this, nor empirical evidence to support the claim. Then, *what if the architecture of an artificial agent does not comply with the structures proposed in the IIT?*

Architecture of any evolvable software system must be modular, maintainable, as well as scalable. Moreover, for an artificial agent to be able to perform complex human-like tasks, its software should be dependable and fault-tolerant. Also, the principles of parallelism and concurrency must be followed, resulting in an efficient implementation.

We argue here, that if the architecture of an artificial agent is not compliant with IIT, and yet produces the conscious behavior similar to that of humans, this would make it possible to perform detailed comparison between the two approaches. It might provide insight useful

for software engineering if we could learn something new from such a comparison. E.g., we might learn that the brain's architecture is more suitable for specific classes of computation.

D. Recognizing Consciousness in Artificial Agents

JP is back at the Santa Fe Railyard on a bright, bustling afternoon to meet his friend Juniper. He walks into Second Street Brewery and sees her sitting at a table already. He sits down next to her and orders a drink. The two begin to chat about the previous week's events at CSSS 2014.

The AI approach presented here is about looking at consciousness in terms of conscious behavior. That is, we do not look to reproduce consciousness, we seek to imitate the functions of consciousness. This begs the question–*how does Juniper, a human, know that the artificial agent JP is conscious?*

As previously stated, we can assume that JP looks and acts exactly like a human, so Juniper cannot immediately determine that he is a robot. We can also assume that our algorithms for artificial consciousness have been perfectly implemented. Therefore, Juniper recognizes consciousness in JP by recognizing the output of the functions of consciousness– his emotions, facial expressions, etc.

V. LIMITATIONS AND FUTURE WORK

In this paper, we have presented mechanisms and architectures that serve as a road map for implementing consciousness in artificial agents, and inform the theories of consciousness in neuroscience. Our approach, nevertheless, has been to keep our discussion at a high level of abstraction.

We realize that in order to uncover major obstacles in the implementation of conscious agents, we need to design and analyze algorithms that uncover low level details, and extend the architecture to various systems involved in conscious behavior that have not been covered in this paper. Therefore, one direction for future research would be to delve deeper into the design process, and move towards implementation of software agents to provide proof of concept.

As pointed out in the Section IV, there are various possible points of connection between

the Integrated Information Theory and our approach to artificial consciousness. These also lead to open questions that are worth pursuing. A high priority task for us would be to use artificial neural networks as toy models for doing an in depth empirical analysis of the structure of complexes and its relationship to information integration. It would be interesting to see how this compares to the concept of modularity, which is a cornerstone of software systems engineering.

ACKLOWLEDGEMENT

This working paper emerged from group work at the Santa Fe Institute Complex Systems Summer School 2014. We wish to thank all the organizers and participants of the summer school, especially those who took part in interdisciplinary discussions about consciousness that helped shape ideas underpinning this paper.

(2010). *The social brain*. The Danish School of Education, Aarhus University, Copenhagen.

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

Arbib, M. A. (1995). *The Handbook of Brain Theory and Neural Networks*. The MIT Press, Cambridge, MA.

Atkinson, A. P., M. S. C. T. e. a. (2000). Consciousness: mapping the theoretical landscape. *Trends in Cognitive Sciences*, 4 (10):372–382.

Bennamoun, M. and Mamic, G. J. (2002). *Object Recognition: Fundamentals and Case Studies*. Springer-Verlag New York, Inc., New York, NY, USA.

Blakemore, S.-J., Boyer, P., Pachot-Clouard, M., Meltzoff, A., Segebarth, C., and Decety, J. (2003). The detection of contingency and animacy from simple animations in the human brain. *Cerebral Cortex*, 13(8):837–844.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies 2*, pages 200–219.

Corbetta, M., Kincade, J. M., Ollinger, J. M., McAvoy, M. P., and Shulman, G. L. (2000). Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nature neuroscience*, 3(3):292–297.

Dennett, D. C. (1991). *Consciousness explained.* Little Brown and Co., Boston, MA.

Fausett, L. (1994). *Fundamentals of Neural Networks-Architectures, Algorithms, and Applications.* Prentice Hall.

Fodor, J. A. (1992). The big idea: Can there be a science of mind? In *Times Literary Supplement.*

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

Frith, U. and Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1431):459–473.

Gazzaniga, M. S. (2005). Forty-five years of split-brain research and still going strong. *Nature Reviews Neuroscience*, 6(8):653–659.

Gong, S., McKenna, S. J., and Psarrou, A. (2000). *Dynamic Vision: From images to face recognition.* Imperial College Press, London.

Grabczewski, K. (2014). *Meta-Learning in Decision Tree Induction (Studies in Computational Intelligence).* Springer.

Graziano, M. S. and Kastner, S. (2011). Human consciousness and its relationship to social neuroscience: a novel hypothesis. *Cognitive neuroscience*, 2(2):98–113.

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

Harnad, S. (1995). Grounding symbolic capacity in robotic capacity. In *Artificial Route to Artificial Intelligence: Building Situated Embodied Agents.*

Haykin, S. S. (2009). *Neural networks and learning machines*, volume 3. Pearson Education Upper Saddle River.

Hofstadter, D. R., Dennett, D. C., and Dennett, D. C. (1981). *The Mind's I: Fantasies and Reflections on Self & Soul.* Basic Books, 0 edition.

Holland, O., editor (2003). *Machine consciousness.* Imprint Academic, New York.

Hu, M.-K. (1962). Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, IT-8:179–187.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45.

Kandel, E. R., Schwartz, J. H., Jessell, T. M., et al. (2000). *Principles of neural science*, volume 4. McGraw-Hill New York.

Luger, G.F., L. J. and Stern, C. (2002). Problem-solving as model refinedemnt: Towards a constructivist epistemology. *Brain, Behavior, and Evolution*, 59:87–100.

McCarthy, J. (1995a). Making robot conscious of their mental states. In *Machine Intelligence*.

McCarthy, J. (1995b). Making robots conscious of their mental states. In *Machine Intelligence 15*, pages 3–17.

Minsky, M. (1991). Conscious machines. In *Machinery of Consciousness*. National Research Council of Canada.

Minsky, M. (2007). *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. SIMON & SCHUSTER.

Moravec, H. P. (1988). Sensor fusion in certainty grids for mobile robots. *AI magazine*, 9(2):61.

Moravec, H. P. (2000). *Robot: Mere Machine to Transcendent Mind*. Oxford University Press, Inc., New York, NY, USA.

Pelphrey, K., Morris, J., and Mccarthy, G. (2004). Grasping the intentions of others: the perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. *Cognitive Neuroscience, Journal of*, 16(10):1706–1716.

Perlis, D. (1997). Consciousness as self-function. *Journal of Consciousness Studies*, 4(5-6):509–525.

Proctor, R. W. and Vu, K.-P. L. (2006). *Stimulus-Response Compatibility Principles: Data, Theory, and Application*. CRC Press.

Pylyshyn, Z. (1984). *Computation and Cognition: Towards a Foundation for Cognitive Science*. MIT Press.

Rudrauf, D., Lutz, A., Cosmelli, D., Lachaux, J.-P., and Le Van Quyen, M. (2003). From autopoiesis to neurophenomenology: Francisco varela's exploration of the biophysics of being. *Biological research*, pages 27–65.

Sayed, A. H. (2008). *Adaptive filters*. John Wiley & Sons.

Schiele, B. and Crowley, J. (1996). Probabilistic object recognition using multidimensional receptive field histograms. In *Proceedings of the 13th International Conference on Pattern Recognition (ICPR 96)*, volume B, pages 50–54.

Schreiber, G. T. and Akkermans, H. (2000). *Knowledge Engineering and Management: The CommonKADS Methodology*. MIT Press, Cambridge, MA, USA.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Shulman, G. L., Pope, D. L., Astafiev, S. V., McAvoy, M. P., Snyder, A. Z., and Corbetta, M. (2010). Right hemisphere dominance during spatial selective attention and target detection occurs

outside the dorsal frontoparietal network. *The Journal of Neuroscience*, 30(10):3640–3651.

Siegwart, R., Nourbakhsh, I. R., and Scaramuzza, D. (2011). *Introduction to Autonomous Mobile Robots*. The MIT Press, 2nd edition.

Simon, H. (1981). *The Sciences of the Artificial*. MIT Press, Cambridge, MA.

Sloman, A. and Chrisley, R. (2003). Virtual machines and consciousness. *Journal of Consciousness Studies*, 10(4-5):133–172.

Sperber, D. (1999). First-person methodologies: What, why, how.

Throop, C. J. and Laughlin, C. D. (2007). Anthropology of consciousness. *Cambridge Handbook of Consciousness*, pages 631–672.

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

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *The Biological Bulletin*, 215(3):216–242.

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

Varela, F. and Shear, J. (1996). Explaining culture: A naturalistic approach. *Journal of Consciousness studies 6*.

von Neumann, J. (1945). First draft of a report on the EDVAC. Technical report, University of Pennsylvania.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.

Wittgenstein, L. (1953). *Philosophical Investigations*. Macmillan, New York.