Applying machine consciousness models in autonomous situated agents

Raúl Arrabales Moreno *, Araceli Sanchis de Miguel

Departamento de Informática, Universidad Carlos III de Madrid, Avda. de la Universidad, 30, 28911 Leganés, Madrid, Spain

Abstract

This paper briefly describes the most relevant current approaches to the implementation of scientific models of consciousness. Main aspects of scientific theories of consciousness are characterized in sight of their possible mapping into artificial implementations. These implementations are analyzed both theoretically and functionally. Also, a novel pragmatic functional approach to machine consciousness is proposed and discussed. A set of axioms for the presence of consciousness in agents is applied to evaluate and compare the various models.

Keywords: Machine consciousness; Situated agents; Attention; Models of mind

1. Introduction to machine consciousness

Throughout centuries philosophers and scientists have developed many theories trying to account for human consciousness. Despite this great historic effort in the search for an explanation for natural consciousness, relatively little effort has been made in the corresponding field of Artificial Intelligence. The scientific advances achieved during the last three decades in the quest for an explanation of consciousness have had a modest influence in bio-inspired artificial systems. However, this matter has attracted growing interest recently, e.g. Holland (2003).

Consciousness can be defined as the relation between attention, reasoning, recognition and behavior, i.e., a conscious being has the capability to pay attention toward something, think about it, wonder what is it, how it is, why is the way it is, etc., with the aim to recognize it. Once the object (or event) is identified, the subject decides what to do with it. The machine consciousness paradigm is inspired by these conscious/unconscious processes observed in the mind of humans and other higher mammals. The main objective of machine consciousness is building artificial systems with the same advantageous abilities and functionalities as those that consciousness has endowed on humans thanks to evolution.

The complex nature of some of the theories of consciousness is the root problem in the process of modeling them computationally. Some of the paradigms applied to account for conscious processes, like quantum mechanics (Hameroff and Penrose, 1996) or electromagnetic components of brain waves (Rakovic, 1990) are practically impossible to reproduce applying classical software techniques. Current implementations discussed here are based on cognitive theories, in which the problem of consciousness is addressed from a functional system-level perspective.

2. Limiting the scope

There exist theories of consciousness mainly based in functional cognitive aspects, like the approaches proposed by Baars (1993) and Dennett (1991), which might be implemented pragmatically using Artificial Intelligence techniques. Nevertheless, given the complexity of the matter, an initial differentiation between the main dimensions of conscious processes has to be defined. Thus, we can clearly establish the aspects of consciousness that we aim to imi-

* Corresponding author. Tel.: +34 916249111; fax: +34 916249129.
E-mail addresses: raul.arrabales@uc3m.es (R. Arrabales Moreno), masm@inf.uc3m.es (A. Sanchis de Miguel).

tate in artificial systems. Established theories of consciousness cover several perspectives or dimensions of consciousness, which are often merged in a confusing fashion. Block (1995) makes a useful distinction between Phenomenal Consciousness (P-Consciousness), Access Consciousness (A-Consciousness), Monitoring Consciousness (M-Consciousness), and Self-Consciousness (S-Consciousness). This quartet comprises all the current scientific perspectives of consciousness. Quite briefly described, P-Consciousness refers to subjective experience or qualia Dennett (1991). A-Consciousness defines the accessibility of contents for reasoning, volition and speech. M-Consciousness is about inner perception or introspection. Finally, the ability of self-recognising and reasoning about the recognized self is called S-Consciousness. In the context of this work we have primarily built on a combination of A-Consciousness, M-Consciousness and S-Consciousness, which we call *Reasoning Consciousness*. According to Block (1995), in some cases it is feasible to have access consciousness without phenomenal consciousness. However, other authors argue that A-Consciousness correlates with P-Consciousness (Chalmers, 1990).

P-Consciousness is the subjective experience that the individual has due to the fact that he/she is conscious, whereas reasoning consciousness is the availability for the use of reasoning, actions and speech. The vast unconscious domain of knowledge and concurrent control routines can be accessed using consciousness as a gateway. The reasoning consciousness dimension is very interesting regarding its possible application in artificial systems. Phenomenal aspects are considered out of the scope of the present analysis.

3. From theories of consciousness to artificial implementations

In the context of the present analysis, we aim to understand how the access to knowledge and perception is managed and how the control of a vast set of complex parallel (unconscious) processes is carried out from a unique sequential (conscious) thread. At any given time there are a great number of neuronal unconscious processes running in parallel in a human brain; however, only certain contents are showed to the consciousness at a particular time, i.e. attention determines the contents of mind that are perceived consciously. Therefore, one of the main features of conscious processes, in contrast with unconscious processes, is that the former are very much limited. In humans conscious mechanisms are based on short term memory and the selection of attention focus. These aspects are clearly limited, two different voluntary actions cannot be done at the same time consciously and working memory cannot manage more than approximately seven elements at the same time, e.g. telephone numbers (Miller, 1956).

Several hypotheses have been developed trying to account for the mechanisms, evolution, function, and features of consciousness (Atkinson et al., 2000). Here we focus on those accounts advocating for processes rather than vehicles, i.e. those which explain consciousness as functional or relational properties of representational vehicles. These kinds of theories, where consciousness is produced by the computations independently of any particular intrinsic properties of the vehicles, are susceptible to being applied in artificial machinery. In order to characterize consciousness, assumptions are usually made on the following aspects: conscious versus unconscious knowledge processing and learning, coherence or coalition mechanisms, goals, and emotions. The way these assumptions are mapped into the machine consciousness domain as functional modules is discussed below.

Depending on the conscious or unconscious nature of the processes, knowledge can be declarative or procedural, localized or distributed, serial or parallel. Knowledge representation is implicit in unconscious processes and explicit in conscious processes. Implicit information is not directly accessible unless interpretative mechanisms are used. Sun (2002) argues that the cognitive processes are structured in these two levels with different mechanisms. Therefore, the results coming from the two levels must be integrated somehow. According to Sun a synergy is obtained between the implicit (unconscious) and explicit (conscious) processing. In (Schacter et al., 1995) evidence of dissociation of different types of knowledge in the brain is analyzed.

Common factors in many theories are concepts like coherence or coalition. These ideas aim to account for the integrative function of consciousness and its emergence from functionally heterogeneous neural networks or separate unconscious processors. As described by Baars (1993) a number of specialized unconscious processors provide information to a global workspace, which coordinates the processors selecting coherent information patterns. Another to some extent analogous view is the 40 Hz synchronized activation of neuron coalitions that Crick and Koch (1991) argued to be the physical base of consciousness in their search for the neural correlates of consciousness. Nevertheless, they refuted this hypothesis after they checked these activations between 35 and 75 Hz in the cerebral cortex and found they were not necessarily related to conscious processes. Also, Damasio et al. (1990) talk about coherence in a similar sense. The reverberation in neuronal areas of sensory convergence integrates information coming from each sense.

Another key aspect of consciousness, very much related with subject goals, is the emotional dimension. According to different psychological theories (Marina, 2002), emotions are the mechanism that humans use to synthesize their situation in the world within the limited scope of their consciousness. As pointed out by Franklin et al. (1998), in a machine consciousness system, feelings or emotions provide an assessment about how the system goals are being accomplished. In humans, emotions influence behavior according to belief systems and personality (Martinez-Miranda and Aldea, 2005). Therefore, under a frustration state some individuals give up their original objectives com-

pletely, while others will opt for trying different alternatives. The significant role that emotions can play is often neglected in machine consciousness models.

4. From metaphors of consciousness to functional design

Metaphors can help us to intuitively understand how consciousness works. Additionally, an analysis of the functions of consciousness can be derived from the study of these metaphors. Two of the most relevant metaphors are briefly introduced in this section. In the Global Workspace Theory Baars (1997) describes a ''theatre'' in which the spotlight represents the focus of consciousness directed by attention. The complete scene corresponds to the working memory, which is the memory system in charge of conscious contents of mind. The information obtained in the spotlight is distributed globally throughout the theater to two types of unconscious processors: the ones forming the audience receive information from the spotlight; whereas, behind scenes, the unconscious contextual systems form the events taking place in the spotlight. The actors compete for appearing in the attention spotlight, in which they appear as completely conscious contents. The spotlight attention selection is done to a great extent behind the scenes. The unconscious processors carry out this selection based on context and sets of beliefs (usually unconscious) that determine the conscious thoughts (the play in scene). Baars also indicates that the spotlight of consciousness is the instrument used by the ''director'' for decision making in the field of working memory guided by goal achievement. Conscious experiences activate unconscious contexts, which help in interpreting future conscious events. In sum, consciousness provides a framework for access to the vast unconscious contents of mind. It seems that the research work done using brain imaging techniques (fMRI, PET, etc.) indicate that this hypothesis could be true (Baars, 2002); however, more neurological analyses are needed to definitely confirm or refute Baars' assumptions.

Another illustrative metaphor is known as the Multiple Draft Model (Dennett, 1991). Dennett proposes an editorial review process metaphor, where coalitions of unconscious processors evolve due to reiterative edition and review until they are presented as the official published content of the mind. Given the highly parallel nature of brain processing, sensory stimuli are processed concurrently. Moreover, the information is continuously being reviewed and ''edited'', i.e. modified in order to adapt it to the new information that is being acquired. The subject is not aware of any of these continuous changes. By contrast, a sequential – already interpreted – version is finally available to subjective experience. Most of the draft versions that competed for consciousness are deleted and only one version, one narrative flow is fixed in memory. Furthermore, contents in memory are also under constant revision.

We have tried to extract a number of reasoning consciousness functional components from the theories of consciousness we have analyzed. We aim to separate the functional dimension from the phenomenal aspects of consciousness, and identify from the former the functions and characteristics that make consciousness an evolutionary advantage for the conscious beings. Baars (1997) identifies nine functions of consciousness according to his theory. The functions identified by Baars are about adaptation, learning, contextualization, access to a self system, prioritization, recruitment of unconscious processors, decision making, error-detection, self-monitoring and optimization. Taking into account these functions, jointly with Dennett's approach, and including a more elaborated affective dimension to the picture, we have defined a set of basic functional components intended to accomplish all the reasoning consciousness functionality: (1) attention, (2) status assessment, (3) global search, (4) preconscious management, (5) contextualization, (6) sensory prediction, (7) modal and multimodal memory management and (8) self-coordination.

Our hypothesis is that these functional components should be integrated in a machine consciousness system in order to have the expected advantages. This component list is specifically related to consciousness. Other cognitive functions required in an intelligent system have not been included in the list. The identified functional components of consciousness are integrated in CERA (Conscious and Emotional Reasoning Architecture), which is a model designed for autonomous robot control (Arrabales and Sanchis, 2006). The functional modules are briefly described as follows: (1) The attention mechanism provides the individual with the capability to pay attention to a determined event or object, and drive its learning and behavior. (2) The status assessment component comprises the ability of the subject to keep a conscious summary of its state. Emotions play this role. (3) Global search implies access to virtually all the knowledge the individual has. This ability is necessary for the access to unconscious control routines and different memory systems. (4) The separation between conscious and unconscious processes has to be based on the differentiation between explicit and implicit knowledge, respectively. In each of these domains a learning capability has to be present, i.e., unconscious implicit learning and conscious explicit learning. Both domains must be coordinated via control and access mechanisms such as attention and global search. (5) Contextualization is needed for the recruitment of adequate unconscious processors by the conscious control system. The retrieval of problem related knowledge is required as part of problem solving processes. Associative memory retrieval mechanism is an example of contextualization. (6) Sensory prediction function provides a continuous unconscious monitoring and prediction process of the information obtained by the senses. When the perceived differs from the expected, the corresponding information has to become conscious. That means that the individual pays attention to the unexpected situation in order to deal successfully with it. (7) Multimodal memory (which corresponds to semantic

memory and working memory), is where all conscious contents from all senses converge temporarily. Consciousness is multimodal. In contrast, modal memories indefinitely keep specific kinds of content (for instance, visual memory). The CERA memory management module is in charge of integrating and providing access to all memory systems. (8) Self-coordination component is in charge of next action selection and it is driven by the accomplishment of the established goals. Mechanisms such as inner talk and introspection could be included in this component as project management tools (where a project is the set of tasks to be done in order to accomplish one or more goals).

There exist several relationships and synergies between the components described above, which we think should also be implemented in any machine consciousness system (see Fig. 1). Novelties require more participation of consciousness for their learning (Grossberg, 2003), i.e. interrelation between modules 1 and 6. Components 5 and 6 represent the top–down and bottom–up control flows, respectively. On one hand, conscious will invoke unconscious processing in order to achieve its goals; On the other hand, unconscious processors that integrate information coming from the senses call the conscious attention in case of an unexpected or novel situation. This is the way machine consciousness provides the means to integrate top–down and bottom-up recognition flows. We believe that coherence and coalition properties of processes mentioned above are covered in component 5 because the association of processors constitutes a kind of coherence at a functional level. Component 7 represents the personal history of the individual. This concept provides the self with the necessary unity to manage its own experience and identity. The mechanism of coordination indicated in component 8 is related to the status assessment module 2 (emotions), because goal and action selection are conditioned by the emotional state.

5. Models of consciousness implementations

There exist several artificial systems partially bio-inspired in the mechanisms of consciousness and attention. Most of them are based on the Global Workspace Theory (Baars, 1993, 1997). The most salient examples are IDA (Intelligent Distribution Agent) (Franklin et al., 1998), LIDA (Learning IDA) (Ramamurthy et al., 2006) and Computational Agent Framework for Consciousness (Moura and Bonzon, 2004). IDA is designed specifically for the optimization of task assignment for US Navy staff, and LIDA is its learning extension, which is intented to be applied to real complex problems that need to be approached in a human-like way. Moura and Bonzon's proposal is potentially a general purpose system. They consider the concepts of plan and condition, and the model is based on deliberation and concurrent context formation.

The obvious progression of machine consciousness implementations is the cognitive design of control systems for autonomous robots, i.e. cognitive robotics. We think that experimentation with physical autonomous agents is key to evaluating and comparing these kinds of machine consciousness architectures. In this domain, a number of recent works are of interest: LIDA-AV (the application of LIDA to an Autonomous Vehicle), Brain-Inspired Architecture for Cognitive Robotics (Shanaham, 2005), and the aforementioned CERA architecture. However, with the current state of the art in this domain, where architectures for consciousness are not yet fully mature, a direct empirical evaluation based on real world problems is unfeasible. What can be done at this stage is a qualitative analysis to evaluate these concept-models from a functional design standpoint. We propose testing them against to a set of axioms for minimal consciousness in agents (Aleksander and Dunmall, 2003). The minimal axiomatic base is composed of the following axioms: I. Depiction

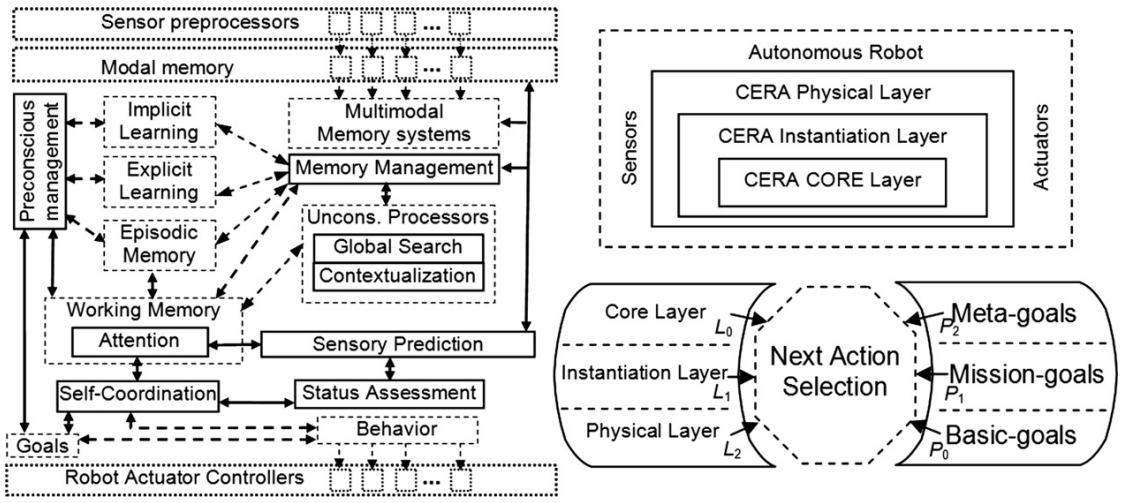

Fig. 1. In the left diagram solid lines represent CERA Core modules. Dashed lines represent CERA instantiation layer (domain-specific modules). Dotted lines represent CERA physical layer. Right diagrams illustrates CERA layered design and next action selection contributions.

(agent has perceptual states), II. Imagination (agent is conscious of imagined as well as perceived events). III. Attention (agent is capable of selecting parts of its perception or imagination), IV. Planning (agent has control over imaginational states in order to plan actions), and V. Emotion (agent has affective states to evaluate actions). As discussed in Arrabales et al. (2007), we believe the CERA model can fulfill all these minimal axioms. Moura and Bonzon's proposal does not fully meet axioms II and IV. Additionally, axiom V is not considered yet. Generally, all the approaches based on the Global Workspace Theory cover axioms I, III and V, as they directly correlate to Baars' account. Imagination and planning over imaginational states are usually lacking in these models. For instance, the LIDA cognitive cycle does not include an explicit imaginational stage at any step between perception and action, i.e. action selection is influenced by feelings and emotions, but goal contexts are only activated as per present perception. No possible (imagined) future is taken into account for goal context formation and emotional evaluation.

6. Conclusions

Consciousness can be the way the nervous system has evolved to deal with novel and unexpected events in the world. This idea surpasses old situational systems, in which an autonomous agent has no need of inner states or knowledge. A common denominator of all the analyzed models is that consciousness is supposed to emerge from functional interaction between specialized modules, rather than having a specialized consciousness module. Consciousness is considered a global event produced in distributed parts of the (artificial) brain. The great challenge in the field of machine consciousness is the design of such a brain. Two main approaches exists: imitating the human nervous system at the neurophysiological level, e.g. CyberChild (Cotterill, 2003) and applying a system level brain modeling, i.e. modeling brain functional areas and their interaction. The latter approach, which corresponds to most of the implementations discussed above, lies between physiological and psychological levels. Which one is the right approach? We think both have important flaws. On the one hand, we do not have an accurate knowledge about all neurophysiological mechanisms. On the other hand, the implementations presented in this paper are primarily based on metaphors that simply help us understand in a holistic way how the human mind works. Actually, a simple metaphor is far away from an established body of scientific knowledge. However, it can be useful as a tool for orienting the research in particular directions, which will confirm or discard the original hypotheses.

Even though the presented functional analysis is just a simple approximation, it provides heuristic clues about what functions of consciousness are key for the improvement of learning and adaptive capabilities of a situated agent. Ideally, an evaluation of an artificial system should be done adding and removing functional components. We propose this as a future project: CERA can be used as a modular test-bed system suitable to evaluate the impact of each consciousness function (and the corresponding relationships between functions). Such a system has to be applied to real world problems, where autonomous robots interact, collaborating and competing in an evolving ecosystem.

References

Aleksander, I., Dunmall, B., 2003. Axioms and tests for the presence of minimal consciousness in agents. J. Conscious. Stud. 10 (4–5).

Arrabales, R., Sanchis, A., 2006. A machine consciousness approach to autonomous mobile robotics. In: 5th International Cognitive Robotics Workshop, AAAI-06.

Arrabales, R., Ledezma, A., Sanchis, A., 2007. Modeling consciousness for autonomous robot exploration. In: IWINAC 2007.

Atkinson, A., Thomas, M., Cleeremans, A., 2000. Consciousness: Mapping the theoretical landscape. In: Trends in Cognitive Sciences. No. 4.

Baars, B.J., 1993. A Cognitive Theory of Consciousness. Cambridge University Press, New York.

Baars, B.J., 1997. In the theatre of consciousness: Global workspace theory, a rigorous scientific theory of consciousness. J. Conscious. Stud. (4), 292–309.

Baars, B.J., 2002. The conscious access hypothesis: Origins and recent evidence. Trends Cogn. Sci. (6), 47–52.

Block, N., 1995. On a confusion about a function of consciousness. Behav. Brain Sci. (18), 227–287.

Chalmers, D., 1990. Availability: The cognitive basis of experience. In: Block, N., Flanagan, O., Guzeldere (Eds.), The Nature of Consciousness: Philosophical Debates. The MIT Press, Cambridge, MA, pp. 3–15.

Cotterill, R., 2003. Cyberchild. A simulation test-bed for consciousness studies. J. Conscious. Stud. 10 (4–5).

Crick, F., Koch, C., 1991. Towards a Neurobiological Theory of Consciousness. Technical Report 9, Computation and Neural Systems, Pasadena, CA.

Damasio, A., Damasio, H., Tranel, D., Brandt, J., 1990. Neural regionalization of knowledge access: Preliminary evidence. Symp. Quant. Biol. 55, 1039–1047.

Dennett, D.C., 1991. Consciousness Explained. Little, Brown and Co., Boston.

Franklin, S., Kelemen, A., McCauley, L., 1998. Ida: A cognitive agent architecture. IEEE Conf. on Systems Man and Cybernet., 14.

Grossberg, S., 2003. The brain's cognitive dynamics: The link between learning, attention, recognition, and consciousness. In: Palade, V., Howlett, R.J., Jain, L.C. (Eds.), Knowledge-based Intelligent Information and Engineering Systems, LNCS, vol. 2773, Springer, pp. 5–12.

Hameroff, S., Penrose, R., 1996. Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness. In: Hameroff, S., Kaszniak, A., Scott, A. (Eds.), Toward a Science of Consciousness. MIT Press.

Holland, O., 2003. Machine Consciousness. Imprint Academic, UK.

Marina, J.A., 2002. El laberinto sentimental. Anagrama, Barcelona.

Martinez-Miranda, J., Aldea, A., 2005. Emotions in human and artificial intelligence. Comput. Human Behav. 21 (2), 323–341.

Miller, G., 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychol. Rev. 63, 81–97.

Moura, I., Bonzon, P., 2004. A computational framework for implementing baars global workspace theory of consciousness. In: BICS 2004, Scotland.

Rakovic, D., 1990. Neural networks vs. brainwaves: Prospects for cognitive theory of consciousness. In: Engineering in Medicine and Biology Society, IEEE.

Ramamurthy, U., Baars, B., D'Mello, S.K., Franklin, S., 2006. Lida: A working model of cognition. In: Danilo Fum, F.D.M., Stocco, A. (Eds.), Cognitive Modeling. Edizioni Goliardiche, pp. 244–249.

Schacter, D.L., Reiman, E., Uecker, A., Roister, M.R., Yun, L.S., Cooper, L.A., 1995. Brain regions associated with retrieval of structurally coherent visual information. Nature 376, 587–590.

Shanaham, M., 2005. Consciousness, emotion, and imagination. A brain-inspired architecture for cognitive robotics. In: AISB Workshop Next Generation Approaches to Machine Consciousness.

Sun, R., 2002. Duality of the Mind. Lawrence Erlbaum, Mahwah, NJ.