International Journal of Machine Consciousness

CONSCIOUSNESS IS COMPUTATIONAL: THE LIDA MODEL OF GLOBAL WORKSPACE THEORY

Bernard J. Baars and Stan Franklin

We argue that the functions of consciousness are implemented in a bio-computational manner. That is to say, the conscious as well as the non-conscious aspects of human thinking, planning, and perception are produced by adaptive, biological algorithms. We propose that machine consciousness may be produced by similar adaptive algorithms running on the machine

Global Workspace Theory is currently the most empirically supported and widely discussed theory of consciousness. It provides a high-level description of such algorithms, based on a large body of psychological and brain evidence. LIDA provides an explicit implementation of much of GWT, which can be shown to perform human-like tasks, such as the interactive assignment of naval jobs to sailors. Here we provide brief descriptions of both GWT and LIDA in relation to the scientific evidence bearing on consciousness in the brain. A companion article explores how this approach could lead to machine consciousness (Franklin et al 2009 to appear).

We also discuss the important distinction between volition and consciously mediated action selection, and describe an operational definition of consciousness via verifiable reportability. These are issues that may well bear on the possibility of machine consciousness.

1. Introduction

Progress in machine consciousness will depend on our understanding of consciousness in humans. Consciousness must be understood in the context of its role in cognition. This will require an appropriate high-level theory embodying a suitable conceptual framework, one that respects both the psycho-biological evidence and its computational functions. Such a framework is provided by Global Workspace Theory (GWT), a proposed information-processing architecture designed to account for the contrasting conscious and unconscious features of human cognition (Baars, 1988, 2002; Baars and Franklin, 2003). Though originally motivated by blackboard models from artificial intelligence, it currently enjoys considerable empirical support in brain and psychological science. A striking feature of GWT is that it accounts for both the massive parallel processing of the human brain, most of which is not conscious (i.e., not reportable), and the surprisingly narrow moment-to-moment capacity of the conscious stream of thought.

The LIDA model is both a conceptual and computational fleshing out of a major portion of GWT. The *conceptual* LIDA partially specifies the adaptive algorithms from this portion, while the as yet partially implemented *computational* LIDA completely specifies them.

In addition, this paper is intended to provide brief descriptions of both GWT and its LIDA implementation, together with pointers to more detailed information. We also discuss the important distinction between volition and consciously mediated action selection, and describe an operational definition of consciousness via verifiable reportability. We believe that a framework like GWT and LIDA provides insights such as these, that may play an important role in achieving machine consciousness.

2. Global Workspace Theory

The idea that consciousness has an integrative function has a long history. Global Workspace Theory suggests a fleeting memory capacity in which only one consistent content can be dominant at any given moment. Dominant information is widely distributed in the brain. This makes sense in a nervous system viewed as a massive distributed set of specialized networks. In such a system, coordination, control, and novel problem solving could take place by way of a central information exchange, allowing some regions – such as sensory cortex – to distribute information to the whole. This strategy is particularly useful for novel problems that do not have a known solution, and which may be solved by the collaborative and/or competitive activation of numerous specialized networks, each of which may present a partial step towards a solution. Such an approach works well in some large-scale computer architectures, which show typical 'limited capacity' behavior when information flows by way of a global workspace (Baars, 1988, 2002)

Hence, Global Workspace Theory (GWT) is a cognitive architecture with an explicit role for consciousness in humans (Baars 1983, 1988). It makes the following assumptions:

(i) That the brain may be viewed as a collection of distributed *specialized networks* (processors);
(ii) That consciousness is associated with a *global workspace* -- a fleeting memory capacity whose focal contents are widely distributed ("broadcast") to many unconscious specialized networks;
(iii) That some unconscious networks, called *contexts*, shape conscious contents (for example, unconscious parietal maps of the visual field modulate feature cells needed for conscious vision);
(iv) That such contexts may work together to jointly constrain conscious events;
(v) That motives and emotions can be viewed as part of *goal contexts*;
(vi) That executive functions work as *hierarchies of goal contexts*.

A sizable body of evidence suggests that consciousness is the primary agent of such a global

access function in humans and other mammals (Baars, 1988, 1997, 2002, etc). The idea is now favored by a number of scientists and philosophers. like Daniel Dennett and others (Damasio 1989, Edelman 1989, Freeman 1991, Edelman and Tononi 1999, Kanwisher 2001, Dennett 2001, Dehaene. and Naccache 2001, John et al.(2001, Llinas and Ribary 2001, Rees 2001, Varela et al. 2001).

GW theory generates explicit predictions for conscious aspects of perception, emotion, motivation, learning, working memory, voluntary control, and self systems in the brain. It has similarities to biological theories such as Neural Darwinism (Edelman 1987) and dynamical theories of brain functioning (Skarda and Freeman 1987). Functional brain imaging now shows that conscious cognition is distinctively associated with wide spread of cortical activity, notably toward frontoparietal and medial temporal regions. Unconscious comparison conditions tend to activate only local regions, such as visual projection areas. Frontoparietal hypometabolism is also implicated in unconscious states, including deep sleep, coma, vegetative states, epileptic loss of consciousness, and general anesthesia. These findings are consistent with the GW hypothesis, which is now favored by a number of scientists and philosophers (see above).

Based on this perspective, Baars and Franklin have made a number of testable empirical proposals about the role of consciousness in Baddeley-type working memory, in spontaneous recall, in the multiple memory systems of the human brain, in action control, and other major brain functions. (Baars & Franklin, 2003, Franklin et al 2005, Baars Ramamurthy and Franklin 2007). In addition, Stanislas Dehaene and his research group have modeled and found fMRI evidence consistent with GWT (Dehaene and Naccache 2001), and Shanahan and Baars have suggested a neural net implementation of GWT to deal with the challenging Frame Problem in AI and robotics. Shanahan amd Baars 2005).

3. The LIDA Model and its Architecture

The LIDA model is a comprehensive, conceptual and computational[1] model covering a large portion of human cognition. Based primarily on Global Workspace theory (Baars 1988), the model implements and fleshes out a number of psychological and neuropsychological theories including situated cognition (Varela et al. 1991), perceptual symbol systems (Barsalou 1999), working memory (Baddeley and Hitch 1974), memory by affordances (Glenberg 1997), long-term working memory (Ericsson and Kintsch 1995), Sloman's cognitive architecture (1999), and transient episodic memory (Conway 2001).

[1] At this writing the LIDA model is not yet completely implemented. We claim it as a computational model since each of its modules and most of its processes have been designed for implementation.

The LIDA computational architecture, derived from the LIDA cognitive model, employs several modules that are designed using computational mechanisms drawn from the "new AI." These include variants of the Copycat Architecture (Hofstadter and Mitchell 1995, Marshall 2002), Sparse Distributed Memory (Kanerva 1988, Rao and Olac 1998), the Schema Mechanism (Drescher 1991, Chaput et al. 2003), the Behavior Net (Maes 1989, Tyrrell 1994), and the Subsumption Architecture (Brooks 1991).

The LIDA model and its ensuing architecture are grounded in the LIDA cognitive cycle. Every autonomous agent (Franklin and Graesser 1997), be it human, animal, or artificial, must frequently sample (sense) its environment and select an appropriate response (action). More sophisticated agents, such as humans, process (make sense of) the input from such sampling in order to facilitate their decision making. The agent's "life" can be viewed as consisting of a continual sequence of these cognitive cycles. Each cycle constitutes a unit of sensing, attending and acting. A cognitive cycle can be thought of as a moment of cognition -- a cognitive "moment." Higher-level cognitive processes are composed of many of these cognitive cycles, each a cognitive "atom."

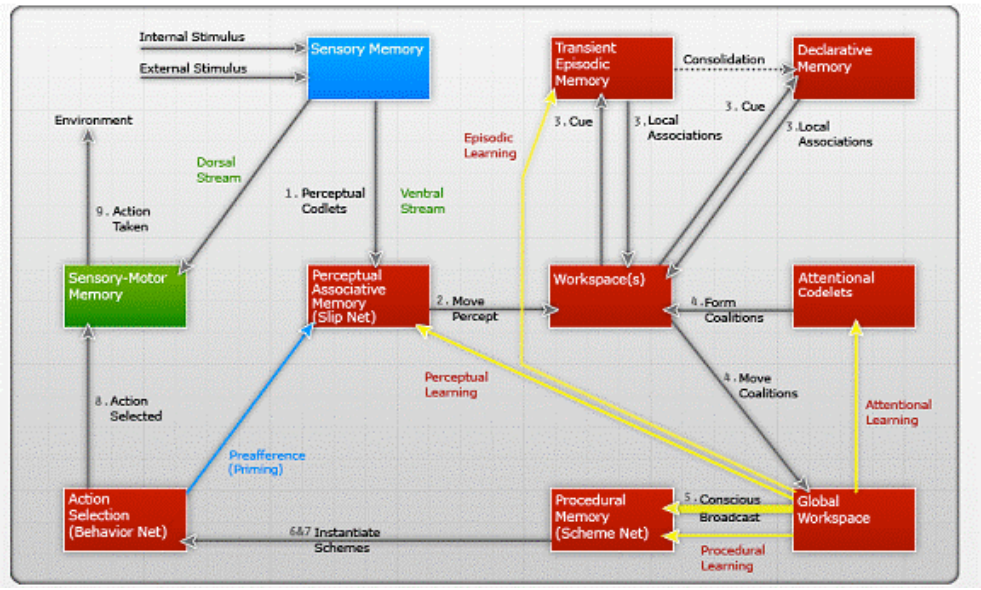

Figure 1. The LIDA Cognitive Cycle

Just as atoms are composed of protons, neutrons and electrons, and some of these are composed of quarks, bosons, muons, etc., these cognitive "atoms" have a rich inner structure. What the LIDA model hypothesizes as the rich inner structure of the LIDA cognitive cycle will now be described briefly. More detailed descriptions are available elsewhere (Baars and Franklin 2003, Franklin et al. 2005).

During each cognitive cycle the LIDA agent first makes sense of its current situation as best as it can *by updating its representation of its world, both external and internal*. By a competitive process, as specified by Global Workspace Theory, it then decides what portion of the represented situation is most in need of attention. Broadcasting this portion, the current contents of consciousness, enables the agent to finally chose an appropriate action and execute it. Thus, the LIDA cognitive cycle can be subdivided into three phases, the understanding phase, the consciousness phase, and the action selection phase. Figure 1 should help the reader follow the description. It starts in the upper left corner and proceeds roughly clockwise.

Beginning the understanding phase, incoming stimuli activate low-level feature detectors in Sensory Memory. The output is sent to Perceptual Associative Memory where higher-level feature detectors feed in to more abstract entities such as objects, categories, actions, events, etc. The resulting percept is sent to the Workspace where it cues both Transient Episodic Memory and Declarative Memory producing local associations. These local associations are combined with the percept to generate a current situational model, the agent's understanding of what's going on right now.

Attention Codelets begin the consciousness phase by forming coalitions of selected portions of the current situational model and moving them to the Global Workspace. A competition in the Global Workspace then selects the most salient, the most relevant, the most important, the most urgent coalition whose contents become the content of consciousness that are broadcast globally.

The action selection phase of LIDA's cognitive cycle is also a learning phase in which several processes operate in parallel. New entities and associations, and the reinforcement of old ones, occur as the conscious broadcast reaches Perceptual Associative Memory. Events from the conscious broadcast are encoded as new memories in Transient Episodic Memory. Possible action schemse, together with their contexts and expected results, are learned into Procedural Memory from the conscious broadcast. Older schemes are reinforced. In parallel with all this learning and using the conscious contents, possible action schemes are recruited from Procedural Memory. A copy of each such is instantiated with its variables bound and sent to Action Selection, where it competes to be the behavior selected for this cognitive cycle. The selected behavior triggers Sensory-Motor Memory to produce a suitable algorithm for the execution of the behavior. Its execution completes the cognitive cycle.

4. Volition vs. Consciously Mediated Action

In this section we'll be concerned with conscious, volitional decision making (volition for short), a higher-level cognitive process for conscious action selection. To understand volition we must carefully distinguish it from 1) consciously mediated action selection, 2) automatized action selection, 3) alarm-driven interrupts, and 4) the execution of actions. Each of the latter three is performed unconsciously, but often has conscious consequences. A conscious machine would, presumably, be capable of at least the first of these, as well as of volition. We'll take these four up individually before moving on to volition. But first a few examples will help ground the discussion.

Consciously planning a driving route from a current location to the airport is an example of deliberative, volitional decision making. Choosing to turn left at an appropriate intersection along the route requires information about the identity of the cross street acquired consciously, but the choice itself is most likely made unconsciously -- on our account, the choice was *consciously mediated* even though it was unconsciously made. While driving along a straight road with little traffic, the necessary slight adjustments to the steering wheel are typically automatized actions selected completely unconsciously. They are usually not even consciously mediated, though unconscious sensory input is used in their selection. If a car cuts in front of the driver, often he or she will have turned the steering wheel and pressed the brake simultaneously with becoming conscious of the danger. An alarm mechanism has unconsciously selected appropriate actions in response to the challenge. The actual turning of the steering wheel, how fast, how far, the execution of the action, is also performed unconsciously though with very rapid sensory input.

Though heavily influenced by the conscious broadcast (the contents of consciousness), action selection during a single cognitive cycle in the LIDA model is not performed consciously. A cognitive cycle is a mostly unconscious process. When speaking, for example, a person usually does not consciously think in advance about the structure and content of the next sentence, and is sometimes even surprised at what comes out. When approaching the intersection in the example above, no conscious thought need be given to the choice to turn left. Consciousness serves to provide information on which such action selection is based, but the selection itself is done unconsciously after the conscious broadcast (Negatu and Franklin 2002). We refer to this very typical single cycle process as *consciously mediated action selection*.

A runner on an unobstructed sidewalk may only pay attention to it occasionally to be sure it remains safe. Between such moments he or she can attend to the beauty of the fall leaves or the music coming from the iPod. The running itself has become automatized, just as the adjustments to the steering wheel in the example above. In the LIDA model such automatization occurs over time with each stride initiating a process that unconsciously chooses the next. With childhood practice the likelihood of conscious mediation between

each stride and the next diminishes. Such automatization in the LIDA model (Negatu, McCauley and Franklin in review) is implemented via pandemonium theory (Jackson 1987).

Sloman (1998) has emphasized the need for an alarm-interrupt mechanism such as that described in the driving example above. The LIDA model implements alarms via learned perceptual memory alarm structures, bypassing the workspace and consciousness, and passing directly to procedural memory. There the appropriate scheme is instantiated immediately into sensory-motor memory, bypassing action selection. This alarm-interrupt mechanism runs unconsciously in parallel with the current, partly conscious, cognitive cycle.

The modes of action selection discussed above operate over different time scales. Volition may take seconds, or even much, much longer. Consciously mediated actions are selected roughly five to ten times every second, since each cognitive cycle takes roughly 300 ms and some three such cycles may overlap in a cascade fashion. Automatized actions are as fast as that, or faster. Alarm mechanisms seem to operate in the sub 50 ms range. In contrast, the execution of an action requires sensory motor communication at roughly ten to forty times a second, all done subconsciously (Goodale and Milner 2004). Let's now look at volition.

We now return to a consideration of deliberative, volitional decision making, having distinguished it from other modes of action selection and execution. Global Workspace Theory specifies that volition occurs via William James' ideomotor theory of volition (1890). James uses an example of getting out of bed on a cold winter morning to effectively illustrate this theory, but in this age of heated homes we will use thirst as an example. James' theory can be interpreted to include processes such as proposers, objectors, and supporters as actors in the drama of acting volitionally. He might have suggested the following scenario in the context of dealing with a feeling of thirst. The idea of drinking orange juice “pops into mind,” propelled to consciousness by a proposer motivated by a feeling of thirst and a liking for orange juice. “No, it's too sweet,” asserts an objector. “How about a beer?” says a different proposer. “Too early in the day,” says another objector. “Orange juice is more nutritious,” says a supporter. With no further objections, drinking orange juice is volitionally selected.

Baars incorporated ideomotor theory directly into his global workspace theory (1988 chapter 7). The LIDA model fleshes out volitional decision making via ideomotor theory within Global Workspace Theory (Franklin 2000) as follows. An idea “popping into mind” in the LIDA model is accomplished by the idea being part of the conscious broadcast of a cognitive cycle, that is, part of the contents of consciousness for that cognitive moment. These contents are the information contained within the winning coalition for that cycle. This winning coalition was gathered by some attention codelet (see above). Ultimately, this attention codelet is responsible for the idea “popping into mind.” Thus we implemented the imaginary characters in James’ scenario as attention codelets, with some acting as proposers, others as objectors, and others as supporters. In the presence of a thirst node in the workspace, one such attention codelet, a proposer codelet, wants to bring drinking orange

juice to mind, that is, to consciousness. Seeing a "let's drink orange juice" node in the workspace, another attention codelet, an objector codelet, wants to bring to mind the idea that orange juice is too sweet. Supporter codelets are implemented similarly.

But, how does the conscious thought of "Let's drink orange juice," lead to a "let's drink orange juice" *node* in the Workspace? Like every higher-order cognitive process in the LIDA model, volition occurs over multiple cycles, and is implemented by a behavior stream in the Action Selection module. This volitional behavior stream is an instantiation of a volitional scheme in Procedural Memory. Whenever a proposal node in its context is activated by a proposal in the conscious broadcast, this volitional scheme instantiates itself. The instantiated volitional scheme, the volitional behavior stream, is incorporated into the Action Selection mechanism. The first behavior in this volitional behavior stream sets up the deliberative process of volitional decision making as specified by ideomotor theory, including writing the "let's drink orange juice" node to the Workspace[3].

Our fleshing out of ideomotor theory in the LIDA model includes the addition of a timekeeper codelet, created by the first behavior in the volitional behavior stream. The timekeeper starts its timer running as a consequence of a proposal coming to mind. When the timer runs down, the action of the proposal contends in the behavior net to be the next selected action, with the weight (activation) of deliberation supporting it. The proposal is most likely to be selected barring an objection or an intervening crisis. The appearance of an objection in consciousness stops and resets the timer, while that of a supporter or another proposal restarts the timer from a new beginning. Note that a single proposal with no objection can be quickly accepted and acted upon.

Here we propose that any conscious machine should be capable of both volition and consciously mediated action.

5. Consciousness as Verifiable Reportability

Psychological and neuroscientific studies of conscious vs. unconscious perceptual events demand an operational definition of consciousness, suitable for an experimental context. Accurate, verifiable report (Baars 1988) is widely accepted as an operational criterion for the consciousness perception of an event. An experimental human subject reports having seen a triangle on the screen. By verifying that, indeed, a triangle had been presented on the screen, an experimenter can conclude that the subject had been consciously aware of the perceptual event of a triangle on the screen. Verifiable reportability can also be used in animal studies,

[3] Alternatively, this node could arrive in the workspace with the percept of the following cycle as a result of internal sensing of the internal speech. In LIDA, this is only an implementation matter, making no functional difference. In humans this is an empirical matter to be decided by experiment. Thus the design decision for LIDA becomes a cognitive hypothesis.

for example with macaques (Seth, Baars, and Edelman 2005). The monkey is trained to report by pressing a particular button when presented with a specific image.

Here we suggest that verifiable reportability may prove to be a useful necessary, but by no means sufficient, criteria for a machine's being conscious of a perceptual event.

6. Conclusions

Progress in machine consciousness will depend on our understanding of consciousness in human brains. We argue that the functions of biological consciousness are implemented in a bio-computational manner. That is to say, the conscious (as well as the non-conscious) aspects of human thinking, planning and perception are produced by adaptive, biological algorithms. We propose that machine consciousness may be produced by similar adaptive algorithms running on the machine.

Global Workspace Theory provides a high-level description of such algorithms, based on a large body of psychological and brain evidence. LIDA provides an explicit implementation of a part of GWT, which has been shown to perform very challenging human tasks. Here we provide brief descriptions of both GWT and LIDA in relation to the scientific evidence bearing on consciousness in the brain.

References

1. Baars, B J. 1983. Conscious contents provide the nervous system with coherent, global information. In *Consciousness & self-regulation*, ed. R J Davidson, G E Schwartz and Daniel O Shapiro:41. New York: Plenum Press.
2. Baars, Bernard J. 1988. *A cognitive theory of consciousness*. Cambridge: Cambridge University Press.
3. Baars, Bernard J and Stan Franklin. 2003. How conscious experience and working memory interact. *Trends in Cognitive Science* 7: 166–172.
4. Baars, Bernard J, Uma Ramamurthy, and Stan Franklin. 2007. How deliberate, spontaneous and unwanted memories emerge in a computational model of consciousness. In *Involuntary memory: New perspectives in memory approach*, ed. John H. Mace:177-207. Oxford: Blackwell.
5. Baddeley, A D and G J Hitch. 1974. Working memory. In *The psychology of learning and motivation*, ed. G A Bower:47–89. New York: Academic Press.
6. Barsalou, L W. 1999. Perceptual symbol systems. *Behavioral and Brain Sciences* 22: 577–609.
7. Brooks, R.A. 1991. Inteligence without represention. *Artificial Intelligence* 47: 139-159.
8. Chaput, Harold H., Benjamin Kuipers, and Risto Miikkulainen. 2003. Constructivist learning: A neural implementation of the schema mechanism. In *Proceedings of WSOM '03: Workshop for Self-Organizing Maps*. Kitakyushu, Japan.
9. Conway, Martin A. 2001. Sensory–perceptual episodic memory and its context: Autobiographical memory. *Philos. Trans. R. Soc. Lond B.* 356: 1375–1384.
10. Damasio, A.R. (1989) Time-locked multiregional retroactivation: a systems-level proposal for the neural substrates of recall and recognition. *Cognition*33, 25–62

11. Dehaene, S. and Naccache, L. (2001) Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*79, 1–37
12. Dennett, D. (2001) Are we explaining consciousness yet? *Cognition*79, 221–237
13. Drescher, Gary L. 1991. Made-up minds: A constructivist approach to artificial intelligence. Cambridge, MA: MIT Press.
14. Edelman, Gerald M. 1987. *Neural Darwinism*. New York: Basic Books.
15. Edelman, G.M. (1989) *The Remembered Present*, Basic Books
16. Edelman, G.M. and Tononi, G. (1999) *A Universe of Consciousness*, Basic Books
17. Ericsson, K A and Walter Kintsch. 1995. Long-term working memory. *Psychological Review* 102: 211–245.
18. Franklin, S. 2000. Deliberation and Voluntary Action in 'Conscious' Software Agents. *Neural Network World* 10:505-521.
19. Franklin, S, B J Baars, U Ramamurthy, and Matthew Ventura. 2005. The role of consciousness in memory. *Brains, Minds and Media* 1: 1–38, pdf.
20. Franklin, Stan, Sidney D'Mello, Bernard J Baars, and Uma Ramamurthy. 2009 to appear. Evolutionary pressures for perceptual stability and self as guides to machine consciousness. *International Journal of Machine Consciousness*.
21. Franklin, Stan and A C Graesser. 1997. Is it an agent, or just a program?: A taxonomy for autonomous agents. In *Intelligent agents iii*: 21–35. Berlin: Springer Verlag.
22. Freeman, W.J. (1991) The physiology of perception. *Sci. Am*. 264,78–85
23. Glenberg, A. M. 1997. What memory is for. *Behavioral and Brain Sciences* 20:1-19.
24. Goodale, M. A., and D. Milner. 2004. *Sight Unseen*. Oxford: Oxford University Press.
25. Hofstadter, D R and M Mitchell. 1995. The copycat project: A model of mental fluidity and analogy-making. In *Advances in connectionist and neural computation theory, vol. 2: Logical connections*, ed. K J Holyoak and J Barnden:205–267. Norwood N.J.: Ablex.
26. Jackson, J. V. 1987. Idea for a Mind. *Siggart* Newsletter, 181:23-26.
27. John, E.R. *et al.*(2001) Invariant reversible QEEG effects of anesthetics. *Conscious. Cogn.* 10,165–183 Kanerva, P. 1988. *Sparse distributed memory*. Cambridge MA: The MIT Press.
28. Kanwisher, N. (2001) Neural events and perceptual awareness. *Cognition:*79, 89–113
29. Llinas, R. and Ribary, U. (2001) Consciousness and the brain: the thalamocortical dialogue in health and disease. *Ann. N. Y. Acad. Sci.* 929,166–175
30. Maes, P. 1989. How to do the right thing. *Connection Science* 1: 291–323.
31. Marshall, J. 2002. Metacat: A self-watching cognitive architecture for analogy-making. In *24th Annual Conference of the Cognitive Science Society*:631-636.
32. Negatu, A., and S. Franklin. 2002. An action selection mechanism for 'conscious' software agents. *Cognitive Science Quarterly* 2:363-386.
33. Rao, Rajesh P.N and Olac Fuentes. 1998. Hierarchical learning of navigational behaviors in an autonomous robot using a predictive sparse distributed memory. *Machine Learning* 31: 87-113.
34. Rees, G. (2001). Seeing is not perceiving. *Nat.Neurosci*. 4, 678–680
35. Seth, A K, B J Baars, and D B Edelman. 2005. Criteria for consciousness in humans and other mammals. *Consciousness and Cognition* 14: 119–139.
36. Shanahan, M.P. & Baars, B.J. (2005). Applying global workspace theory to the frame problem, Cognition 98 (2), 157–176.
37. Skarda, C and Walter J Freeman. 1987. How brains make chaos in order to make sense of the world. *Behavioral and Brain Sciences* 10: 161–195.

38. Sloman, A. 1998. Damasio, Descartes, Alarms and Meta-management. In *Proceedings Symposiumon Cognitive Agents: Modeling Human Cognition*. San Diego: IEEE
39. Sloman, A. 1999. What Sort of Architecture is Required for a Human-like Agent? In *Foundations of Rational Agency*, ed. M. Wooldridge, and A. S. Rao. Dordrecht, Netherlands: Kluwer Academic Publishers.
40. Tyrrell, Toby. 1994. An evaluation of maes's bottom-up mechanism for behavior selection. *Adaptive Behavior* 2: 307-348.
41. Varela, F, J P. Lachaux, E Rodriguez, and J Martinerie. 2001. The brainweb: Phase synchronization and large-scale integration. *Nature Reviews Neuroscience* 2: 229–239.