International Journal of Machine Consciousness

SELF SYSTEM IN A MODEL OF COGNITION

UMA RAMAMURTHY

Department of Pediatrics and the Dan L. Duncan Institute for Clinical and Translational Research, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA
uramamur@bcm.edu

STAN FRNKLIN

Department of Computer Science and Institute for Intelligent Systems, The University of Memphis, Memphis, Tennessee, 38152, USA
franklin@memphis.edu

PULIN AGRAWAL

Department of Computer Science and Institute for Intelligent Systems, The University of Memphis, Memphis, Tennessee, 38152, USA
pagrawal@memphis.edu

Philosophers, psychologists and neuroscientists have proposed various forms of a "self" in humans and animals. All of these selves seem to have a basis in some form of consciousness. The Global Workspace Theory (GWT) [Baars, 1988, 2003] suggests a mostly unconscious, many layered self-system. In this paper, we consider several issues that arise from attempts to include a self-system in a software agent/cognitive robot. We explore these issues in the context of the LIDA model [Baars and Franklin, 2009], [Ramamurthy, *et al*., 2006] which implements the Global Workspace Theory..

Keywords: Consciousness; Self System; Global Workspace Theory; LIDA model.

1. Introduction

The LIDA model is both a conceptual and computational model implementing and fleshing out a major portion of Global Workspace Theory (GWT) [Baars, 1988]. The model also implements a number of other psychological and neuropsychological theories including situated cognition [Varela, 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 H-CogAff [1999], and transient episodic memory [Conway, 2001].

As is true with any computational/conceptual model of human cognition, the LIDA model has gaps, areas in which it cannot yet offer explanations. One such gap is the self-system.

Baars [1988] sees the self as an unconscious executive that receives conscious input and controls voluntary actions. There is a direct connection between self and consciousness If one damages the self-system of a human, then conscious contents may also disappear. Recall also, that in people with split brains, the dissociated executive loses access to the conscious contents of the other executive [Baars, 1988], [Baars, *et al*., 2003]. Our goal is to implement a self-system in the LIDA model that is in tune with GWT, while attempting to understand how the self system works in humans/animals.

2. Self System

In the spirit of GWT, a self-system in an autonomous agent may be constituted by three major components namely, the Proto-Self, the Minimal (Core) Self and the Extended Self as shown in Figure 1.

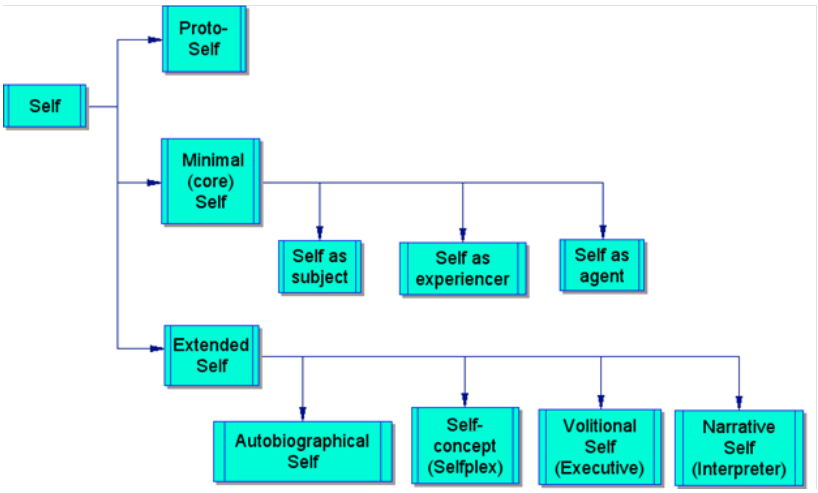

Figure 1. The Self System for LIDA

Neuroscientist Antonio Damasio conceived of a proto-self as a short-term collection of neural patterns of activity representing the current physical state of the organism [1999]. This proto-self receives neural and hormonal signals from visceral changes.

The minimal or core self is attributed to many animals by biologists, philosophers and neuroscientists [Bekoff and Sherman, 2004], [Damasio, 1999], [Gallagher, 2000], [Goodale and Milner, 2004]. The core consciousness is continually regenerated in a series of pulses (one in each of LIDA's cognitive cycles [Franklin and Ramamurthy, 2006]), which blend together to give rise to a continuous stream of consciousness. The minimal or core self is partitioned into the self-as-agent (the acting self), the self-as-experiencer

(the experiencing self) and the self-as-subject (the self that can be acted upon by other entities in the environment).

The extended self consists of the autobiographical self, the self-concept, the volitional or executive self, and the narrative self. This extended self is ascribed to humans and, possibly, to higher animals. The autobiographical self develops directly from episodic memory [Baddeley, *et al*., 2001], [Franklin, *et al*., 2005]. The self concept, also referred to as the self context [Baars, 1988] or the selfplex [Blackmore, 1999], consists of enduring self beliefs and intentions, particularly those relating to personal identity and properties. The volitional self provides executive function [Baars, 1988]. Finally, the narrative self is able to report, sometimes equivocally, contradictorily or self-deceptively, on actions, intentions, etc., [Gazzaniga, 1998].

3. LIDA Model

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], the Schema Mechanism [Drescher, 1991], [Chaput, *et al*., 2003], the Behavior Net [Maes, 1989], and the Subsumption Architecture [Brooks, 1991]. As the architecture implements GWT, the various modules in this system have processors executing and accomplishing small, simple and more complex tasks. These processors are often represented by codelets, which are small pieces of code that accomplish one specific task. The LIDA model has been detailed in several publications [Franklin, *et al*., 2007; Baars and Franklin, 2009; Ramamurthy, *et al*., 2006].

LIDA's processing can be viewed as consisting of a continual iteration of Cognitive Cycles [Franklin and Ramamurthy, 2006; Baars and Franklin, 2009; Madl *et al*., 2011]. Each cycle is composed of units of understanding, attending and acting. During each cognitive cycle a LIDA-based agent first makes sense of (understands) 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 choose an appropriate action, which it then executes. Thus, the LIDA cognitive cycle can be subdivided into three phases, the understanding phase, the consciousness (attending) phase, and the action selection phase.

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 into representations of more abstract entities such as objects, categories, actions, events, etc. The resulting percept is sent to the (preconscious)

Workspace where it cues both Transient Episodic Memory and Declarative Memory producing local associations. These local associations are combined with the percept to generate or update 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 coalition whose contents become the content of consciousness that is broadcast globally.

In the action selection phase of LIDA's cognitive cycle, relevant 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 provide the action selected for this cognitive cycle. The selected instantiated scheme triggers Sensory-Motor Memory to produce a suitable motor plan for the execution of the action. Its execution completes the cognitive cycle.

4. Implementing a Self System in LIDA

In the context of the LIDA model briefly described in the previous section, let us consider how the various parts of a Self-System as seen in Figure 1 can be implemented in this model.

4.1. *Implementing Proto-Self*

The Proto-Self for a software agent or cognitive robot can be viewed as the set of global and relevant parameters in the various modules of the autonomous agent. In LIDA, these are the parameters in the Behavior Net, the memory systems, and the underlying computer system's memory and operating system. These aspects, which constitute the Proto-Self, are already present in the LIDA model.

4.2. *Implementing Minimal/Core Self*

All the three parts of Minimal Self can be implemented as sets of entities in the LIDA ontology [Franklin and Ferkin, 2006], that is, computationally as collections of nodes and links in LIDA's Perceptual Associative Memory (PAM).

One of the features of consciousness is subjectivity, the first person point of view. The self-as-agent accomplishes some aspects of such subjectivity.[1] Self-as-agent was earlier thought to be implemented as the set of self-action nodes in PAM [Ramamurthy and

[1] Subjective consciousness is often used synonymously with phenomenal consciousness. We are not doing so here. We make no claim for phenomenal consciousness in LIDA based agents.

Franklin, 2011], i.e., nodes representing actions by the agent such as lie-down, stand, roll-over, walk, glance-left, etc. Having such action nodes in PAM would allow actions –

- to be part of structure building in LIDA's Workspace;
- to be included in cues to episodic memories;
- to come to consciousness;
- to be written to episodic memory as parts of events, and
- to be available for the creation of new schemes by the procedural learning mechanism.

This kind of implementation would give such actions first-class status among the ontological entities of the LIDA model. Self-as-agent would then be realized as the set of all self-action nodes in PAM.

Since the structure of representation of events in LIDA has changed [McCall, *et al.*, 2010], there is a need to modify the previous conception of minimal self. LIDA now implements events in the form of an event node to which its various attributes are connected via thematic role links. Attributes may include agent, subject, action etc. as shown in Figure 2.

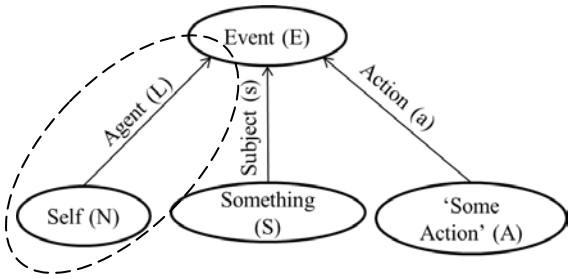

Figure 2. Event Representation and Self-as-Agent

Self-as-agent can now be implemented as follows. Consider a set consisting of a node, say N, and a thematic role link, say L, which links to an event node, say E (see Figure 2). The link L specifies agency in that event. The node N is the 'self' node. The self-as-agent will now be this 'Self' node together with the 'Agent' links connected to all event nodes. This modification will still keep the five characteristics of action-nodes, described above, intact.

Expectation codelets are a specific type of attention codelets that are produced with every action selected in LIDA. The expectation codelet attempts to bring to consciousness items in the Workspace that bear on the success of the given action achieving its expected result. Thus LIDA's expectation codelets will be part of the self-as-agent implementation.

Self-as-subject was earlier thought to be implemented as the set of acted-upon nodes in PAM [Ramamurthy and Franklin, 2011], i.e., nodes representing actions by other entities upon the agent such as being pushed, stroked, hugged, slapped, yelled-at, fallen-upon, etc.

Using the new representation of events, self-as-subject can be implemented as follows. Consider the 'self' node, say N and a thematic role link, say L, out of it, which links to an event node, say E (see Figure 3). The link L linking to the event node specifies the subject in that event. Now the self-as-subject is the 'Self' node along with all the 'Subject' links connected to all event nodes.

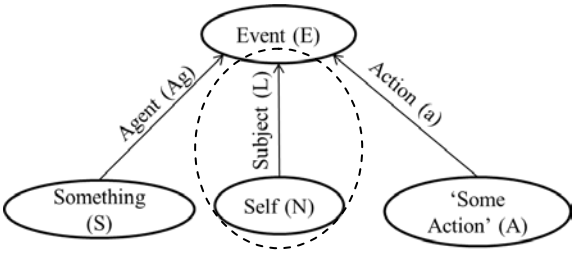

Figure 3. Self-as-subject

Self-as-experiencer might be thought of as being comprised of all of the rest of PAM, that is, of everything that can be recognized. Thus, the Minimal Self can be implemented simply from the existing modules in the LIDA model.

4.3. *Implementing the Extended Self*

Here we consider the four parts of the Extended Self from Figure 1. The Autobiographical Self is the collection of episodic memories of events that one has about himself or herself, rather than only about others. These memories have to have come from consciousness. This is a requirement because the agent must have been conscious of these events for them to have become part of Autobiographical Self. In LIDA, the local associations from transient episodic memory and declarative memory come to the Workspace during every cognitive cycle. From there they may become part of a conscious broadcast A memory that has come from consciousness requires a verifiable report (of that memory coming to consciousness). Not all of them may be operationally verifiable.

The Selfplex consists of personal beliefs and intentions. In the LIDA model, the agent's beliefs are in semantic memory. Intentions are represented in LIDA by event structures

in the Workspace that are tagged as intentions. Such intention (goal) structures are produced by volitional decision making. We also speak of the expected result of each selected behavior as being the LIDA agent's current goal or intention. Each such behavior selection also produces an intention codelet, a variety of attention codelet that looks for any opportunity to bring information concerning the goal to the Global Workspace.

Action that is taken volitionally, that is, as the result of conscious deliberation, is an instance of action by the Volitional Self. Deliberate actions, requiring multiple cognitive cycles, occur in LIDA and are represented as behavior streams. Deliberative acts have to be conscious, in the sense that the process of deliberation has to be conscious before the act itself. Thus LIDA has a volitional self.

Actions that are affected by the Narrative Self convey something meaningful about the agent. These actions are characterized by presence of personal pronouns in self-reports generated by the agent. These pronouns may appear explicitly or may be implied. First, a LIDA-based agent has to understand such self-report requests. This can occur in LIDA's Workspace. Then the agent has to generate the reports based on its understanding of such requests. The LIDA model can in principle accomplish this with existing modules. A LIDA-based agent must have motivations to report on itself and to enjoy responding to such queries about itself, implemented with feeling nodes in PAM. The agent has to become conscious of such a request, by means of attention codelets specifically built for such a task. We need reporting behavior streams in Procedural Memory that can generate reports from the contents of consciousness.

Effectively, the LIDA model provides for the basic blocks with which to implement the various parts of a multi-layered self system as hypothesized in GWT. There are several interesting issues that such an implementation would bring up at which we will look in the following discussion section.

5. Discussion

The main goal of our research work is to understand how minds work. Implementing a self system in the LIDA model provides a better and more complete understanding of cognition and of Global Workspace Theory.

We see that the Proto-Self is already part of the LIDA model; it is not built as a separate module/structure. This may be the case with most cognitive software agents/cognitive robots. The very nature of these systems requires global parameters for the functioning of these agents, thus affecting the state of the software agent or robot.

In contrast, the Minimal/Core Self and the Extended Self need to be implemented in the LIDA model. While the Minimal Self can be easily accommodated in the LIDA model

with the existing modules, the Extended Self requires new structures to be added. Implementing the various pieces of the self system would take us one step closer to a comprehensive model of cognition.

An autonomous agent / cognitive robot based on the LIDA model that also has a self system might be suspected of becoming close to being pbenomenally conscious for several reasons. First, such an agent/robot would be functionally conscious [Franklin, 2003]. Further, it could be made to fulfil the coherent, stable perceptual world condition [Merker, 2005; Franklin, 2005]. We claim that such an agent/robot will take us one step closer to realizing phenomenal consciousness in these cognitive models.

Today researchers at the Brain Mind Institute at EPFL are using virtual reality and brain imaging to understand how the human body is represented in the brain and how this affects the conscious mind [2011]. The self system is directly linked to consciousness and, as we implement models of machine consciousness, it is imperative that we include the self system in these models.

References

Baars, Bernard J. [1988] *A cognitive theory of consciousness*. Cambridge: Cambridge University Press.

Baars, B.J. [1997] *In the Theater of Consciousness: The Workspace of the Mind*. NY: Oxford University Press.

Baars, B J. [2003] How brain reveals mind: Neural studies support the fundamental role of conscious experience. Journal of Consciousness Studies 10: 100–114.

Baars, Bernard J and Stan Franklin [2003] How conscious experience and working memory interact. Trends in Cognitive Science 7: 166–172.

Baars, B J, T Ramsoy, and S Laureys [2003] Brain, conscious experience and the observing self. Trends Neurosci. 26: 671–675.

Baars, Bernard J and Stan Franklin [2009] Consciousness is Computational: The LIDA model of global workspace theory. International Journal of Machine Consciousness, Vol 1, Issue 1, pp. 23-32.

Baddeley AD, Hitch GJ [1974] Working memory. In Bower GA (Ed), *The Psychology of Learning and Motivation*. New York: Academic Press, pp 47–89.

Baddeley, Alan, Martin Conway, and John Aggleto [2001] *Episodic memory*. Oxford: Oxford University Press.

Barsalou, L. W. [1999] Perceptual symbol systems. Behavioral and Brain Sciences, 22, 577–660.

Bekoff , M., and P. W. Sherman [2004] Reflections on animal selves. Trends in Ecology and Evolution 19: 176-180.

Blackmore, Susan [1999] *The meme machine*. Oxford: Oxford University Press.

Brooks RA. [1991] How to build complete creatures rather than isolated cognitive simulators. In VanLehn K (ed), *Architectures for Intelligence*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp 225–239.

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.

Conway, M. A. [2001] Sensory-perceptual episodic memory and its context: Autobiographical memory. In A. Baddeley, M. Conway, & J. Aggleton (Eds.), *Episodic memory*. Oxford: Oxford University Press.

Damasio, Antonio R [1999] *The feeling of what happens.* New York: Harcourt Brace.

Drescher, Gary L. [1991] *Made-up minds: A constructivist approach to artificial intelligence*. Cambridge, MA: MIT Press.

Ericsson KA, Kintsch W. [1995] Long-term working memory. Psychological Review 102: 211–245.

Franklin, S. [2003] IDA: A Conscious Artifact? Journal of Consciousness Studies, 10, 47–66.

Franklin, S. [2005] Evolutionary Pressures and a Stable World for Animals and Robots: A Commentary on Merker. Consciousness and Cognition, 14, 115–118.

Franklin, S, B J Baars, U Ramamurthy, and Matthew Ventura [2005] The role of consciousness in memory. Brains, Minds and Media 1: 1–38.

Franklin, S. and Ramamurthy U [2006] Motivations, Values and Emotions: 3 sides of the same coin. Proceedings of the Sixth International Workshop on Epigenetic Robotics, Paris, France, September 2006, Lund University Cognitive Studies, 128; p. 41-48.

Franklin, S., & Ferkin, Michael H. [2006]. An Ontology for Comparative Cognition: a Functional Approach. Comparative Cognition & Behavior Reviews, 1, 36–52.

Stan Franklin, Uma Ramamurthy, Sidney K. D'Mello, Lee McCauley, Aregahegn Negatu, Rodrigo Silva L., and Vivek Datla [2007] LIDA: A Computational Model of Global Workspace Theory and Developmental Learning, AAAI 2007 Fall Symposium - AI and Consciousness: Theoretical Foundations and Current Approaches.

Gallagher, Shaun [2000] Philosophical conceptions of the self: Implications for cognitive science. Trends in Cognitive Science 4: 14–21.

Gazzaniga, Michael S. [1998] *The mind's past.* Berkeley: University of California Press.

Glenberg AM. [1997] What memory is for. Behavioral and Brain Sciences 20:1–19.

Goodale, M. A., and D. Milner [2004] *Sight Unseen*. Oxford: Oxford University Press.

Hofstadter DR, Mitchell M. [1995] The Copycat Project: A model of mental fluidity and analogy-making. In Holyoak K.J., and Barnden J. (eds), Advances in connectionist and neural computation theory, Vol. 2: logical connections. Norwood N.J.: Ablex, pp 205–267.

Kanerva P [1988] *Sparse Distributed Memory*. Cambridge MA: The MIT Press.

Madl, T., Baars, B. J., & Franklin, S. [2011]. The Timing of the Cognitive Cycle. PLoS ONE, 6(4), e14803. doi: 10.1371/journal.pone.0014803

Maes, P. [1989] How to do the right thing. Connection Science 1: 291–323.

Marshall, J. [2002] Metacat: A self-watching cognitive architecture for analogy-making. In 24th Annual Conference of the Cognitive Science Society:631-636.

McCall, R., Franklin, S., & Friedlander, D. [2010]. Grounded Event-Based and Modal Representations for Objects, Relations, Beliefs, Etc. Paper presented at the FLAIRS-23, Daytona Beach, FL.

Merker, Bjorn. [2005] The liabilities of mobility: A selection pressure for the transition to consciousness in animal evolution. Consciousness and Cognition 14: 89–114.

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.

Shanahan, M P. [2006] A cognitive architecture that combines internal simulation with a global workspace. Consciousness and Cognition 15: 433-449.

Sloman A. [1999] What Sort of Architecture is Required for a Human-like Agent? In Wooldridge M, Rao AS (eds), Foundations of Rational Agency. Dordrecht, Netherlands: Kluwer Academic Publishers, pp 35–52.

Strawson, G. [1999] The self and the sesmet. In *Models of the Self*, ed. Shaun Gallagher and J Shear:483–518. Charlottesville, VA: Imprint Academic.

Uma Ramamurthy, Bernard J Baars, Sidney K. D'Mello and Stan Franklin [2006] LIDA: A Working Model of Cognition. Proceedings of the 7th International Conference on Cognitive Modeling, Eds: Danilo Fum, Fabio Del Missier and Andrea Stocco, p 244-249.

Ramamurthy, U., Franklin, S. [2011]. Self System in a model of Cognition. Proceedings of Machine Consciousness Symposium at the Artificial Intelligence and Simulation of Behavior Convention (AISB'11), University of York, UK, p 51-54.

The Science of Self – Philosophy and Neurobiology [2010]: http://actualites.epfl.ch/newspaper-article?np_id=1648&np_eid=114

EPFL: The real avatar: [2011] Researchers use virtual reality and brain imaging to hunt for the science of the self: http://www.physorg.com/news/2011-02-real-avatar-virtual-reality-brain.html

Varela, F. J, Thompson, E., & Rosch, Eleanor [1991] *The embodied mind*. Cambridge, MA: MIT Press.