Self System in a Model of Cognition

Uma Ramamurthy* and Stan Franklin**

* St Jude Children's Research Hospital, Memphis, TN 38105, USA.
Email: **uma.ramamurthy@stjude.org**.
**Dept. of Computer Science and Institute for Intelligent Systems,
The University of Memphis, Memphis, TN 38152, USA.
Email: **franklin@memphis.edu**.

Abstract. 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) [1 - 3] 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 [4], [15] which implements the Global Workspace Theory.

1 INTRODUCTION

The LIDA model is both a conceptual and computational model implementing and fleshing out a major portion of Global Workspace Theory (GWT) [1]. The model also implements a number of other psychological and neuropsychological theories including situated cognition [20], perceptual symbol systems [21], working memory [23], memory by affordances [24], long-term working memory [25], Sloman's H-CogAff [26], and transient episodic memory [22].

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 [1] 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 that in people with split brains, the dissociated executive loses access to the conscious contents of the other executive [1], [6]. 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.

Neuroscientist Antonio Damasio conceived a *proto-self* as a short-term collection of neural patterns of activity representing the current state of the organism [9]. This proto-self receives neural and hormonal signals from visceral changes.

The *minimal* or *core self* is attributed to all animals by biologists, philosophers and neuroscientists [9], [12], [19]. The core consciousness is continually regenerated in a series of pulses (LIDA's cognitive cycles [11]), 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 [7], [10]. The self concept, also referred to as the self context [1] or the selfplex [8] consists of enduring self beliefs and intentions, particularly those relating with personal identity and properties. The volitional self provides executive function [1]. Finally, the narrative self is able to report, sometimes equivocally, contradictorily or self-deceptively, on actions, intentions, etc., [13].

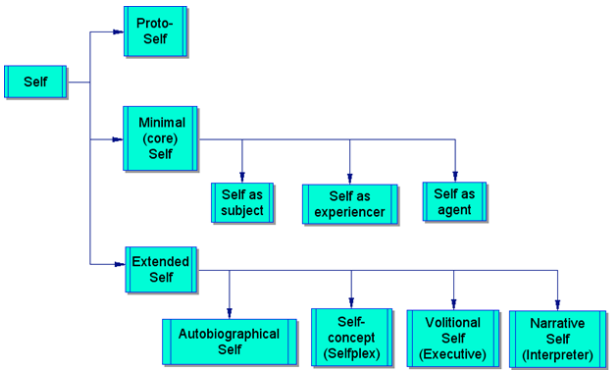

Figure 1. The Self System for LIDA

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 [27], [30], Sparse Distributed Memory [28], the Schema Mechanism [31], [33], the Behavior Net [29], and the Subsumption Architecture [32]. As the architecture implements GWT, the various modules in this system have processors executing and accomplishing small, simple and 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 [34], [35], [36].

LIDA's processing can be viewed as consisting of a continual iteration of Cognitive Cycles [11], [35]. Each cycle constitutes units of understanding, attending and acting. During each cognitive cycle a LIDA-based 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 choose an appropriate action which it then executes. Thus, the LIDA cognitive cycle can be subdivided into three phases, *the understanding phase, the consciousness 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 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 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 algorithm for the execution of the action. Its execution completes the cognitive cycle.

4 IMPLEMENTING 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 in Figure 1 can be implemented in this model.

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.

Implementing Minimal/Core Self: All the three parts of Minimal Self can be implemented as sets of entities in the LIDA ontology, that is, computationally as collections of nodes in the slipnet of LIDA's perceptual associative memory.

One of the features of consciousness is subjectivity, the first person point of view. The self-as-agent accomplishes some aspects of such subjectivity. Self-as-agent can be implemented as the set of self-action nodes in the slipnet, i.e., nodes representing actions by the agent such as lie-down, stand, roll-over, walk, glance-left, etc. Having such action nodes in the slipnet would allow actions –

- to be part of structure building in working memory;
- 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 the slipnet.

Expectations 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 can be implemented as the set of acted-upon nodes in the slipnet, i.e., nodes representing actions by other entities upon the agent such as being pushed, stroked, hugged, slapped, yelled-at, fallen-upon, etc.

Self-as-experiencer might be thought of as being comprised of all of the rest of the slipnet. The Minimal Self can be implemented simply from the existing modules in the LIDA model.

Implementing 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. In LIDA, the local associations from transient episodic memory and declarative memory come to the workspace in every cognitive cycle. This requires a verifiable report (of that memory coming to consciousness). Not all of them may be operationally verifiable.

The Selfplex is personal beliefs and intentions. In the LIDA model, the agent's beliefs are in the semantic memory. Intentions are represented by the intentions codelets. These are processes that get generated at each volitional goal selection. They look for opportunity to bring information concerning the goal to the Global Workspace. In LIDA, each volitional goal has an intention codelet.

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

An action to be influenced by the Narrative Self must intend to convey something meaningful about the speaker; it can be

determined by the presence of either explicit or implied personal pronouns. First, a LIDA-based agent has to understand such self-report requests. This can be implemented in the perceptual associative memory using perception codelets, slipnet and working memory. Then the agent has to generate the reports based on its understanding of such requests. The LIDA model facilitates this with existing modules. A LIDA-based agent can have motivations to report on itself and enjoy responding to such queries about itself, with feeling nodes in its perceptual associative memory. The agent has to become conscious of such a request, by its attention codelets, specifically built for such a task. We need reporting behavior streams in the procedural memory that can generate reports from the contents of consciousness.

Effectively, the LIDA model provides for the basic blocks 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, which we will look at in the discussion section of this paper.

5 DISCUSSION

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

We see that the Proto-Self is already part of the LIDA model and 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 the 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 facilitated in the LIDA model with the existing modules, the Extended Self requires new structures to be added to the existing modules. Implementing the various pieces of the self system would take us one step closer to a complete model of cognition.

An autonomous agent/cognitive robot based on the LIDA model that also has a self system might be suspected of being close to subjectively conscious for several reasons. First, such an agent/robot would be functionally conscious. Further, it could be made to fulfil the coherent, stable perceptual world condition [14]. 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 [37]. 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

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

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

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

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

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

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

[7] Baddeley, Alan, Martin Conway, and John Aggleton (2001), *Episodic memory*. Oxford: Oxford University Press.

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

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

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

[11] Franklin, S. and Ramamurthy U. (2006), Motivations, Values and Emotions: 3 sides of the same coin. *Proc. 6th International Workshop on Epigenetic Robotics*, Paris, September 2006, Lund University Cognitive Studies, 128: 41-48.

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

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

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

[15] U Ramamurthy, B J Baars, S K D'Mello and S Franklin (2006), LIDA: A Working Model of Cognition. *Proc. 7th International Conference on Cognitive Modeling*, 244-249.

[16] 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, (2005).

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

[18] 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.

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

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

[21] Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577–609.

[22] 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.
[23] Baddeley AD, Hitch GJ. (1974), Working memory. In Bower GA (Ed), The Psychology of Learning and Motivation. New York: Academic Press, pp 47–89.
[24] Glenberg A. M. (1997) What memory is for. *Behavioral and Brain Sciences* 20:1–19.
[25] Ericsson KA, Kintsch W. (1995). Long-term working memory. *Psychological Review* 102: 211–245.
[26] 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: Kluwer Academic Publishers, pp 35–52.
[27] Hofstadter DR, Mitchell M. (1995), The Copycat Project: A model of mental fluidity and analogy-making. In Holyoak KJ, Barnden JA (eds), *Advances in connectionist and neural computation theory, Vol. 2: logical connections*. Norwood N.J.: Ablex, pp 205–267.
[28] Kanerva P (1988) *Sparse Distributed Memory*. Cambridge MA: The MIT Press.
[29] Maes, P. (1989), How to do the right thing. Connection Science 1: 291–323.
[30] Marshall, J. (2002), Metacat: A self-watching cognitive architecture for analogy-making. In *24th Annual Conference of the Cognitive Science Society*:631-636.
[31] Drescher, Gary L. (1991). *Made-up minds: A constructivist approach to artificial intelligence.* Cambridge, MA: MIT Press, (1991).
[32] 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.
[33] 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.
[34] 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*, Menlo Pk, CA: AAAI.
[35] Baars, Bernard J and Stan Franklin, (2009). Consciousness is computational: The LIDA model of global workspace theory. *International Journal of Machine Consciousness*, (2009).
[36] Uma Ramamurthy, Bernard J. Baars, Sidney K. D'Mello, and Stan Franklin (2006). LIDA: A Working Model of Cognition. *7th International Conference on Cognitive Modeling*, Eds: Danilo Fum, Fabio Del Missier and Andrea Stocco, p. 244-249.
[37] The Science of Self – Philosophy and Neurobiology: http://actualites.epfl.ch/newspaper-article?np_id=1648&np_eid=114
[38] The real avatar: 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