Memory Systems for Cognitive Agents

Uma Ramamurthy[1] and Stan Franklin[2]

Abstract. The several different memory systems in human beings play crucial roles in facilitating human cognition. To build artificial agents that have cognitive capabilities similar to those of human beings, we have to develop these agents based on architectures modelling what we know of human cognition from neuroscience, psychology and cognitive science. In this paper we describe the various memory systems in the LIDA Architecture, which implements Global Workspace Theory. We discuss the interaction between these memory systems, feelings and emotions, and consciousness in the context of cognitive cycles. Finally, we look at our current work on spatial memory in the LIDA model.

1 INTRODUCTION

Human memory seems to come in myriad forms: sensory, procedural, working, declarative, episodic, semantic, long-term memory, long-term working memory and perhaps others. To achieve human-like cognitive capabilities in artificial agents, we have to build them with principles of human cognition and learning. When an autonomous artificial agent [19] is equipped with computational versions of human cognitive features, such as multiple senses, perception, various forms of memory including transient episodic memory and declarative memory, learning, emotions, multiple drives, it is called a cognitive agent [18]. Such cognitive agents promise to be more flexible, more adaptive, more human-like than classical software systems because of their ability to learn, and to deal with novel input and unexpected situations. One way to design and implement cognitive agents is to build them within the constraints of the Global Workspace Theory (GWT) [2], [3], a psychological theory that gives a high-level, abstract account of human consciousness and cognition.

Per Global Workspace Theory, one of the most fundamental functions of consciousness is to provide access among separate sources of information. Effectively, consciousness creates access to various memory systems of a cognitive agent. In the following sections, we will discuss the various human memory systems that play a role in the Learning Intelligent Distribution Agent (LIDA), a model of cognition that implements Global Workspace Theory [7], [37]. The main aims of the LIDA model include understanding how the mind works as well as building smarter and better artificial cognitive systems. The LIDA model, which is both computational and conceptual, includes modules for perception, various types of memories, "consciousness", action selection, deliberation, volition, and several types of learning technologies [23].

[1] St Jude Children's Research Hospital, Memphis, TN 38105, USA. Email: uma.ramamurthy@stjude.org

[2] Dept. of Computer Science and Institute for Intelligent Systems, The University of Memphis, Memphis, TN 38152, USA. Email: franklin@memphis.edu

2 MEMORY SYSTEMS

The memory modules in LIDA are not unique to this model. Other cognitive architectures like SOAR, ACT-R and Clarion for example, have multiple memory systems in them. In LIDA, the approach to memory is more systemic and granular. Let us consider the different memory systems of the LIDA model, short-term to long-term.

Sensory memory holds incoming sensory data in sensory buffers and performs the initial processing. It provides a workspace for integrating the features from which representations of objects and their relations are constructed. There are different sensory memory registers for different senses and probably a separate sensory memory for integrating multimodal information. Sensory memory decays at the fastest rate, measured in tens of milliseconds.

Working memory is the scratchpad of the mind. It holds sensory data, including visual images and inner speech, together with their interpretations. There are separate working memory components associated with the different senses, the visuo-spatial sketchpad and the phonological loop [8], [5]. Its decay rate is in tens of seconds.

Episodic or autobiographical memory is memory for events having features of a particular time and place [10]. This memory system is associative and content-addressable.

An unusual aspect of the LIDA model is its transient episodic memory (TEM), an episodic memory with a decay rate measured in hours. Our hypothesis is that a conscious event is stored in transient episodic memory by a broadcast from a global workspace. A corollary to this

hypothesis says that conscious contents can only be encoded in long-term declarative memory via consolidation from transient episodic memory.

Humans have a variety of long-term memory types that may decay exceedingly slowly. Memory research distinguishes between procedural memory, the memory for motor skills including verbal skills, and declarative memory. Declarative memory (DM) is composed of autobiographical memory and semantic memory, memories of fact or belief typically lacking a particular source with a time and place of acquisition. Declarative memory systems are accessed by means of cues from working memory.

We see a clear distinction between perceptual memory (recognition memory [34]) and sensory memory (similar to Taylor [42]). Our model distinguishes between semantic memory and perceptual associative memory (PAM) and hypothesizes distinct mechanisms for each [20]. PAM memory is a memory for individuals, categories, actions, feelings, events, and their relations. PAM plays the major role in recognition, categorization, and more generally the assignment of interpretations. Upon presentation of features of an incoming stimulus, PAM returns precepts, the beginnings of meaning. We venture that PAM is evolutionarily older than TEM or declarative memory. This further points to the likelihood, though not at all certain, that they have different neural mechanisms. Since the contents of TEM consolidate into DM, which contains semantic memory, these facts suggest the possibility of separate mechanisms for PAM and semantic memory.

Procedural memory in LIDA is a modified and simplified form of Drescher's schema mechanism [14], the scheme net. The scheme net is a directed graph whose nodes are (action) schemes and whose links represent the 'derived from' relation. A scheme consists of an action, together with its context and its result. At the periphery of the scheme net lie empty schemes (schemes with a primitive action, but no context or results), while more complex schemes consisting of actions and action sequences are discovered as one moves inwards.

3 TEM AND DM IN LIDA

Transient episodic and declarative memories have distributed representations in the LIDA model. There is evidence that this is also the case in the nervous system [20]. In this model, these two memories are implemented computationally using a modified version of Kanerva's Sparse Distributed Memory (SDM) architecture [26], [36]. The SDM architecture has several similarities to human memory [26] and provides for "reconstructed memory" in its retrieval process:

- Fast divergence in SDM is equivalent to knowing that one does not know.
- Neither converging nor diverging indicates the tip-of-the-tongue state.
- Rehearsal happens by writing a datum many times to memory. A datum rehearsed well is retrieved with fewer iterations than an item that is stored only once.
- Full and overloaded memories exhibit momentary feelings of familiarity that fade away rapidly.
- Forgetting increases with time because of intervening write operations (interference), as well as decay.

A preconscious percept consisting of a selection of the contents of sensory memory, together with recognitions, categorizations and other interpretations produced in PAM, are stored in working memory. Only the conscious portion of the contents of working memory (actually long-term working memory [16]) is stored in TEM. Information from the same conscious content is used to update PAM, TEM, and procedural memory. The undecayed contents of TEM are consolidated into DM at a later time offline. Retrieval from the content-addressable, associative TEM and DM memories uses recently stored unconscious contents of working memory as cues.

In the next section, we will describe LIDA's cognitive cycle and the role played by the various memory systems in effecting human-like cognitive processing in this artificial agent.

4 LIDA'S COGNITIVE CYCLE

LIDA's processing can be viewed as consisting of a continual iteration of Cognitive Cycles. Each cycle consists of units of understanding, attending and acting. 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 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 PAM 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 understands 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, possibly 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.

The LIDA model hypothesizes that all human cognitive processing is via a continuing iteration of such cognitive cycles. The unconscious elements of these cycles are proposed to occur asynchronously, with each cognitive cycle taking roughly 200-300 milliseconds. These cycles cascade, that is, several cycles may have different processes running simultaneously in parallel. This cascading must, however respect the serial nature of conscious processes that are necessary to maintain the stable, coherent image of the world [21], [32]. The cascading cycles, which partially overlap, allows a rate of cycling in humans of five to ten cycles per second. There is considerable evidence from cognitive psychology and neuroscience that is consistent with such cognitive cycling in humans [28], [41], [46], [48].

5 FORGETTING IN MEMORY SYSTEMS

Forgetting is a fundamental aspect of memory. Historically, decay [15], [12], [35] and interference [30], [27], [47] have been proposed as two theories on forgetting. Retrieval failures have also been proposed as the possible basis for forgetting – memories never disappear; they just cannot be retrieved [43]. We do not take this view, and build decay into every memory system.

Altmann and Gray [1] have proposed a functional theory of decay, which says that decay and interference are functionally related. If a memory trace decays, it interferes less with future memory traces. This theory states that when an attribute is to be updated frequently in memory, its current value decays to prevent interference with later values; and the decay rate adapts to the rate of memory writes. Wixted [49] has proposed that recently formed memories which have not yet consolidated are vulnerable to interference from mental activity and memory formation.

Memory researchers hypothesize about decay in working memory [25]. While there is debate and controversy over decay in declarative/autobiographical memory, decay in transient episodic memory is a hypothesis that the LIDA model offers.

Decay plays two roles in these cognitive agents: modelling the cognitive processes in memory (assuming the hypothesis that there is decay in human memory systems) and providing the solution to the memory capacity problem of the SDM architecture. Decay is essential in the modified SDM architecture utilized in the LIDA model for Transient Episodic Memory (TEM). Decay ensures that the detailed memory traces of episodes that have occurred in the past few hours are retrievable. Without decay, the SDM architecture will retrieve a high-level, aggregate of all the traces written to that region of the binary space, and not the specific trace that is expected from a TEM. To be able to retrieve details of episodes with cues such as 'where did we park our car this morning?' or 'what did we have for dinner yesterday night?' we hypothesize that decay is required in the modified SDM that will be used as transient episodic memory.

We have tested different types of decay mechanisms in our modified SDM module, including linear decay, exponential decay and inverse sigmoid decay [38]. The inverse sigmoid decay function models the memory hypotheses of decay mechanism by rapid decay of the less rehearsed episodes while episodes which were rehearsed most experienced a very slow decay. Those episodes rehearsed most were retrievable after several decay cycles while all other episodes written fewer times decayed away in the first couple of decay cycles. This high grade filtering ensures that only relevant, important, unique, urgent and highly emotion-based episodes are retained in transient episodic memory, as they come to consciousness many times and are thus written many times to TEM.

6 MEMORY CONSOLIDATION

The Memory Consolidation hypothesis has been discussed and debated from the time it was proposed over a hundred years ago by Müller and Pilzecker [33]. In this hypothesis, it is believed that the hippocampal complex acts as a temporary indexer linking traces in other cortical regions. With repeated reference and retrieval of the memory traces, direct cortico-cortical connections get established and these connections are independent of the hippocampal function [45]. The exact processes and purpose of this mechanism are still unclear. Many believe that consolidation occurs over hours and days, and during

our REM sleep. There is also debate about this process being conscious vs. subconscious. The LIDA model conjectures the need for two episodic memories, transient episodic memory and long term declarative memory. As pointed out in the previous section, the first is needed to recall details of events that would, over time, be wiped out by interference from similar events. In the LIDA model, events reach DM only by consolidation from TEM.

We use the LIDA model to propose a design for memory consolidation. We hypothesize that in cognitive agents based on the LIDA model, the memory traces which have not decayed away from transient episodic memory (TEM) are consolidated into the agent's declarative memory (DM). The contents of every conscious broadcast get stored in TEM. Over time and without rehearsing that information, those memory traces in TEM will decay. On the other hand, when those traces are rehearsed and hence strengthened, they will remain in the TEM. We hypothesize that at regular intervals (perhaps equivalent to human sleep cycles), the cognitive agent transfers the contents of its TEM to its DM.

The two memories – TEM and DM – based on the modified SDM architecture have identical address spaces. The TEM employs a faster inverse sigmoid decay function tuned to the domain in which the cognitive agent lives. The DM has a variable decay rate based on the inverse sigmoid decay function but with parameters different from those of TEM. The decay mechanism in TEM is crucial in ensuring that only memory traces that are significant, relevant and important to the cognitive agent are consolidated to DM. A ball seen under a bush on a morning walk will be encoded in TEM, but is unlikely to be consolidated into DM unless some particular meaning gives it an affective boost, or brought it to consciousness multiple times leading to multiple encodings.

At specific intervals, defined by the parameter 'consolidation time', the consolidation mechanism goes into action. Since the two memories have identical address space, there will be a one-to-one correspondence between their hard locations. The consolidation mechanism transfers the contents of the bit-counters of each hard location in the modified SDM used in the TEM to the corresponding hard location in DM. The parameter 'consolidation time' may be tuned dependent on the domain in which the cognitive agent lives. We hypothesize that this will be in the order of a few hours. The consolidation mechanism may also be triggered by other internal or external states.

7 DISCUSSION

The main goal of our research work in the LIDA model is to understand how minds work, be they human, animal or artificial. In that spirit, the LIDA model has a very granular architecture accounting for various cognitive processes. The cognitive cycle of the LIDA model provides an important tool for fine-grained analyses of cognitive processes. We have several memory systems in the model as described in this paper, based on both psychological, neuroscience and evolutionary evidence as well as on the interactions these memories have with consciousness per Global Workspace Theory [20].

As must be true with any computational/conceptual model of human cognition, the LIDA model is replete with gaps, areas in which it cannot yet offer explanations. One such gap with reference to human memory systems and artificial agents that we are currently working on is *spatial memory*.

In the human brain, two neural systems facilitate encoding of self-location [13]: they are (1) the *place cells* in the hippocampus for encoding unique environments and (2) *grid cells, border cells* and *head-direction cells* in the parahippocampal and entorhinal cortices for mapping positions and directions in all environments. Humans and many animals construct multiple spatial maps, also called cognitive maps [31] generated by these two neural systems. These spatial maps can be extended by adding multiple maps together.

Episodic memory is for the recording the 'what', 'where' and 'when' of events. The 'where' component of episodic memory results in cognitive maps. We hypothesize that a separate memory module/mechanism is needed in the LIDA model to account for such spatial memory/cognitive maps. While considering this memory module, we have to address several issues related to this memory:

- What is the interaction between the spatial memory and the other memory systems in the LIDA model?
- How does consciousness interact with spatial memory?
- What will be the basic representation of a spatial map, and how will it be accessed?
- If complex spatial maps are created from smaller fragments, how are the different fragments linked together and where are they stored?
- How do we represent very large environments in these spatial maps?
- Is there a decay mechanism in spatial memory and if so, what type of decay is to be employed in this memory?

As yet we have only tentative answers to a few of these questions. Taking advantage of this so far relatively rare occurrence of neuroscience providing a mechanism, a primitive spatial map will be represented in a picture like fashion inhabited by land-marks (objects). The representation will denote the size, shape and orientation of the object as well as its position and distance relative to

other landmarks. Each object will also be connected back to its corresponding node in LIDA's PAM, so as to make connections with features and relations of the object that are known to LIDA.

LIDA's spatial memory must interact with its PAM as well as with its two episodic memories so as to provide locations for events [29]. We envision much of this interaction taking place through LIDA's preconscious working memory, but just how is still an open question.

Spatial memory will be a long term memory system. Like most of those in the LIDA model, it will have a network structure with nodes corresponding to spatial maps and links to inclusion (being a subset of). Again as in other forms of long term memory in LIDA, spatial learning will have to be both selectionist (reinforcing existing spatial maps) and instructionalist (creating new spatial maps, or updating the content of existing maps).

Consciousness will play the same role with spatial memory as it does with all other memory systems. We learn that to which we attend, that is, the contents of consciousness.

As we continue work on understanding, designing and implementing spatial memory in the LIDA model, we hope that it will take us one step closer to realizing a more comprehensive and complete model of cognition. Using this model to build artificial agents will enhance our understanding of the interaction amongst these various memory systems, and between these memory systems and consciousness.

REFERENCES

[1] Altmann, E. M. & Gray, W. D. 2002. Forgetting to remember: The functional relationship of decay and interference. Psychological Science, 13(1), 27-33.

[2] Baars, B. J. 1988. A Cognitive Theory of Consciousness. Cambridge.: Cambridge University Press.

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

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

[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] 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.

[8] Baddeley, A. D. 1993. Working memory and conscious awareness. In Theories of memory, ed. A. Collins, S. Gathercole, M. Conway, and P. Morris. Howe: Erlbaum.

[9] Baddeley, A. D. 2000. The episodic buffer: a new component of working memory? Trends in Cognitive Science 4:417-423.

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

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

[12] Brown, J. 1958. Some tests of the decay theory of immediate memory. Quarterly Journal of Experimental Psychology,10:12-21.

[13] D Derdikman and E. I. Moser. 2010. A manifold of spatial maps in the brain. Trends in Cognitive Sciences, December 2010, Vol. 14, No. 12, p. 561-569.

[14] Drescher, G. 1991. Made Up Minds: A Constructivist Approach to Artificial Intelligence. Cambridge, MA: MIT Press.

[15] Ebbinghaus, H. 1885/1964. Memory: A contribution to experimental psychology. New York: Dover.

[16] Ericsson, K. A., and W. Kintsch. 1995. Long-term working memory. Psychological Review 102:211-245.

[17] Franklin, S. 1995. Artificial Minds. Cambridge MA: MIT Press.

[18] Franklin, S. 1997. Global Workspace Agents. Journal of Consciousness Studies 4:322-334.

[19] Franklin, S., and A. C. Graesser. 1997. Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. In Intelligent Agents III. Berlin: Springer Verlag. 21-35.

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

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

[22] 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.

[23] Franklin, Stan, 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.

[24] Freeman, W J. 2002. The limbic action-perception cycle controlling goal-directed animal behavior. Neural Networks 3: 2249-2254.

[25] James, Michael. 2002. Modelling Working Memory Decay in Soar. Online Proceedings of the 22nd North American Soar Workshop, Ann Arbor, MI (http://www.eecs.umich.edu/~soar/sitemaker/workshop/22/James-S22.PDF)

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

[27] Keppel, G. and Underwood, B. J. 1962. Proactive inhibition in short-term retention of single items. Journal of Verbal Learning and Verbal Behavior, 1:153-161.

[28] Massimini M, Ferrarelli F, Huber R, Esser Steve K, Singh H, et al., 2005. Breakdown of Cortical Effective Connectivity During Sleep. Science. 309: 2228–2232.

[29] 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.

[30] McGeoch, J.A. 1932. Forgetting and the law of disuse. Psychological Review, 39:352-370.

[31] McNaughton, B. L., Battaglia, Francesco P., Jensen, O., Moser, Edvard I., & Moser, M.-B. 2006. Path inte-gration and the neural basis of the 'cognitive map'. Nature Reviews Neuroscience, 7, 663–678.

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

[33] Müller G.E. and Pilzecker A. 1900. Z. Psychol. 1, 1.

[34] Nadel, L. 1992. Multiple memory systems: What and why. J. Cogn. Neurosci., 4, 179–188.

[35] Peterson, L. R. and Peterson, M. J. 1959. Short-term retention of individual verbal items. Journal of Exper-imental Psychology, 58:193-198.

[36] Ramamurthy, U., S. D'Mello, and S. Franklin. 2004. Modified Sparse Distributed Memory as Transient Episodic Memory for Cognitive Software Agents. IEEE International Conference on Systems, Man and Cy-bernetics (SMC2004).
[37] Ramamurthy,U., B J Baars, S K D'Mello and S Franklin. 2006. LIDA: A Working Model of Cognition. Proceedings of the 7th International Conference on Cognitive Modelling, p 244-249.
[38] Ramamurthy, Uma, Sidney K. D'Mello, and Stan Franklin. 2006. Realizing Forgetting in a Modified Sparse Distributed Memory System. Proceedings of the 28th Annual Conference of the Cognitive Science Society, p. 1992-1997.
[39] 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.
[40] Shanahan, M P. 2006. A cognitive architecture that combines internal simulation with a global workspace. Consciousness and Cognition 15: 433-449.
[41] Sigman M, Dehaene S, 2006. Dynamics of the Central Bottleneck: Dual-Task and Task Uncertainty. PLoS Biol. 4.
[42] Taylor, J. G. 1999. The Race for Consciousness. Cambridge, MA: MIT Press.
[43] Tulving, E. 1968. Theoretical issues in free recall. In T.R. Dixon & D.L. Horton (eds.) Verbal Behaviour and General Behaviour Theory, Prentice Hall, Englewood Cliffs, N.J.
[44] Tulving, E. 1985. Memory and consciousness. Canadian Psychology 26:1-12.
[45] Tulving, E. and Craik, I.M. Fergus. 2000. The Oxford Handbook of Memory. Editors. Oxford University Press.
[46] Uchida N, Kepecs A, Mainen Zachary F, 2006. Seeing at a glance, smelling in a whiff: rapid forms of perceptual decision making. Nature Reviews Neuroscience. 7: 485–491.
[47] Waugh, N. C. and Norman, D. A. 1965. Primary Memory. Psychological Review, 72:89-104.
[48] Willis J, Todorov A, 2006. First Impressions: Making Up Your Mind After a 100-Ms Exposure to a Face. Psychological Science. 17: 592–599.
[49] Wixted, John T. 2004. The Psychology and Neuroscience of Forgetting. Annual Rev. Psychol. 2004. 55:235-69.