A Foundational Architecture for Artificial General Intelligence

Stan FRANKLIN
Computer Science Department &
Institute for Intelligent Systems
The University of Memphis

Abstract. Implementing and fleshing out a number of psychological and neuroscience theories of cognition, the LIDA conceptual model aims at being a cognitive "theory of everything." With modules or processes for perception, working memory, episodic memories, "consciousness," procedural memory, action selection, perceptual learning, episodic learning, deliberation, volition, and non-routine problem solving, the LIDA model is ideally suited to provide a working ontology that would allow for the discussion, design, and comparison of AGI systems. The LIDA architecture is based on the LIDA cognitive cycle, a sort of "cognitive atom." The more elementary cognitive modules and processes play a role in each cognitive cycle. Higher-level processes are performed over multiple cycles. In addition to giving a quick overview of the LIDA conceptual model, and its underlying computational technology, we argue for the LIDA architecture's role as a foundational architecture for an AGI. Finally, lessons For AGI researchers drawn from the model and its architecture are discussed.

Introduction

Early AI researchers aimed at what was later called "strong AI," the simulation of human level intelligence. One of AI's founders, Herbert Simon, claimed (circa 1957) that "… there are now in the world machines that think, that learn and that create." He went on to predict that with 10 years a computer would beat a grandmaster at chess, would prove an "important new mathematical theorem, and would write music of "considerable aesthetic value." Science fiction writer Arthur C. Clarke predicted that, "[AI] technology will become sufficiently advanced that it will be indistinguishable from magic" [1]. AI research had as its goal the simulation of human-like intelligence.

Within a decade of so, it became abundantly clear that the problems AI had to overcome for this "strong AI" to become a reality were immense, perhaps intractable. As a result, AI researchers concentrated on "weak AI" (now often referred to as "narrow AI"), the development of AI systems that dealt intelligently with a single narrow domain. An ultimate goal of artificial human-level intelligence was spoken of less and less.

As the decades passed, narrow AI enjoyed considerable success. A killer application, knowledge-based expert systems, came on board. Two of Simon's predictions were belatedly fulfilled. In May of 1997, Deep Blue defeated grandmaster and world chess champion Garry Kasparov. Later that year, the sixty-year-old Robbins conjecture in mathematics was proved by a general-purpose, automatic theorem-prover [2]. Narrow AI had come of age.

More recently, and perhaps as a result, signs of a renewed interest in a more human-like, general artificial intelligence began to appear. An IEEE Technical Committee on Autonomous Mental Developmental was formed, aimed at human-like learning for software agents and mobile robots. Motivated by the human autonomic nervous system, IBM introduced self-managed computer systems, called autonomic systems, designed to configure themselves, to heal themselves, to optimize their performance, and to protect themselves from attacks. In April of 2004, DARPA, the Defense Advanced Research Projects Agency sponsored a workshop on Self-Aware Computer Systems which led to a call for proposals to create such systems. AAAI-06 had a special technical track on integrated intelligent capabilities, inviting papers that highlight the integration of multiple components in achieving intelligent behavior. All these are trends toward developing an artificial, human-like, general intelligence.

The next major step in this direction was the May 2006 AGIRI Workshop, of which this volume is essentially a proceedings. The term AGI, artificial general intelligence, was introduced as a modern successor to the earlier strong AI.

Artificial General Intelligence

What is artificial general intelligence? The AGIRI website lists several features, describing machines

- with human-level, and even superhuman, intelligence.
- that generalize their knowledge across different domains.
- that reflect on themselves.
- and that create fundamental innovations and insights.

Even strong AI wouldn't push for this much, and this general, an intelligence. Can there be such an artificial general intelligence? I think there can be, but that it can't be done with a brain in a vat, with humans providing input and utilizing computational output. Well, if it can't be a brain in a vat, what does it have to be? Where can one hope to create artificial general intelligence (AGI)?

Autonomous Agents

Perhaps it can be created as an autonomous agent. And what's an autonomous agent? Biological examples include humans and most (all?) animals. Artificial autonomous agent can include software agents such as the bots that serve to populate Google's databases, and our IDA that does personnel work for the U.S. Navy ([3], [4]). Other such examples include some mobile robots and computer viruses. But what do I mean by *autonomous agent*? Here's a definition [5]. It's a system that

- is embedded in an environment,
- is a part of that environment,
- which it senses,
- and acts on,
- over time,
- in pursuit of its own agenda (no human directs its choice of actions),
- so that its actions may affect its future sensing (it's structurally coupled to its environment ([6], [7])).

The earlier features of an autonomous agent listed in the definition are typically accepted in the community of agent researchers. The final feature is needed to distinguish an autonomous agent from other kinds of software, like a check-writing program that reads a personnel database once a week and produces a check for each employee.

Why must an AGI system be an autonomous agent? In order for an AGI system to generalize its knowledge across different, and likely novel, domains, it will have to learn. Learning requires sensing, and often acting. An autonomous agent is a suitable vehicle for learning [8], particularly for human-like learning ([9], [10], [11]).

An Agent in its Environment

So the picture I'm going to propose as the beginning of a suggested ontology for AGI research is developed from that of an agent that senses its environment and acts on it, over time, in pursuit of its own agenda.

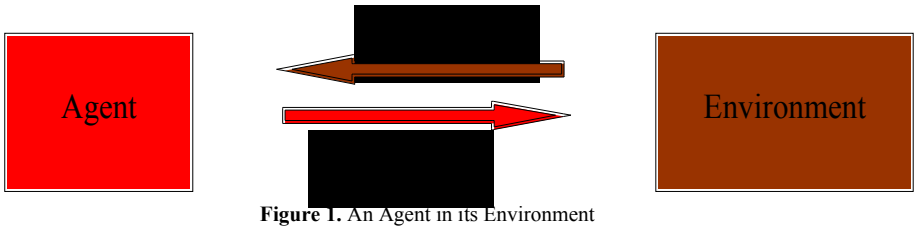

Figure 1. An Agent in its Environment

In order to do all this, it must have built in *sensors* with which to sense, it must have *effectors* with which to act, and it must have *primitive motivators* (which I call *drives*), which motivate its actions. Without motivation, the agent wouldn't do anything. Sensors, effectors and drives are primitives that must be built into, or evolved into, any agent.

Cognition

Next we replace the agent box in Figure 1 with a box called Cognition (see Figure 2).

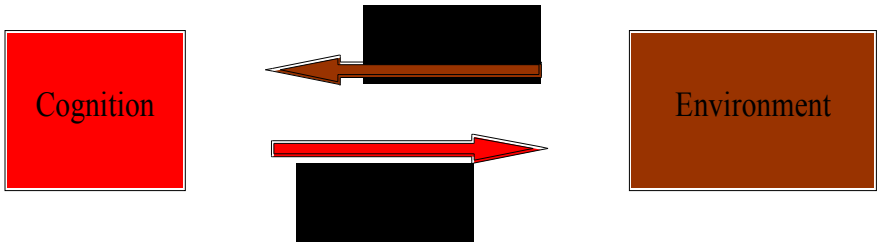

Figure 2. Cognition

For any autonomous agent, including we humans, the one significant, and constantly recurring question is "what to do next" ([12], Chapter 16). Cognition is the term we will use for this unending succession of deciding "what to do next," in some sense the only question there is. We humans face that question every moment of our existence, as must our artificial autonomous agents. In humans, such selected actions include movements in the hands, turning of the head, saccades of the eyes, movements of the lips and tongue when speaking, and all sorts of other actions.

Do note that this use of the term cognition is broader than the way the psychologists use it. When they talk about cognition, they normally don't include perception or the process of actually taking the action.

In what follows, we'll break out modules and processes, one or two at a time, from the Cognition box, replacing it with a gray Rest of Cognition box, and talk about such modules and processes individually. Each such module or process will become part of the developing ontology, hopefully acting as a foundation for AGI. I know you'll think this won't sound anything like AGI. It's far from general intelligence, but in my view this is where we have to start. This simple ontology will lead us to the beginnings of a foundational architecture for AGI.

Perception

Let's first break out *perception* (see Figure 3), that is, the process of assigning of meaning to incoming sensory data. All this sensory data coming in must, somehow, be made meaningful to the agent itself. The agent must recognize individual objects, must classify them, and must note relations that exist between objects. The agent must make sense of the "scene" in front of it, perhaps including a visual scene, an auditory scene, an olfactory scene, etc. Perception is the process of agent making sense of its world.

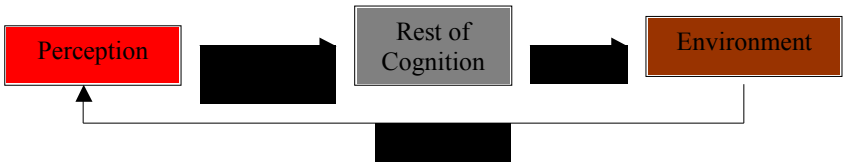

Figure 3. Perception

And, what does meaning mean? How do you measure meaning? In my view, it's best measured by how well the meaning assists the agent in deciding what to do next, in action selection. This process of assignment of meaning can be bottom-up, that is, drawn immediately from the sensation. It can also be top-down with older meanings coming back around in additional cycles and contributing to later meanings. Perception can also be top-down within a single cycle, in that it can look back more than once at the incoming sensory data. In the diagram of Figure 3, we break out perception at one end, and have it construct a *percept* that sends the meaningful aspects of the scene forward. Top-down influences of both types may contribute to the percept, contributing to the sense of the scene.

Procedural Memory

Next we're going to break out *procedural memory* (see Figure 4), by which I mean a repertoire of tasks (actions, procedures). Actions from this repertoire can be executed singly, in parallel, in series, and even in more complex streams of actions. In addition to the actions themselves, procedural memory might also keep track of the context in which the action may prove useful, as well as the expected result of performing the action.

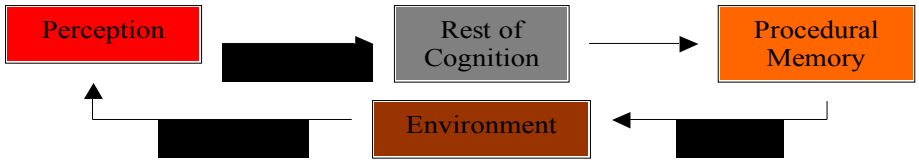

Figure 4. Procedural Memory

But don't confuse procedural memory, which has to do with WHAT to do next, with sensory-motor memory, which keeps track of HOW to do it. How do I pick up a glass, turn it over and put it back? Deciding to pick it up is one thing; actually doing it is something different. These two kinds of memory should require different mechanisms.

Episodic Memory

Next we'll break out *episodic memory* (see Figure 5), that is, the content-addressable, associative memory of events, of the what, the where and the when of the previous happening. These episodic memories may include semantic, locational, temporal, emotional, and causal aspects of the event.

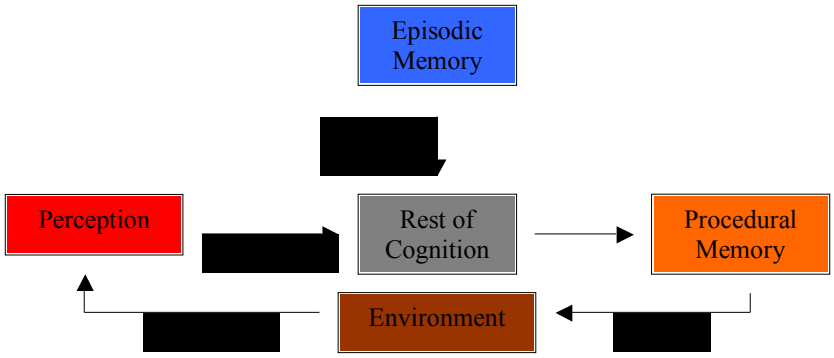

Figure 5. Episodic Memory

Normally in humans, recall from episodic memory is accomplished by some kind of internal virtual reality. We build mental images, using them to partially re-live the event. These images may be visual, auditory, or whatever. Might such virtual imagery recall be possible, or useful, in artificial agents, I don't know. Animal cognition researchers often try to avoid controversy about animal consciousness by referring to "episodic-like" memory, defining it as the storing of the what, the when, and the where without any assumption of mental imaging ([13], [14]).

Episodic memories come in several varieties. *Transient episodic memory* has a decay rate measured in hours or perhaps a day ([15], [4], [8], [16], [17]). Long-term episodic memories have the potential of storing information indefinitely. Long-term *declarative memory* includes *autobiographical memory*, the memory of events as described above, and *semantic memory*, the memory for facts.

Attention & Action Selection

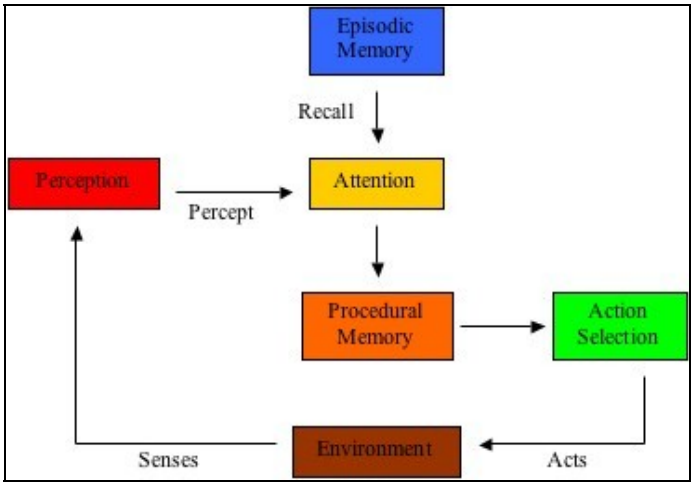

Figure 6. Attention and Action Selection

In this section the gray "rest of cognition box" has disappeared, to be replaced by attention and action selection. *Attention* is the process that brings information built from perception and from episodic memory to consciousness. The global workspace theory of consciousness ([18], [19], [20]) postulates a competition for consciousness (see Cognition as Filtering below). The competition aims at selecting the most relevant, the most important, the most urgent, or the most insistent information to become conscious. This is a functional view of consciousness, and takes no stand on the possibility of subjective machine consciousness in an AGI [21].

The winning conscious information serves to recruit internal resources from which the next task is selected by the *action selection* mechanism. For an AGI such action selection must be quite sophisticated. In particular, it must be able to choose well between tasks serving different concurrent goals. It also must be able to bounce between seeking two such concurrent goals so as to take advantage of opportunities offered by the environment.

Cognition as Filtering

Following the *cognitive cycle* displayed in Figure 6 above, we can usefully think of each step as a filtering process. An agent's sensory receptors filter all of the possible

sensory data available in the environment, letting through only that to which the agent's sensors respond. Perception, as described above, is also a filtering process. Some sensory data coming in are ignored, while others are processed into possibly useful information, and become part of the percept that moves forward. The recall associations returned from a possibly huge episodic memory, accumulated over sometimes extremely long time periods, are also the result of a filtering process, What's wanted is information relevant to, and important for, the agent's current situation, including its goals. Hopefully, that's what comes out of this filtering process so far. Attention is yet another filtering process that decides what part of the recent percepts and episodic recall to bring to consciousness. Again the criteria for this filtering include relevance, importance, urgency, and insistence. Procedural memory then uses the contents of consciousness, what comes to attention, to recruit only those actions that might be possible and useful in the current situation, yet another filtering process. Our final filtering process is action selection, the process of choosing what single action to perform next.

The more complex the environment, and the more complex the agent, the more filtering is needed. One can think of the whole cognitive cycle, and even of cognition itself, as being essentially a complex, compound, filtering process.

Learning

Since, by its very nature, an AGI must learn, we next add several sorts of learning to the cognitive cycle (see Figure 7 below). Our assumption is that the agent learns that to which it attends ([18] section 5.5). Thus the learning arrows, in red, immerge from the Attention box. You'll see three different kinds of learning denoted here, though there are others. There's *perceptual learning*, the learning of new meanings, that is of objects, categories, relations, etc., or the reinforcement of existing meanings. The *episodic learning* of events, the what, the where and the when, is denoted by its encoding. Finally, *procedural learning* improves skills and/or to learns new skills.

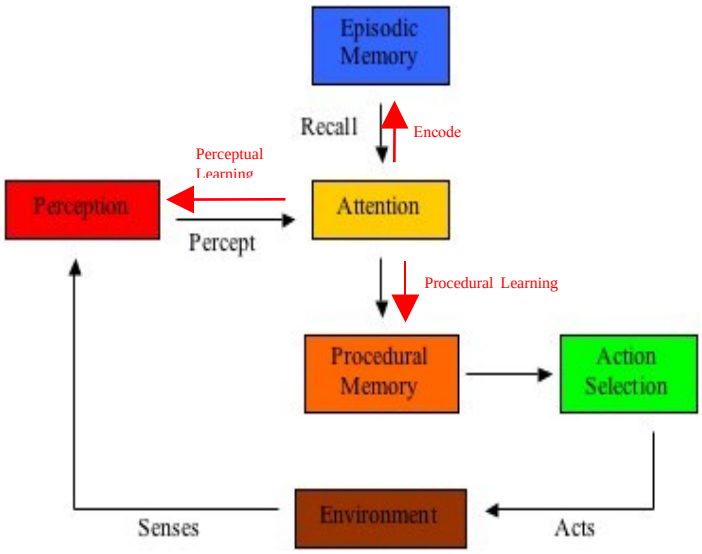

Figure 7. Learning

A Foundational Architecture for AGI

So, if we're going to aim for an AGI, where do you look for it? How should we go about trying to build an AGI agent? In my view, if you want smart software, copy it after a human. That is, model the early AGI agents on what we know about human cognition. In the previous sections we've discussed modules and processes, derived from human cognition, that we believe must be included in any AGI architecture. Where do we go from there? One possibility is to dive right in and attempt to build a full-blown AGI directly. This strategy, while surely ambitious, may well succeed. A second possible strategy might be to construct a sequence of increasingly complex, intelligent, and general, artificial agents, culminating in a true AGI. This second strategy may prove to be even more likely to succeed.

Here we suggest a means of enabling this second strategy by way of a common foundational architecture for each agent in the sequence. Such a foundational architecture would allow each successive agent to be built by adding higher-level cognitive processes to its predecessor. Let's assume, as we must ([22], [23], [24], [25], [9], [10]), that learning via a developmental period must be an integral part of the life cycle of any AGI. The strategy suggested might allow what's learned by one robot to be initially incorporated into its immediate successor.

Any autonomous agent, and hence any AGI, must operate by means of a continuing iteration of cognitive cycles, the sense-cognize-act cycles described above. Each such cognitive cycle acts as a cognitive moment, an atom of cognition, in that

each higher-level cognitive process is performed via the execution of a sequence of cognitive cycles. Higher-level cognitive processes are built of these cognitive cycles as cognitive atoms. Thus, a foundational architecture for AGI must implement a cognitive cycle to be continually iterated, and must provide mechanisms for building higher-level cognitive processes composed of sequences of these cognitive cycles. The LIDA architecture, to be described next, accomplishes both.

The LIDA Architecture

IDA denotes a conceptual and computational model of human cognition. LIDA, short for Learning IDA, denotes another such model with learning added. Let's start with a brief description of IDA.

The US Navy has about 350,000 sailors. As each sailor comes to the end of a certain tour of duty, he or she needs a new billet, a new job. The Navy employs some 300 detailers, as they call them, personnel officers who assign these new billets. A detailer dialogs with sailors, usually over the telephone, but sometime by email. These detailers read personnel data from a sailor's record in a Navy personnel database for items bearing on qualifications. They check job requisition lists in another Navy database to see what jobs will come available and when. They enforce the Navy's policies and try to adhere to the sailors' wishes, as well as looking to the needs of the particular job. Eventually, the detailer offers one, two or, rarely, three jobs to the sailor. Some back and forth negotiations ensue, involving several communications. Hopefully the sailor agrees to takes a job offered by the detailer. If not, the detailer simply assigns one.

IDA, an acronym for Intelligent Distribution[1] Agent, is an autonomous software agent, which automates the tasks of the detailers as described in the previous paragraph ([25], [26]). Built with Navy funding, IDA does just what a human detailer does. In particular, she communicates with sailors in natural language, in English, though by email rather than telephone. The sailor writes anyway he or she wants to write. There's no prescribed protocol or format, no form to fill out. IDA understands what the sailor writes in the sense of knowing how to pick out relevant and important pieces of information from the email message, and what to do with it. IDA is implemented, up and running, and tested to the Navy's satisfaction.

To accomplish the tasks of a human detailer, IDA employs a number of higher-level cognitive processes. These include constraint satisfaction [27], deliberation ([28], [29]), sophisticated action selection [30] and volition [29].

Both in its cognitive cycle and its implementation of higher-level cognitive processes, IDA, and its learning extension LIDA, implement a number of mostly psychological theories of cognition. We'll very briefly describe each, and its role in the LIDA architecture.

Over the past couple of decades, research in AI, and more generally cognitive science, has moved towards situated or embodied cognition [31]. The idea is that cognition should be studied in the context of an autonomous agent situated within an environment. Being an autonomous software agent, IDA is embodied [32]. Similarly, software agents, autonomous robots or AGI's built on the foundation of a LIDA architecture would be embodied.

[1] Distribution is the Navy's name for the process of assigning new jobs to sailors at the end of a tour of duty.

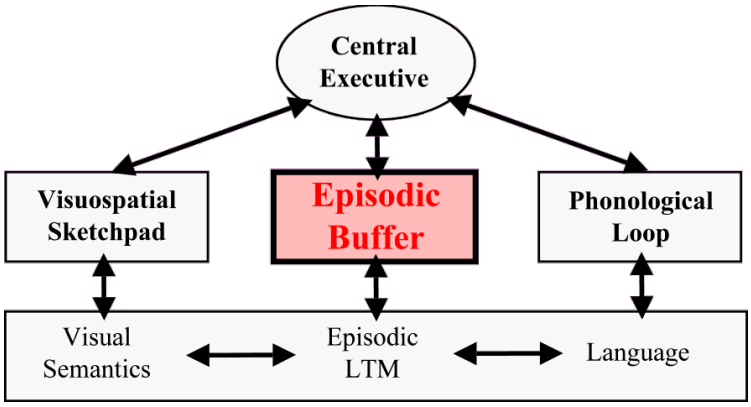

Figure 8. Working Memory

Barsalou, in his theory of perceptual symbol systems [33] postulates that there are no amodal symbols involved in human cognition. Rather, *all* such information is represented by perceptual symbols. Put another way, all cognitive symbols are ultimately grounded in perception[34]. The LIDA architecture represents perceptual entities, objects, categories, relations, etc., using nodes and links in a slipnet [35]. These serve as perceptual symbols acting as the common currency for information throughout the various modules of the LIDA architecture.

In cognitive psychology the term working memory refers to a theoretical framework specifying and describing structures and processes used for temporarily storing and manipulating information [36]. Among these structures are the visuospacial sketchpad, the phonological loop, and a central executive responsible for the integration of information. More recent working memory structures include a consciousness mechanism [37] and the episodic buffer [38] (see Figure 8). All of the various modules and processes working memory are implemented in the LIDA architecture, mostly in its perceptual module and its workspace (see below) [39].

Glenberg's theory [40] stresses the importance of patterns of behavior to conceptualization and to understanding. For example, an object is understood via its affordances [41]. In the LIDA architecture templates for such pattern of behavior are found in perceptual memory. Their instantiations as sequences of actions contribute to perceptual learning, including conceptualization, leading to further understanding.

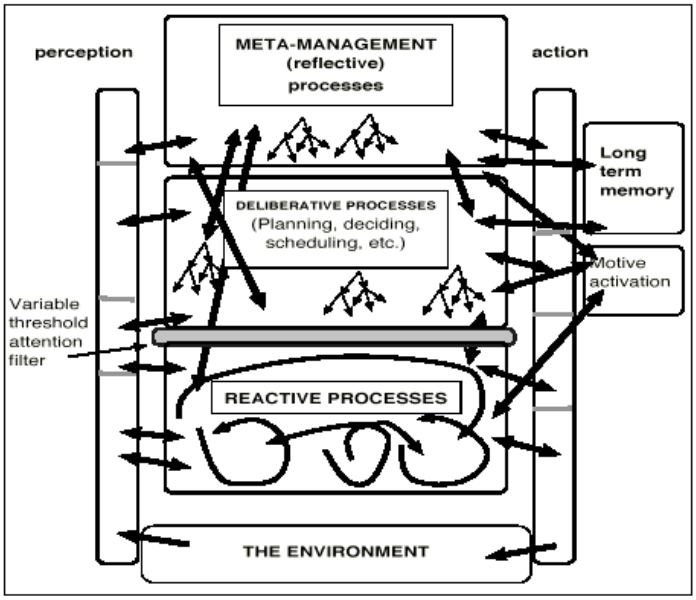

Figure 9. Sloman's Architecture

The long-term working memory of Ericsson and Kinstch [42] is incorporated into LIDA's workspace (see below), in which local associations recalled from episodic memory are combined with percepts to produce higher-level perceptual structures. In an AGI this workspace would include the various working memory structures mentioned above.

By far the single most significant influence on the LIDA architecture from cognitive psychology came from Baars' global workspace theory (GWT) of consciousness and cognition ([18], [19], [29], [39]). GWT postulates attention, bringing important, relevant information to consciousness (the global workspace). Its contents are then broadcast to the entire system in order to recruit internal resources with which to deal appropriately with the current situation. The LIDA architecture implements precisely this function, as will become clear during the discussion of the LIDA cognitive cycle below.

Finally, the LIDA architecture can be thought of a fleshing out and an implementation of Sloman's architecture for a human-like agent [28]. One can construct a concordance between most of the various modules and processes shown in Figure 9 and corresponding modules and processes of the LIDA cognitive as shown in Figure 10 below. Those that won't fit in to such a concordance correspond to higher-level, multi-cyclic cognitive processes in LIDA (see Multi-cyclic Cognitive Processes below). Sloman's meta-management process, that is, what the psychologist call metacognition, has not yet been designed for the LIDA model, but it certainly can be.

The LIDA Cognitive Cycle

LIDA operates as any autonomous must, with a continuously iterating cognitive cycle. Higher-level cognitive processes are composed of sequences of several or many of these cognitive cycles. Such higher-level cognitive processes might include deliberation, volition, problem solving, and metacognition.

Let's take a quick, guided tour through LIDA's cognitive cycle, which is based on Figure 7 above. Figure 10 below will provide a useful map for our tour. Note, that this cognitive cycle is highly complex, and yet all of this must be accomplished in every cognitive moment. Computational resources may well prove an issue.

Beginning at the upper left of Figure 10, we see stimuli coming in from both the internal and the external environment. Recall that, by definition, every autonomous agent is a *part* of its environment. LIDA is modeled after humans; we have to deal with both external and internal stimuli. Any AGI will likely have to also do so.

In *Sensory Memory* (SM) one would find the sensors themselves and primitive, that is, built-in, feature detectors. It would also include early learned, and therefore not primitive, feature detectors that provide the beginnings of understanding of the stimuli. Note that information from SM goes both to Perceptual Associative Memory, which we'll discuss next, and to the Effectors via the SMA (sensory-motor automatisms). In the later role, SM is crucially involved in quickly providing the kind of precise spatial, temporal, egocentric information that permit such actions as successfully hitting an oncoming fast ball, or even the grasping of a cup. Such SMA's in humans operate on their own direct sensory-motor cycles at about five times the rate of the larger LIDA cognitive cycle.

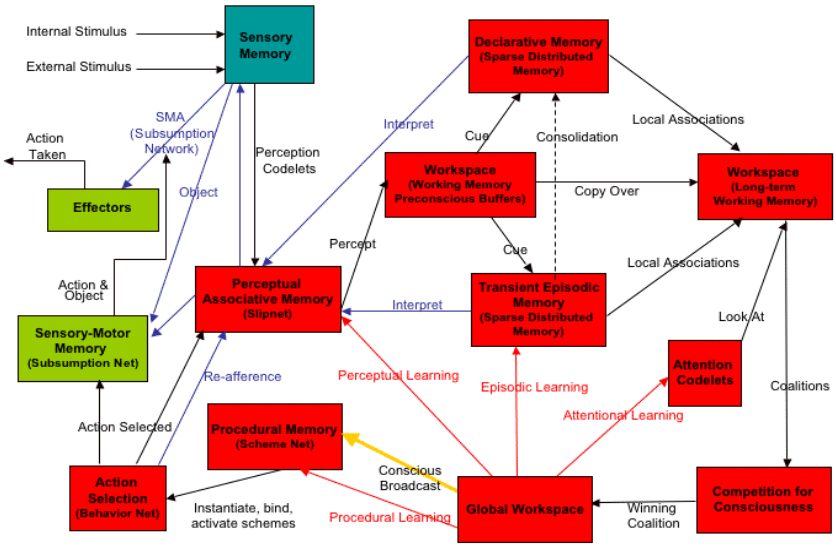

Figure 10. The LIDA Cognitive Cycle

From SM, information travels to *Perceptual Associative Memory* (PAM), which we implement as a slipnet [35]. Here the next stage of constructing meanings occur, that is the recognition of further features, of objects, and of categories. Passing activation brings some nodes and links over threshold, and thus into the *percept*.

The LIDA cognitive cycle includes two episodic memory modules, the short-term *Transient Episodic Memory* (TEM), and the potentially long-term *Declarative Memory* (DM) ([15], [4], [8], [16]). Recording such information as where I parked my car in the garage this morning, TEM encodings decay in humans within hours or a day. DM encodings only occur through offline consolidation from TEM. Though they can decay away, when sufficiently reinforced DM encodings can last a lifetime. Both episodic memories are computationally implemented using a modified sparse distributed memory ([43], [44], [45]).

The percept produced by PAM (described two paragraphs above) is moved into the *Workspace*, an amalgam of the preconscious working memory buffers and long-term working memory (in Figure 10 the Workspace is split into two boxes). Here additional, more relative, less precise, scene understanding structures are built. As well as the current percept, the Workspace also contains previous percepts and recent local associations recalled from both TEM and DM, all in various stages of decaying away. These contents of the Workspace serve to cue TEM and DM for current local associations. An understanding of the current scene is produced in the Workspace using additional, quick, two-way communication, that is, including down-stream communication, with TEM, DM, PAM and even SM.

Next, the *Attention Codelets*[2], whose job it is to bring relevant and important information to consciousness, come into play. An attention codelet has its own special interests, to which it wishes to draw attention. Each attention codelet searches the workspace for items (objects, relations, situations) of interest, and creates coalitions[3] of these items if it finds them.

These coalitions move into the Global Workspace where there's a competition for consciousness. This competition constitutes the final filtering of input. The idea is to attend to filter most relevant, the most important, the most urgent, the most insistent aspects of the current situation. Once the competition for consciousness is resolved, GWT call for a global broadcast of the contents of consciousness.

Aside from learning, which we'll discuss later, the major recipient of the global broadcast is *Procedural Memory* (PM), which we implement as a scheme net modeled after the schema mechanism [46]. M uses the contents of the global broadcast to pick out those possible actions that might be relevant to the current situation. Each scheme in PM is a template for an action together with its context and result. The schemes that might be relevant, that is, those whose context and/or results intersect with the contents of the global broadcast, including goals, instantiate themselves and bind their variables with information from the broadcast.

These instantiated schemes then go to Action Selection (AS), which is implemented as a behavior net ([47], [30]), a very sophisticated kind of action selection mechanism. In AS, instantiated schemes compete to be the single action selected, possibly a compound of sub-actions in parallel. Over multiple cognitive cycles, AS may select a sequence of actions to accomplish a given goal. It might also bounce opportunistically between sequences of actions serving different goals.

The single action chosen during a given cognitive cycle is sent, along with the object(s) upon which it is to act, to Sensory-Motor Memory (S-MM), which contains procedures for actually performing the selected action, the so called sensory-motor automatisms. Our representation for these sensory-motor automatisms is as yet undecided, but we're leaning toward a net built from subsumption networks [48].

In our tour through the LIDA cognitive cycle we postponed a discussion learning, which we'll take up now. Our basic premise is that we learn that to which we attend ([18] pp 213-214). Thus learning occurs as a consequence of, or at least in conjunction with, the conscious broadcast from the Global Workspace. Learning is modulated by affect following an inverted U curve. Learning is strengthened as affect increases up to a point. After that the affect begins to interfere and the learning rate diminishes with further increases in affect ([49], [50]).

The LIDA cognitive cycle includes four types of learning, three of which were discussed earlier in the chapter [9]. The perceptual learning of object, categories, relations, etc., takes place in PAM[8]. Episodic learning of what, where and when are encoded in TEM, while procedural learning of tasks takes place in PM [10]. The hitherto unmentioned form of learning is attentional learning, the learning of what to attend, which takes place in the Attention Codelets. We know little about attentional learning, which is an object of current research.

Each of these types of learning has its selectionist and its instructionalist form [51]. Selectionist learning reinforces existing memory traces positively or negatively. Instructionalist learning adds new entities to the various memories, often by altering or combining existing entities.

---

[2] Taken from the Copycat Architecture (Hofstadter and Mitchell 1995), "codelet" refers to a small, special-purpose piece of computer code, often running as a separate thread. They implement the processors of GWT (Baars 1988). There are many different varieties of codelet in the LIDA model.

[3] The term "coalition" comes from GWT, where it always refers to a coalition of processors.

Following our strategy, mentioned above, of producing smart software by copying humans, the LIDA cognitive cycle was modeled after what we hypothesize happens in humans ([39], [4], [8], [16]). Though asynchronous, each cognitive cycle runs in about 200 milliseconds. But they can cascade, so a new cycle can begin while earlier cycles are completing. As a consequence of this cascading, the rate of this cognitive cycle processing is five to ten cycles per second. Though asynchronous, the seriality of consciousness must be preserved. Though none of it is conclusive, there considerable evidence from neuroscience suggestive or, or supportive of, these cognitive cycles in nervous systems ([52], [53], [54], [55]).

Multi-cyclic Cognitive Processes

In the LIDA model cognitive cycles are the atoms out of which higher-level cognitive processes are built. Here we'll briefly describe several of these higher-level processes: deliberation, volition, atomization, non-routine problem solving, metacognition and self-awareness. Each of these is a multi-cyclic process that can be implemented over multiple cognitive cycles using the LIDA architecture as a foundation. Let's take them up one at a time, beginning with deliberation.

Deliberation refers to such activities as planning, deciding, scheduling, etc. that require one to consciously think about an issue. Suppose I want to drive from a new location in a city I know to the airport. It will be a route I've never taken, so I may imagine landmarks along the way, which turns to take and so, deliberate about how best to get there. When IDA thinks about whether she can get a sailor from a current job to a specific new job with leave time, training time, travel time and so forth all fitted in between, that's deliberation. This higher-level deliberative process takes place in IDA (and LIDA) over multiple cognitive cycles using behavior streams instantiated from PM into the behavior net (AS) [29].

As specified by GWT, conscious, volitional, decision-making, a kind of deliberation, is implemented via William James' ideomotor theory ([56], [18], [29]). Once again, *volition* uses an instantiated behavior stream over several cognitive cycles. For example, suppose that, being thirsty one morning, I consciously considered the possibilities of coffee, tea, and orange juice, weighing the advantages and disadvantages of each, perhaps by arguing with myself. My eventually deciding to drink tea is a volitional decision, as opposed to my typing of this phrase, which was not consciously decided on ahead of time. IDA decides volitionally on which jobs to offer sailors.

How do we get from consciously going through all the steps of learning to drive an automobile, to the effortless, frequently unconscious, automatic actions of an experienced driver? We call this higher-level cognitive process *automization*, and implement it in the LIDA model via pandemonium theory ([57], [58]). Once again automization is accomplished over multiple cognitive cycles using the LIDA architecture as a framework.

In the LIDA architecture Procedural Memory (PM) consists of templates for actions, including their contexts and expected results. Actions are selected from among action templates instantiated in response to a conscious broadcast. What if PM doesn't contain any action templates to be recruited to deal with the current situation? In this case *non-routine problem solving* would be required. The LIDA architecture serves as a foundation for an, as yet unpublished, non-routine problem solving algorithm based on an extension of partial order planning [59].

Defined by psychologists as thinking about thinking, *metacognition*[4] has, in recent years become of interest to AI researchers ([60], [28], [61]). There's even a website for Metacognition in Computation (www.cs.umd.edu/~anderson/MIC). Metacognition is often used to update a strategy. Suppose I think that I was too hard on my daughter in our interaction last night, and decide that next time I want to be more empathetic with her. That's an example of metacognition. After early, misguided attempts (see for example, [62]), we now know how to build metacognition as a collection of higher-level cognitive processes on a foundation of the LIDA architecture and its cognitive cycle. This work is currently in an early stage and not yet published.

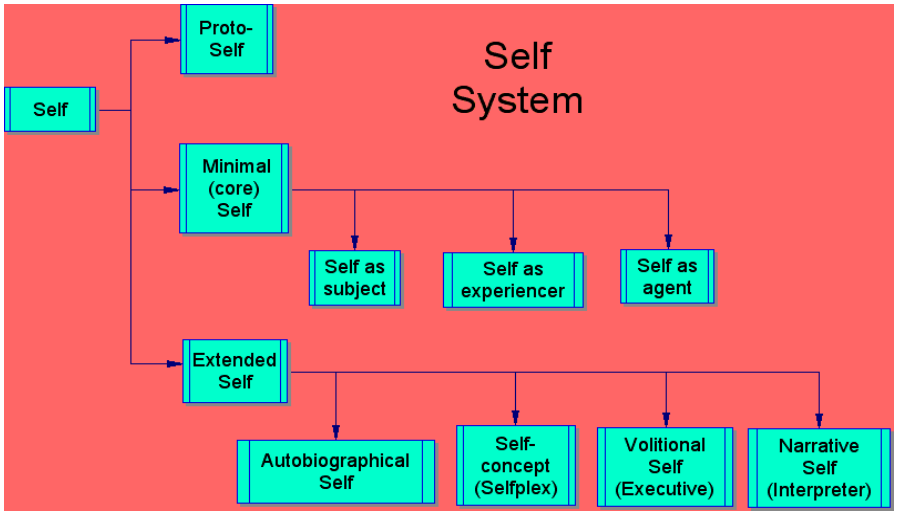

Figure 11. Various Selves

Philosophers, psychologists and neuroscientists have defined and studied a number of varieties of selves ([63], [64], [65], [66], [67]) (see Figure 11) Finally, it's possible to implement several of the various varieties of *self* as higher-level cognitive processes on a foundation of the LIDA architecture. Again, this work has currently just begun and is as yet unpublished.

All of these and many, many more multi-cyclic processes can be built using the LIDA architecture's cognitive cycles as cognitive atoms. It's this possibility that supports the strategy of producing an AGI as a sequence of ever more intelligent, adaptable and versatile autonomous agents each containing the previous, and each based on the LIDA architecture.

[4] Sloman calls it meta-management (see Figure 9).

Lessons for Building an AGI

Suppose we accept the strategy of building an AGI as the culmination of an increasing sequence of ever more intelligent and adaptable AGI agents, each built on the foundation of the LIDA architecture with its cognitive cycles as atoms. What general lessons can we learn as a result? Here are a few.

We must choose a suitable domain for our agent. A domain? A domain for an AGI agent? I thought an AGI was supposed to generalize. It certainly must generalize, but it's still an autonomous agent. Every such agent must come with built-in sensors, motivators, and effectors. That means the agent must have an environment on which to sense and act, that is, a domain. What is needed is a well-chosen domain from which it can generalize. This would entail a broad enough domain with a number of sub-domains from which it can generalize. The successor of each agent in the sequence may be situated in a more inclusive domain, and may be provided with additional sensors, motivators and effectors.

In my view, an AGI agent is much too much to handcraft. By definition, it's supposed to generalize, that is, to add to its store of knowledge and skill. Therefore it must learn. And, how shall it learn? At least at the start, I suggest that it learn like a human, that we build-in human-like learning capabilities. Later on we, or it, may find better ways of learning. Let's note some principles of human learning that can be adapted to human-like learning in an AGI agent, and in its predecessors.

There's no learning from scratch, from a blank slate. For example, human infants come equipped to recognize faces. The practice of the more sophisticated machine learning research community is to build in whatever you can build in. This same principle should be followed when attempting to build an AGI. Learning, yes. Learning from scratch, no.

With trivial exceptions, we learn that to which we attend, and only that. The implication is that an AGI must come equipped with an attention mechanism, with some means of attending to relevant information. This implies the need for some form of functional consciousness, but not necessarily subjective consciousness [21].

Human learning is incremental and continual. It occurs at every moment, that is, during every cognitive cycle. And, it's unsupervised. Supervised machine learning typically involves a training period during which the agent is taught, and after which it no longer learns. In contrast, an AGI agent will need to learn incrementally and continually as human's do. Though such an agent may go through a developmental period of particularly intense learning, it must also be a "lifelong" learner.

Humans learn by trial and error, that is, by what we in AI call a generate-and-test process. The LIDA model hypothesizes that we learn potential new objects in PAM quickly and on the flimsiest of excuses [8]. This is the process of generation. New objects that are reinforced by being attended to survive, while others decay away. This is the testing process. All this is done incrementally and continually, that is, in every cognitive cycle. And, this perceptual learning by generate-and-test is not restricted to new objects, but applies to categories, relations, etc. Similar processes are in place for episodic and procedural learning as well. I suggest that such generate-and-test learning will be needed in AGI agents as well.

According to the LIDA model, much if not all of human memory is content addressable. We don't access an item in memory by knowing its address or index. Rather we access it using a portion of its content as a cue. Sensory memory is cued by the incoming content of the sensors. In PAM detected features allow us to access objects, categories, relations, etc. The contents of LIDA's workspace cue both episodic

memories, TEM and DM, recalling prior events associated with the cue. Action templates in PM are cued by the contents of the conscious broadcast. Such content addressable memories must surely be a part of any AGI agent.

Above we distinguished and spoke of selectionist and instructionalist learning within the LIDA architecture. I suggest that an AGI agent must also learn in each of these learning methods in each of its learning modes, perceptual, episodic, and procedural. In summary, the LIDA model suggests that an AGI agent must initially be copied after humans, must have a rich and broad domain, must employ many multi-cyclic processes, and must be capable of using both learning methods in the several different modes of learning.

Questions for AGI Researchers

Must an AGI agent be functionally conscious? As noted above, the LIDA model suggests an affirmative answer. Though functional consciousness as derived from GWT may not prove necessary, I suspect some form of attentional mechanism will.

Must an AGI agent be phenomenally conscious? That is, must it have subjective experiences as we do? I think not. I suspect that we may be able to build an AGI agent that's not phenomenally conscious. However, subjective consciousness may prove necessary to deal with the problem of distinguishing perceived motion due to changes in the environment from perceived motion due to movement of the agent's sensors ([68], [16]). Subjective consciousness provides an agent with a coherent, stable internal platform from which to perceive and act on its world. We may be pushed into trying to build AGI agents that are phenomenally conscious.

Must an AGI agent be capable of imagination? That is, must it be able to produce an internal virtual reality? Humans often deliberate in this way. An AGI agent must certainly be capable of deliberation. However, deliberation has been shown to be implementable without subjective consciousness ([8], [21]).

Must an AGI agent come equipped with feelings? In humans, feelings include, for example, thirst and pain, as well as emotions[5] such as fear or shame [67]. In humans and animals, feelings implement motivation. It's feelings that drive us to do what we do ([68], [69]). We think we select actions rationally, but such decisions, though surely influenced by facts and expected consequences, are ultimately in the service of feelings. And, feelings modulate our learning by increasing affect, as discussed above [68]. Must an AGI agent have artificial feelings to serve these purposes? I think not. There are other ways to modulate learning; there are other ways to implement drives as primitive motivators. Feelings have proved useful evolutionarily in complex dynamic environments because they provide a lot of flexibility in learning and action selection. They may provide a good solution to these problems in AGI agents, or we may find a better way.

Acknowledgements

The author is indebted to the many members, past and present, of the Cognitive Computing Research Group at the University of Memphis. Their individual and join efforts contributed significantly to the ideas presented in this chapter. In particular, thanks are due to Sidney D'Mello for a graphic on which Figure 10 is based, and to Michael Ferkin who contributed to the development of the suggested ontology.

[5] Following [69] we think of emotions as feelings with cognitive content. For example, one might be afraid of the rapidly approaching truck, the cognitive content.

References

[1] Clarke, A. C. 1962. Profiles of the Future; an Inquiry into the Limits of the Possible. New York: Harper & Row.

[2] McCune, W. 1997. Soltion of the Robbins Problem. *Journal of Automated Reasoning* 19:263-275.

[3] Franklin, S. 2001. Automating Human Information Agents. In *Practical Applications of Intelligent Agents*, ed. Z. Chen, and L. C. Jain. Berlin: Springer-Verlag.

[4] Franklin, S. 2005a. A "Consciousness" Based Architecture for a Functioning Mind. In *Visions of Mind*, ed. D. N. Davis. Hershey, PA: Information Science Publishing.

[5] 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.

[6] Maturana, H. R. 1975. The Organization of the Living: A Theory of the Living Organization. *International Journal of Man-Machine Studies* 7:313-332.

[7] Maturana, R. H. and F. J. Varela. 1980. *Autopoiesis and Cognition: The Realization of the Living, Dordrecht*. Netherlands: Reidel.

[8] Franklin, S. 2005b. Perceptual Memory and Learning: Recognizing, Categorizing, and Relating. Symposium on Developmental Robotics. American Association for Artificial Intelligence (AAAI). Stanford University, Palo Alto CA, USA. March 21-23, 2005.

[9] D'Mello, S. K., S. Franklin, U. Ramamurthy, and B. J. Baars; 2006. A Cognitive Science Based Machine Learning Architecture. AAAI 2006 Spring Symposium Series. American Association for Artificial Intelligence. Stanford University, Palo Alto, California, USA. March.

[10] D'Mello, S. K., U. Ramamurthy, A. Negatu, and S. Franklin. 2006. A Procedural Learning Mechanism for Novel Skill Acquisition. In *Workshop on Motor Development*: *Proceeding of Adaptation in Artificial and Biological Systems, AISB'06*, vol. 1, ed. T. Kovacs, and J. A. R. Marshall. Bristol, England: Society for the Study of Artificial Intelligence and the Simulation of Behaviour; April 2006.

[11] Ramamurthy, U., S. K. D'Mello, and S. Franklin. 2006. LIDA: A Computational Model of Global Workspace Theory and Developmental Learning. BICS 2006: Brain Inspired Cognitive Systems. October 2006.

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

[13] Clayton, N. S., D. P. Griffiths, and A. Dickinson. 2000. Declarative and episodic-like memory in animals: Personal musings of a scrub jay or When did I hide that worm over there? In *The Evolution of Cognition*, ed. C. M. Heyes, and L. Huber. Cambridge, MA: MIT Press.

[14] Ferkin, M. H., A. Combs, J. delBarco-Trillo, A. A. Pierce, and S. Franklin. in review. Episodic-like memory in meadow voles, *Microtus pennsylvanicus*. *Animal Behavior*.

[15] Conway, M. A. 2001. Sensory-perceptual episodic memory and its context: autobiographical memory. *Philos. Trans. R. Soc. Lond B.* 356:1375-1384.

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

[17] Franklin, S., B. J. Baars, U. Ramamurthy, and M. Ventura. 2005. The Role of Consciousness in Memory. *Brains, Minds and Media* 1:1-38, pdf.

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

[19] Baars, B. J. 1997. *In the Theater of Consciousness*. Oxford: Oxford University Press.

[20] Baars, B. J. 2002. The conscious access hypothesis: origins and recent evidence. *Trends in Cognitive Science* 6:47-52.

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

[22] Posner, M. I. 1982. Cumulative Development of Attentional Theory. *American Psychologist* 37:168-179.

[23] Franklin, S. 2000a. Learning in "Conscious" Software Agents. In *Workshop on Development and Learning*, ed. J. Wang. Michigan State University; East Lansing, Michigan, USA: NSF; DARPA; April 5-7, 2000.

[24] Weng, J. 2004. Developmental Robotics: Theory and Experiments. *International Journal of Humanoid Robotics* 1:119-234.

[25] Franklin, S., A. Kelemen, and L. McCauley. 1998. IDA: A Cognitive Agent Architecture. In *IEEE Conf on Systems, Man and Cybernetics*. : IEEE Press.

[26] McCauley, L., and S. Franklin. 2002. A Large-Scale Multi-Agent System for Navy Personnel Distribution. *Connection Science* 14:371-385.

[27] Kelemen, A., S. Franklin, and Y. Liang. 2005. Constraint Satisfaction in "Conscious" Software Agents - A Practical Application. *Applied Artificial Intelligence* 19:491-514.

[28] 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.

[29] Franklin, S. 2000b. Deliberation and Voluntary Action in 'Conscious' Software Agents. *Neural Network World* 10:505-521.

[30] Negatu, A., and S. Franklin. 2002. An action selection mechanism for 'conscious' software agents. *Cognitive Science Quarterly* 2:363-386.
[31] Varela, F. J., E. Thompson, and E. Rosch. 1991. *The Embodied Mind*. Cambridge, MA: MIT Press.
[32] Franklin, S. 1997. Autonomous Agents as Embodied AI. *Cybernetics and Systems.* 28:499-520.
[33] Barsalou, L. W. 1999. Perceptual symbol systems. *Behavioral and Brain Sciences* 22:577-609.
[34] Harnad, S. 1990. The Symbol Grounding Problem. *Physica D* 42:335-346.
[35] 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. A. Barnden. Norwood N.J.: Ablex.
[36] Baddeley, A. D., and G. J. Hitch. 1974. Working memory. In *The Psychology of Learning and Motivation*, ed. G. A. Bower. New York: Academic Press.
[37] Baddeley, A. 1992. Consciousness and Working Memory. *Consciousness and Cognition* 1:3-6.
[38] Baddeley, A. D. 2000. The episodic buffer: a new component of working memory? *Trends in Cognitive Science* 4:417-423.
[39] Baars, B. J., and S. Franklin. 2003. How conscious experience and working memory interact. *Trends in Cognitive Science* 7:166-172.
[40] Glenberg, A. M. 1997. What memory is for. *Behavioral and Brain Sciences* 20:1-19.
[41] Gibson, J. J. 1979. *The Ecological Approach to Visual Perception*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
[42] Ericsson, K. A., and W. Kintsch. Ericsson, K. A., and W. Kintsch. 1995. Long-term working memory. Psychological Review 102:211-245. Long-term working memory. Psychological Review 102:211-245.
[43] Kanerva, P. 1988. Sparse Distributed Memory. Cambridge MA: The MIT Press.
[44] Ramamurthy, U., S. K. D'Mello, and S. Franklin. 2004. Modified Sparse Distributed Memory as Transient Episodic Memory for Cognitive Software Agents. In *Proceedings of the International Conference on Sstems, Man and Cybernetics*. Piscataway, NJ: IEEE.
[45] D'Mello, S. K., U. Ramamurthy, and S. Franklin. 2005. Encoding and Retrieval Efficiency of Episodic Data in a Modified Sparse Distributed Memory System. In *Proceedings of the 27th Annual Meeting of the Cognitive Science Society. Stresa, Italy*.
[46] Drescher, G. L. 1991. *Made-Up Minds: A Constructivist Approach to Artificial Intelligence*. Cambridge, MA: MIT Press.
[47] Maes, P. 1989. How to do the right thing. *Connection Science* 1:291-323.
[48] Brooks, R. A. 1991. How to build complete creatures rather than isolated cognitive simulators. In *Architectures for Intelligence*, ed. K. VanLehn. Hillsdale, NJ: Lawrence Erlbaum Associates.
[49] Belavkin, R. V. 2001. Modelling the inverted-U effect with ACT-R. In *Proceedings of the 2001 Fourth International Conference on Cognitive Modeling*, ed. E. M. Altmann, W. D. Gray, A. Cleeremans, and C. D. Schunn. Hillsdale, NJ: Lawrence Erlbaum Associates.
[50] Cochran, R. E., F. J. Lee, and E. Chown. 2006. Modeling Emotion: Arousal's Impact on memory. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society (pp. 1133-1138) . Vancouver, British Columbia, Canada*. Vancouver, British Columbia, Canada.
[51] Edelman, G. M. 1987. *Neural Darwinism*. New York: Basic Books.
[52] Lehmann, D., H. Ozaki, and I. Pal. 1987. EEG alpha map series: brain micro-states by space-oriented adaptive segmentation. *Electroencephalogr. Clin. Neurophysiol.* 67:271-288.
[53] Lehmann, D., W. K. Strik, B. Henggeler, T. Koenig, and M. Koukkou. 1998. Brain electric microstates and momentary conscious mind states as building blocks of spontaneous thinking: I. Visual imagery and abstract thoughts. *Int. J. Psychophysiol.* 29:1-11.
[54] Halgren, E., C. Boujon, J. Clarke, C. Wang, and P. Chauvel. 2002. Rapid distributed fronto-parieto-occipital processing stages during working memory in humans. *Cerebral Cortex* 12:710-728.
[55] Freeman, W. J., B. C. Burke, and M. D. Holmes. 2003. Aperiodic Phase Re-Setting in Scalp EEG of Beta-Gamma Oscillations by State Transitions at Alpha-Theta Rates. *Human Brain Mapping* 19:248-272.
[56] James, W. 1890. *The Principles of Psychology*. Cambridge, MA: Harvard University Press.
[57] Jackson, J. V. 1987. Idea for a Mind. *Siggart* Newsletter, 181:23-26.
[58] Negatu, A., T. L. McCauley, and S. Franklin. in review. Automatization for Software Agents.
[59] McAllester, D. A. and D. Rosenblitt. 1991. Systematic nonlinear planning. In *Proceedings of the Ninth National Conference on Artificial Intelligence* (AAAI-91), Vol.2. 634-639. Anaheim, CA: AAAI Press.
[60] Minsky, M. 1985. *The Society of Mind*. New York: Simon and Schuster.
[61] Cox, M. T. 2005. Metacognition in computation: a selected research review. *Artificial Intelligence* 169:104-141.
[62] Zhang, Z., D. Dasgupta, and S. Franklin. 1998. Metacognition in Software Agents using Classifier Systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*. Madison, Wisconsin: MIT Press.
[63] Damasio, A. R. 1999. *The Feeling of What Happens*. New York: Harcourt Brace.
[64] Strawson, G. 1999. The self and the SESMET. In *Models of the Self*, ed. S. Gallagher, and J. Shear. Charlottesville, VA: Imprint Academic.

[65] Gallagher, S. 2000. Philosophical conceptions of the self: implications for cognitive science. *Trends in Cognitive Science* 4:14-21.
[66] Baars, B. J., T. Ramsoy, and S. Laureys. 2003. Brain, conscious experience and the observing self. *Trends Neurosci.* 26:671-675.
[67] Goldberg, I. I., M. Harel, and R. Malach. 2006. When the Brain Loses Its Self: Prefrontal Inactivation during Sensorimotor Processing. *Neuron* 50:329-339.
[68] Merker, B. 2005. The liabilities of mobility: A selection pressure for the transition to consciousness in animal evolution. *Consciousness and Cognition* 14:89-114.
[67] Johnston, V. S. 1999. *Why We Feel:The Science of Human Emotions*. Reading MA: Perseus Books.
[68] Franklin, S., and L. McCauley. 2004. Feelings and Emotions as Motivators and Learning Facilitators. In *Architectures for Modeling Emotion: Cross-Disciplinary Foundations, AAAI 2004 Spring Symposium Series*, vol. Technical Report SS-04-02. Stanford University, Palo Alto, California, USA: American Association for Artificial Intelligence; March 22-24, 2004.
[69] Franklin, S., and U. Ramamurthy. 2006. Motivations, Values and Emotions: 3 sides of the same coin. In *Proceedings of the Sixth International Workshop on Epigenetic Robotics*, vol. 128. Paris, France: Lund University Cognitive Studies.