IDA: A Cognitive Agent Architecture[1,]

Stan Franklin[2], Arpad Kelemen[2] and Lee McCauley
Institute for Intelligent Systems
The University of Memphis
Memphis TN 38152, USA

ABSTRACT

Here we describe an architecture for an intelligent distribution agent being designed for the Navy. This autonomous software agent will implement global workspace theory, a psychological theory of consciousness. As a result, it can be expected to react to novel and problematic situations in a more flexible, more human-like way than traditional AI systems. If successful, it will perform a function, namely billet assignment, heretofore reserved for humans. The architecture consists of a more abstract layer overlying a multi-agent system of small processors. The mechanisms implementing the architecture are quite varied and diverse, and are drawn mostly from the "new" AI. This paper is intended as a progress report.

INTRODUCTION

For most of its four decades of existence, artificial intelligence has devoted its attention primarily to studying and emulating individual functions of intelligence. During the last decade, researchers have expanded their efforts to include systems modeling a number of cognitive functions (Albus, 1991, 1996; Ferguson, 1995; Hayes-Roth, 1995; Jackson, 1987; Johnson and Scanlon, 1987; Laird, Newall, and Rosenbloom, 1987; Newell, 1990; Pollack, 1989; Riegler, 1997; Sloman, 1995). There's also been a movement in recent years towards producing systems situated within some environment (Akman, 1998; Brooks, 1990; Maes, 1990b). Some recent work of the first author and his colleagues have combined these two trends by experimenting with cognitive agents (Bogner, Ramamurthy, and Franklin to appear; Franklin and Graesser forthcoming; McCauley and Franklin, to appear; Song and Franklin , forthcoming; Zhang, Franklin and Dasgupta, 1998; Zhang et al, 1998). This paper briefly describes the architecture of one such agent. It's intended as a progress report.

By an autonomous agent (Franklin and Graesser 1997) we mean a system situated in, and part of, an environment, which senses that environment, and acts on it, over time, in pursuit of its own agenda. It acts in such a way as to possibly influence what it senses at a later time. That is, the agent is structurally coupled to its environment (Maturana 1975, Maturana and Varela 1980). Biological examples of autonomous agents include humans and most animals. Non-biological examples include some mobile robots, and various computational agents, including artificial life agents, software agents and computer viruses. Here we'll be concerned with autonomous software agents 'living' in real world computing systems.

Such autonomous software agents, when equipped with cognitive (interpreted broadly) features chosen from among multiple senses, perception, short and long term memory, attention, planning, reasoning, problem solving, learning, emotions, moods, attitudes, multiple drives, etc., will be called cognitive agents (Franklin 1997). Such agents promise to be more flexible, more adaptive, more human-like than any currently existing software because of their ability to learn, and to deal with novel input and unexpected situations. But, how do we design such agents?

On3 way is to model them after humans. We've chosen to design and implement such cognitive agents within the constraints of the global workspace theory of consciousness, a psychological theory that gives a high-level, abstract account of human consciousness and broadly sketches it architecture (Baars, 1988, 1997). We'll call such agents "conscious" software agents. Global workspace theory postulates that human cognition is implemented by a multitude of relatively small, special purpose processes, almost always unconscious. (It's a multiagent system.) Coalitions of such processes find their way into a global workspace (and into consciousness). This limited capacity workspace serves to broadcast the message of the coalition to all the unconscious processors, in order to recruit other processors to join in handling the current

[1] With indispensable help from the other members of the Conscious Software Research Group including Ashraf Anwar, Miles Bogner, Scott Dodson, Art Graesser, Derek Harter, Aregahegn Negatu, Uma Ramamurthy, Hongjun Song, and Zhaohua Zhang.
[2] Supported in part by ONR grant N00014-98-1-0332.

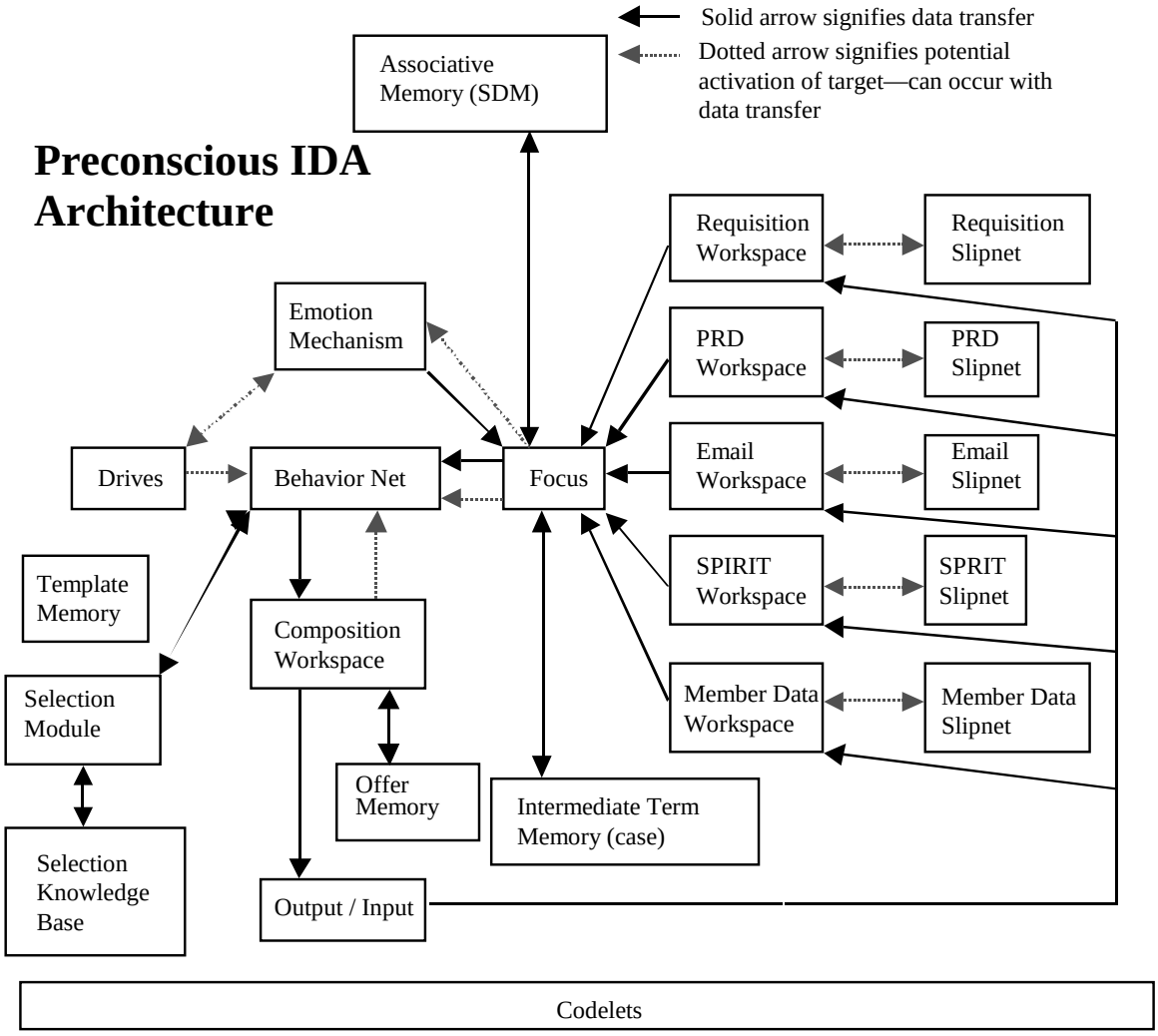

Figure 1. Preconscious IDA Architecture

novel situation, or in solving the current problem. All this takes place under the auspices of contexts: goal contexts, perceptual contexts, conceptual contexts, and/or cultural contexts. Each context is, itself, a coalition of processes. There's much more to the theory, including attention, learning, action selection, and problem solving. Conscious software agents should implement the major parts of the theory, and should always stay within its constraints.

IDA's ARCHITECTURE

IDA (Intelligent Distribution Agent), is to be such a conscious software agent developed for the Navy. At the end of each sailor's tour of duty, he or she is assigned to a new billet. This assignment process is called distribution. The Navy employs some 200 people, called detailers, full time to effect these new assignments. IDA's task is to facilitate this process, by playing the role of detailer as best she can.

Designing IDA presents both communication problems and constraint satisfaction problems. She must communicate with sailors via email and in natural language, understanding the content. She must access a number of databases, again understanding the content. She must see that the Navy's needs are satisfied, for example, the required number of sonar technicians on a destroyer with the required types of training. She must hold down moving costs. And, she must cater to the needs and desires of the sailor as well as is possible.

Here we'll briefly describe a design for IDA including a high level architecture and the mechanisms by which it's to be implemented. While the mechanisms will be referenced individually as they occur, brief accounts of each can be found in *Artificial Minds* (Franklin 1995). With the help of diagrams we'll describe a preconscious version of IDA, and then discuss the additional mechanisms needed to render her conscious.

IDA will sense her world using three different sensory modalities. She'll receive email messages, she'll read database screens and, eventually, she'll sense via operating system commands and messages. Each sensory mode will require at least one knowledge base and a workspace. The mechanism here will be based loosely on the Copycat Architecture (Hofstadter1995; Hofstadter and Mitchell 1994; Zhang et al 1998). Each knowledge base will be a slipnet, a fluid semantic net. The workspace (working memory) will allow perception (comprehension), a constructive process. See the right side of Figure 1 for five such pairs. Each, other than the email, will understand material from a particular database, for example personnel records, a list of job openings, a list of sailors to be assigned. Sensing the operating system isn't present in Preconscious IDA.

Note that each of IDA's senses is an active sense, like our vision rather than our hearing. They require actions on IDA's part before sensing can take place, for example reading email or accessing a database. IDA selects her actions by means of an enhanced version of the behavior net (Maes 1990a; Song and Franklin forthcoming). See Figure 1. The behavior net is a directed graph with behaviors as verticies and three different kinds of links along which activation spreads. Activation originates from internal, explicitly represented drives, from IDA's understanding of the external word through the Focus, and from internal states. The behavior whose activation is highest among those with all prerequisites satisfied becomes the next goal context as specified in global workspace theory. The several small actions typically needed to complete a behavior are taken by codelets, of which more later.

IDA's behaviors are partitioned into streams, the connected components of the digraph, each in the service of one or more drives. Streams of behaviors are like plans, except that they may not be linear, and might well be interrupted during their execution or possibly not completed. Examples of IDA's streams include Access EAIS, Access Personnel Record, Offer Assignments, Send Acknowledgement, Produce Orders.

IDA is very much a multi-agent system, the agents being the codelets that underlie all the higher level constructs and that ultimately perform all of IDA's actions. The term was taken from the Copycat system. Their organization and structure were inspired by pandemonium theory (Jackson 1987), though there are significant differences. We've mentioned the codelets that underlie behaviors. Others underlie slipnet nodes and perform actions necessary for constructing IDA's understanding of an email message or of a database screen (Zhang et al 1998). Still other codelets will play a vital role in consciousness, as we'll see below. Codelets come in two major varieties. The demon codelets are always active, looking for opportunities where they become relevant. Instantiated codelets are generated by demon codelets. Their variables are bound in order that they can perform a particular task. When the task is done, they disappear. The codelets are represented in Figure 1 by a long box at the bottom, since they underlie essentially everything else.

Having gathered all relevant information, IDA must somehow select which assignments she'll offer a given sailor. See the lower left of Figure 1. Being a constraint satisfaction problem, considerable knowledge will be required to make these selections. This knowledge could be in the form of a traditional, rule-based expert system, but more likely will be implemented in some form suitable for reinforcement learning. The constraint satisfaction mechanism will be housed in the selection module. The choice of mechanism for it is currently being researched. A stream of behaviors will set the selection mechanism in motion when appropriate.

IDA's emotion module (McCauley and Franklin 1998), like a human's, provides a multi-dimensional method for ascertaining how well she's doing. We'll experiment with building in mechanisms for emotions. Examples might include anxiety at not understanding a message, guilt at not responding to a sailor in a timely fashion, and annoyance at an unreasonable request from a sailor. Emotions in humans and in IDA influence all decisions as to action (Damasio 1994). IDA's action selection will be influenced by emotions via their effect on drives. Including emotional capabilities in non-biological autonomous agents is not a new idea (Bates, Loyall, and Reilly 1991, Sloman and Poli 1996, Picard 1997). Some claim that truly intelligent robots or software agents can't be effectively designed without emotions.

As a glance at Figure 1 shows, IDA has a number of different memories. The offer memory is a traditional database that keeps track of the assignments IDA has offered various sailors. The template memory is another that holds the various templates that IDA uses compose commands to access databases or issue orders, and to compose messages to sailors. IDA's intermediate term memory acts as an episodic memory, providing context for email messages and for the contents of database screens. It'll be implemented as a case-based memory to facilitate case-based learning at a later stage. IDA's associative memory does what you'd expect. It associates memories, emotions and actions with incoming percepts and with outgoing actions. It's implemented by an extension of sparse distributed memory (Kanerva 1988).

The operation of these last two, more complex, memory systems deserves more explanation. As IDA's most recent percept reaches the perception register (See Figure 2) having been constructed (comprehended)

When the most recent perception register is filled by one of the perception modules, several events occur in simulated parallel. Activation is sent to the behavior net, that is, the environment influences action selection. The associative memory is read using the percept as the cue. Since sparse distributed memory is content addressable,

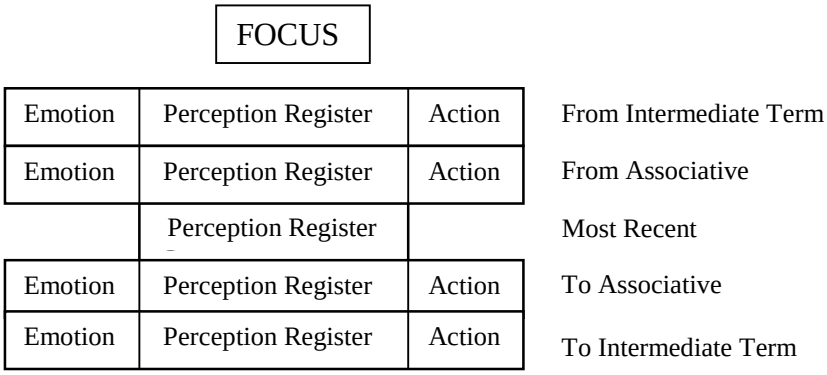

Figure 2. IDA's Focus

associations with the percept, including an emotional overtone and an action previously taken in a similar situation are typically returned into an expanded copy of the perception registers (see Figure 2). These associations also activate the behavior net and the emotion module. Associations influence action selection. At the same time intermediate term memory is read with the same cue. The most similar case is returned, again with emotion and action, into yet another copy of the expanded perception registers. In the full version, consciousness will come into play at this point. Now, an action and an emotion are selected into the two remaining copies of the expanded perception registers along with the current percept. Each is then written to its appropriate memory. IDA has processed a single percept.

A similar, but simpler process takes place with IDA's actions. Recall that IDA can consult databases, compose and send email and orders, and later, try to protect herself from system crashes. An action is taken as a result of a behavior being activated, or perhaps a stream of behaviors. Often the last behavior in such a stream causes the action to be placed, with other information, in the writing registers of the focus. See Figure 2. This results in the action being written to both associative and intermediate-term memory. Thus the action will be available to help set a context for future percepts, and for learning as will be discussed below.

Our brief description of the preconscious form of IDA is as complete as it's going to be in this short paper. She could well be implemented as described, and should be expected to work reasonably well. She would not, however, show the kind of flexibility and more human-like behavior in the face of novel or problematic situations that was claimed in the third paragraph of this paper. To accomplish this, and to implement global workspace theory, will require a fair amount more machinery.

Global workspace theory postulates the contents of consciousness to be coalitions of codelets shined on by a spotlight. Imagine a codelet workspace populated by many active codelets each working, unconsciously and in parallel, on its own agenda. The spotlight seeks out coalitions of codelets that arise from novel or problematic situations. When the spotlight picks out some coalition of codelets, the information contained therein is broadcast to all the codelets, active or not. The idea is to recruit resources, that is, relevant codelets to help in dealing with the situation. It seems that in humans almost any resource may be relevant depending on the situation. The global workspace method attacks the problem of finding the relevant resources by brute force. Broadcast to them all. IDA will use this method. To do so, she'll need a coalition manager, a spotlight controller, and a broadcast manager (Bogner, Ramamurthy, and Franklin to appear).

Metacognition includes knowledge of one's own knowledge and cognitive processes, and the ability to actively monitor and consciously regulate them. Metacognition is important for humans since it guides people to select, evaluate, revise, and abandon cognitive tasks, goals, and strategies (Hacker 1997). If we want to build more human-like software agents, we'll need to build metacognition into them.

Following Minsky's terminology (1985) let's partition IDA's "brain" into two parts, the A-brain and the B-brain. The A-brain performs all cognitive activities. Its environment is the outside world, a dynamic, but limited, real world environment. The B-brain, sitting on top of the A-brain, monitors and regulates the A-brain. The B-brain performs all metacognitive activities; its environment is the A-brain's activities. IDA's metacognition module will be implemented using a classifier system (Holland 1986) in order that it may learn.

Metacognition isn't the only IDA module that can learn. Codelets learn a la pandemonium theory by forming associations (Jackson 1987). Two codelets that share time in the consciousness spotlight either create or strengthen an association between them. The strength of this association can effect the likelihood of their being together in a coalition at a later time. These associations can eventually spark the learning of "concept" codelets,

coalitions of codelets that are chunked together into a higher level codelet.

Yet another form of learning is provided by IDA's associative memory. Its sparse distributive memory mechanism learns associations as a side effect of its structure.

IDA's intermediate term memory uses case-based memory in order that more sophisticated concepts and behaviors can be learned via case-based learning (Kolodner 1993; Bogner, Ramamurthy, and Franklin to appear). This kind of learning takes place as a result of interactions with sailors or with a human detailer to which IDA is apprenticed. New concepts learned in this way appear as new nodes in a slipnet, while new behaviors appear as a part of learned streams in the behavior net. In each case, both new links and new underlying codelets must also be learned. Learned concepts, behaviors, links and codelets are all modeled after existing concepts, behaviors, links and codelets respectively. The current situation is compared to that of the most similar case retrieved from intermediate term memory, and the concepts, behaviors, links and codelets modified so as to accommodate the differences between the two cases. Cases resulting in learned concepts and behaviors typically will result from human interactions. Research on this learning strategy is ongoing.

Consciousness will play a critical role in all theses different modes of learning. Associations between codelets are learned or strengthened only as a result of shared time in consciousness. The contents of the focus are written to associative and intermediate term memory only after having come to consciousness. And, dialogue with humans from which learning occurs is only initiated as a result of conscious recognition of an unknown word, or an unfamiliar situation. Learning in the selection knowledge base and metacoginitive learning will also result from consciousness. All of IDA's learning takes place as a result of her consciousness apparatus.

FURTHER WORK

Designing IDA's architecture is only the bare beginnings of a complete implementation. A tremendous amount of knowledge acquisition and representation must be accomplished. Let's quickly outline what's needed.

The various slipnets on the perception side must be created. They'll be of two different kinds, database and email. A database slipnet must know about each possible value of each field. In order of magnitude estimates, each of the dozen database records will contain a dozen fields each with a dozen values. Large and relatively complex slipnets will be needed. Information gathering for these slipnets has begun.

An even larger and more complex slipnet will be required to help understand email messages from sailors in natural language. The range of topics that can appear, while bounded, is quite varied. They will include issues such a geography, sea or shore duty, job description, training, rating, level of responsibility, housing, education for children, vocational opportunities for spouses, etc. Still another complex slipnet will know about interpreting messages from a human detailer to whom IDA is apprenticed. These messages, again in natural language, will be critical to her learning.

On the behavior side, drives must be determined, and streams of behavior designed to bring about the needed actions. One behavior stream will be needed to consult each database, another to acknowledge messages, another to compose and send messages to a sailor, a different one for messages to a detailer, one to write orders, etc. The behavior net will also be large and complex. Work on it has begun.

The selection knowledge base will require knowledge engineering with a human detailer, as well as being added to by learning during an apprenticeship, yet another major task.

How is such a daunting task justified? From the author's point of view, the IDA project is a proof-of-concept project for conscious software. We expect it to lead us to further knowledge of both human and agent cognition. We also expect it to show that conscious software can perform tasks heretofore reserved only for humans. From the Navy's point of view, the two hundred odd human detailers cost something like $20,000,000 per year. There's lot's of room for the kind of savings IDA promises.

REFERENCES

Akman V. (1998) (guest ed.) *Minds and Machines* (Special Issue: Situations and AI.

Albus, J.S. (1991), "Outline for a Theory of Intelligence," *IEEE Transactions on Systems, Man and Cybernetics,* Vol. 21, No. 3, May/June.

Albus, J.S. (1996). "The Engineering of Mind," *Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior: From Animals to Animats 4,* Cape Cod, MA, September.

Baars, B. J. (1988). *A Cognitive Theory of Consciousness.* Cambridge: Cambridge University Press.

Baars, B. J. (1997). *In the Theater of Consciousness.* Oxford: Oxford University Press.

Bates, Joseph, A. Bryan Loyall, and W. Scott Reilly (1991). "Broad Agents," *Proceedings of the AAAI Spring Symposium on Integrated Intelligent Architectures,* Stanford University, March. These proceedings are available in *SIGART Bulletin,* Volume 2, Number 4, August 1992.

Bogner, Myles, Uma Ramamurthy, and Stan Franklin (to appear), **"**Modeling Global Workspace Perception In A Socially Situated Agent," in Kerstin Dautenhahn ed. *Human Cognition and Social Agent Technology*

Brooks, Rodney A. (1990), "Elephants Don't Play Chess," In Pattie Maes, ed., *Designing Autonomous Agents*, Cambridge, MA: MIT Press

Damasio, A. R. (1994), *Descartes' Error*, New York: Gosset/Putnam Press.

Ferguson, I. A. (1995). "On the role of DBI modeling for integrated control and coordinated behavior in autonomous agents." *Applied Artificial Intelligence*, 9(4).

Franklin, Stan (1994), *Artificial Minds*, Cambridge MA: MIT Press.

Franklin, Stan (1997). "Autonomous Agents as Embodied AI," *Cybernetics and Systems* Special issue on Epistemological Aspects of Embodied AI, 28:6 499-520.

Franklin, Stan and Graesser, Art (1997) "Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents," *Intelligent Agents III*, Berlin: Springer Verlag, 21-35,

Franklin, Stan and Graesser, Art (forthcoming), "Models of Consciousness"

Hacker, Douglas, (1997), "Metacognitive: Definitions and Empirical Founditions," In Hacker, D., Dunlosky, J., Graesser A. (Eds.) *Metacogniton* in *Educational Theory and Practice.* Hillsdale, NJ: Erlbaum.

Hayes-Roth, B. (1995). "An architecture for adaptive intelligent systems." *Artificial Intelligence*, 72 329-365.

Hofstadter, D. R. (1995), *Fluid Concepts and Creative Analogies*, Basic Books.

Hofstadter, D. R. and Mitchell, M. (1994), "The Copycat Project: A model of mental fluidity and analogy-making." In Holyoak, K.J. & Barnden, J.A. (Eds.) *Advances in connectionist and neural computation theory*, Vol. 2: Analogical connections. Norwood, N.J.: Ablex.

Holland, J. H. (1986), "A Mathematical Framework for Studying Learning in Classifier Systems." In D., Farmer et al, *Evolution, Games and Learning: Models for Adaption in Machine and Nature*. Amsterdam: North-Holland

Jackson, John V. (1987), "Idea for a Mind," *SIGGART Newsletter*, no. 181, July, 23-26.

Johnson, M. and Scanlon, R. (1987) "Experience with a Feeling-Thinking Machine", *Proceedings of the IEEE First International Conference on Neural Networks*, San Diego.71-77.

Kanerva, Pentti (1988), *Sparse Distributed Memory*, Cambridge, MA: The MIT Press.

Kolodner, Janet (1993), *Case-Based Reasoning* , Morgan Kaufman

Laird, John E., Newall, Allen, and Rosenbloom, Paul S. (1987). "SOAR: An Architecture for General Intelligence." *Artificial Intelligence*, 33: 1-64.

Maes, Pattie (1990a), 'How to do the right thing', Connection Science, 1:3.

Maes, Pattie (1990b) ed., *Designing Autonomous Agents, Cambridge*, MA: MIT Press

Maturana, H. R. (1975). "The Organization of the Living: A Theory of the Living Organization. *"International Journal of Man-Machine Studies*, 7:313-32.

Maturana, H. R. and Varela, F. (1980). *Autopoiesis and Cognition: The Realization of the Living.* Dordrecht, Netherlands: Reidel.

McCauley, Thomas L. and Stan Franklin (to appear) "An Architecture for Emotion," ????

Minsky, Marvin (1985), *Society of Mind*, New York: Simon and Schuster.

Newell, Allen (1990), *Unified Theories of Cognition*, Cambridge, Mass: Harvard University Press.

Picard, Rosalind (1997), *Affective Computing*, Cambridge MA: The MIT Press.

Pollack, John (1989), *How to Build a Person : A Prolegomenon*, Cambridge MA: MIT Press.

Riegler, A. (1997) "Ein kybernetisch-konstruktivistisches Modell der Kognition." In: Müller, A, Müller, K. H. & Stadler, F. (eds.) *Konstruktivismus und Kognitionswissenschaft. Kulturelle Wurzeln und Ergebnisse.* Wien, New York: Springer.

Sloman, Aaron (1995). "Exploring Design Space and Niche Space." *Proceedings 5th Scandinavian Conf on AI*, Trondheim May 1995, Amsterdam: IOS Press.

Sloman, Aaron and Poli, Riccardo (1996). "SIM_AGENT: A toolkit for exploring agent designs in Intelligent Agents," Vol. II *ATAL-95*, Eds. Mike Wooldridge, Joerg Mueller, Milind Tambe, Springer-Verlag, pp. 392--407.

Song, Hongjun and Stan Franklin (forthcoming), "Action Selection Using Behavior Instantiation"

Zhang, Zhaohua, Stan Franklin and Dipankar Dasgupta (1998), "Metacognition in Software Agents using Classifer Systems," *Proc AAAI 98*,

Zhang, Zhaohua, Stan Franklin, Brent Olde, Yun Wan and Art Graesser (1998) "Natural Language Sensing for Autonomous Agents," *Proc. IEEE Joint Symposia on Intelligence and Systems*, Rockville, Maryland, 374-81.