# IDA, a Conscious Artifact?

Stan Franklin
Institute for Intelligent Systems
The University of Memphis

## Abstract

After discussing various types of consciousness, several approaches to machine consciousness, software agent, and global workspace theory, we describe a software agent, IDA, that is "conscious" in the sense of implementing that theory of consciousness. IDA perceives, remembers, deliberates, negotiates, and selects actions, sometimes "consciously." She uses a variety of mechanisms, each of which is briefly described. It's tempting to think of her as a conscious artifact. Is such a view in any way justified? The remainder of the paper considers this question.

## Consciousness

The term "consciousness" is used with so many different meanings and connotations that at least one philosopher/computer scientist is reluctant to use the term at all for fear of equivocation (Aaron Sloman, personal communication). In this section we'll explore some of those meanings in the hope of making clear the usage of the term in the following sections. In ordinary, day to day language the various meanings are distinguished by context, more or less successfully.

Perhaps the most common usage of the term refers to *phenomenal consciousness*, sometimes called sentience, or subjective (first person) experience. Like trying to explain the color purple to a person who has been sightless from birth, phenomenal consciousness is impossible to convey in words, but must be experienced to be understood. The term "conscious" as used in the title is intended to convey phenomenal consciousness.

But phenomenal consciousness comes in several varieties, which are experienced differently. There are the altered states of consciousness produced by various hallucinogenic drugs. There is the phenomenal consciousness of dreams. There are the in between states sometimes experienced while going to sleep or when waking up. There's the phenomenal consciousness of an infant as yet mostly uninfluenced by language. There's the phenomenally conscious state of mental quiet achieved by meditators who can turn of their inner voices and experience the world without the usual running verbal commentary. There is the phenomenal consciousness of my cat who certainly experiences things differently from me if only by virtue of the differences in our visual systems. There's the phenomenal consciousness of a snake that is incapable of fusing information derived from its various senses. And what about the phenomenal consciousness of a frog, likely so primitive? Does a mosquito have any at all? If not how does it seek heat and avoid certain smells? What about an ant following a pheromone trail? Many people want to attribute phenomenal consciousness only to humans, others to those organisms possessing. neuropil (Freeman 1995), and yet others only to biological organisms. A recent paper argues for phenomenal consciousness in all mammals (Baars 2001).

Consciousness has many functions. It helps us deal with novel or problematic situations for which we have no automatized response. It makes us aware of potentially dangerous situations. It alerts us to opportunities presented by the environment. It allows us to perform tasks that require knowledge of location, shape, size or other features of objects in our environments. And there are a number of other functions of consciousness. I say that an agent (see below) possesses *functional consciousness* if its architecture and mechanisms allow it a number of these and, perhaps, other functions. (Functional consciousness is an example of a cluster concept, that is one that can be ascribed to something on the basis of it satisfying some sufficient but unspecified subset of a given collection of features (Sloman 1995). Surprisingly such concepts, though not well defined, turn out to be useful.) The "consciousness" attributed in this article to IDA and other software agents is meant to be functional consciousness, and the quotes are meant to so indicate. Consciousness attributed to humans must be distinguished as to type by the context.

Can an agent (organism or otherwise) have functional consciousness without phenomenal consciousness? This question has vexed philosophers for decades (Chalmers 1996). They refer to a humanoid such agent as a "zombie." Are there zombies? Could there be? We'll return to a version of this question in the next section.

Another type of consciousness that's often spoken of is self-consciousness, that is being aware of oneself. This often includes one's self image. Presumably, self-consciousness presupposes phenomenal consciousness. Damasio has recently described self-consciousness as beginning with the proto-self, that is, unconscious knowledge about the current state of the body (Damasio 1999). The question of self-consciousness in primates, particular in the other apes has been studied using a

"does he recognize himself in a mirror test" (Gallup 1982). The latest results seem to indicate that all apes are self-conscious, and that no other primate is.

There's recently been a great deal of interest in possible ways of ascertaining or measuring phenomenal consciousness. One means that has been suggested is reportability (Baars 1988). Being able to report, verbally or otherwise, the contents of one's awareness is often given as evidence of phenomenal consciousness, particularly when the reports can be confirmed by an experimenter. We'll discuss reportability again below.

Another approach is via neural correlates of consciousness as measured by PET or fMRI scans ????. If it were found that activity in a particular neural structure correlated well with phenomenal consciousness, the extent of such activity could be used as a measure independent of reportability. To my knowledge such correlates have not been found. To the contrary, several researchers hypothesize that activity over large areas, particularly cortical and thalamic, is required for consciousness, both phenomenal and functional (Freeman 1999, Edelman & Tononi 2000, Baars 2002).

## Machine consciousness

What about the possibility of machine consciousness? Is it possible for a robot or a software agent (see below) to be functionally conscious? Phenomenally conscious? The answer to one question seems clearly yes. I will argue below that IDA is functionally conscious. The question of machine phenomenal consciousness seems much more difficult. The philosophers have engaged in considerable debate over this issue (Chalmers 1996). In May of 2000 a workshop attended by two dozen leading researchers, psychologists, neuroscientists, philosophers, and AI scientists, were unable to make any noticeable headway with the problem.

There are several serious, ongoing projects aimed a producing machine consciousness. One such, headed by Igor Aleksander, is relatively far along, having produced a working system, MAGNUS, using neural modeling that's arguably capable of imagination (2000). A second such is my own IDA project to be described in some detail below. IDA is currently up and running, and exhibiting functional consciousness. A third such project, conceived by Rodney Cotterill and also based on neural modeling, aims at developing machine consciousness in a manner analogous to the way a human child develops An early version is being demonstrated (Cotterill 2001). For a fourth, Owen Holland and Rodney Goodman have embarked on a bottom up approach of add additional capabilities to a robotic system until it shows signs of consciousness REF. A fifth, due to Lee McCauley, builds consciousness into a neural schema system REF.

A major question confronting all the builders of all theses systems is how to determine if, or when, their system become phenomenally conscious. Though there seems to be no good answer, we'll return to this critical question later.

## Global Workspace Theory

The material in this section is from Baars' two books (1988, 1997) (1988, 1997) and superficially describes his global workspace theory of consciousness.

In his global workspace theory, Baars, along with many others (e.g. (Minsky 1985, Ornstein 1986, Edelman 1987)) , postulates that human cognition is implemented by a multitude of relatively small, special purpose processes, almost always unconscious. (It's a multiagent system.) Communication between them is rare and over a narrow bandwidth. 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 novel situation, or in solving the current problem. Thus consciousness in this theory allows us to deal with novel or problematic situations that can't be dealt with efficiently, or at all, by habituated unconscious processes. In particular, it provides access to appropriately useful resources, thereby solving the relevance problem.

This theory offers an explanation for consciousness being serial in nature rather than parallel as is common in the rest of the nervous system. Messages broadcast in parallel would tend to overwrite one another making understanding difficult. It similarly explains the limited capacity of consciousness as opposed to the huge capacity typical of long-term memory and other parts of the nervous system. Large messages would be overwhelming to small, special-purpose processors.

All this activity of processors takes place under the auspices of contexts (see Figure 1): goal contexts, perceptual contexts, conceptual contexts, and/or cultural contexts. Baars uses goal hierarchies, dominant goal contexts, a dominant goal hierarchy, dominant context hierarchies, and lower level context hierarchies. Each context is, itself, a coalition of processes. Though contexts are typically unconscious, they strongly influence conscious processes. A key insight of global workspace says that each context is, in fact, a coalition of processors.

Long back, William James proposed the ideomotor theory of voluntary action (James 1890). James suggests that any idea (internal proposal) for an action that come to mind (to consciousness) is acted upon unless it provokes some opposing idea or some counter proposal. He speaks at length of the case of deciding to get out of a warm bed into an unheated room in the dead of winter. "This case seems to me to contain in miniature form the data for an entire psychology of volition." Global workspace theory adopts James' ideomotor theory as is, and provides a functional architecture for it (Baars 1997 , Chapter 6).

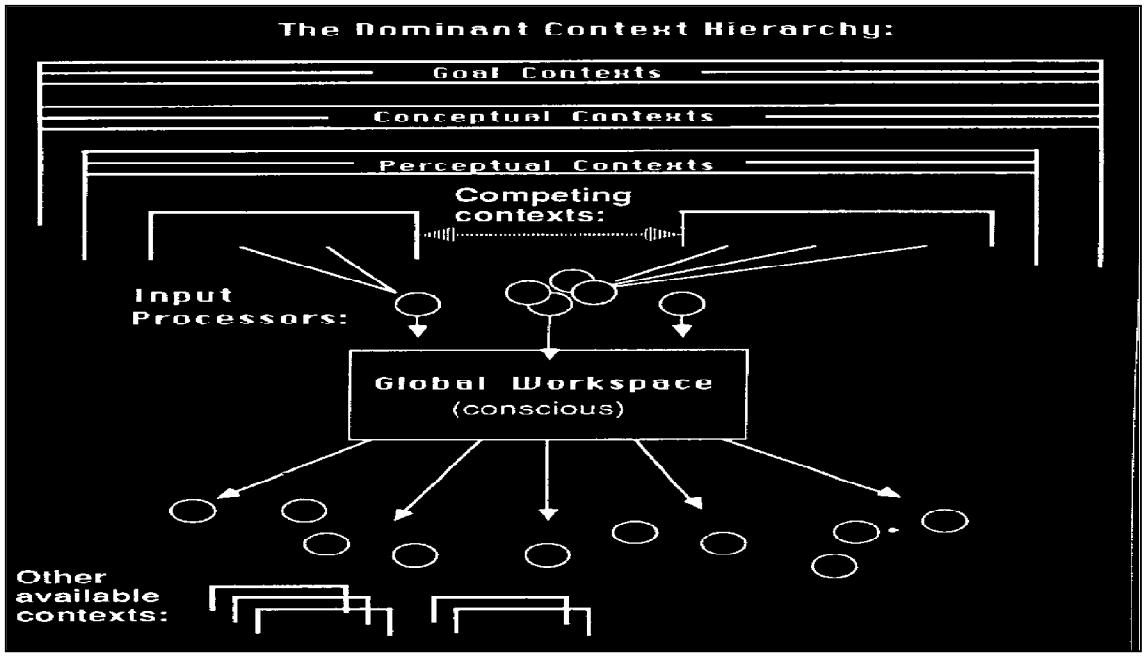



Figure 1. Global Workspace Theory

Baars postulates that learning results simply from conscious attention, that is, that consciousness is sufficient for learning. There's much more to global workspace theory, including attention, action selection, emotion, voluntary action, metacognition and a sense of self. I think of it as a high level theory of cognition.

## Software agents

Artificial intelligence pursues the twin goals of understanding human intelligence and of producing intelligent software and / or artifacts. Designing, implementing and experimenting with autonomous agents furthers both these goals in a synergistic way. An *autonomous agent* (Franklin & Graesser 1997) is 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. In biological agents, this agenda arises from evolved in drives and their associated goals; in artificial agents from drives and goals built in by its creator. Such drives, which act as motive generators (Sloman 1987) must be present, whether explicitly represented, or expressed causally. The agent also acts in such a way as to possibly influence what it senses at a later time. In other words, it is structurally coupled to its environment (Maturana 1975, Maturana et al. 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 many computer viruses. We'll be concerned with autonomous software agents, designed for specific tasks, and 'living' in real world computing systems such as operating systems, databases, or networks.

A *"conscious" software agent* is one that implements global workspace theory. The scare quotes are included to remind the reader that it's functional consciousness that's being claimed, not phenomenal consciousness. In addition to modeling this theory (Franklin & Graesser 1999), such "conscious" software agents should be capable of more adaptive, more human-like operations, including being capable of creative problem solving in the face of novel and unexpected situations.

## Quick overview of the IDA model

IDA is one recent instantiation of our theory of intelligent software agents. IDA (Intelligent Distribution Agent) is a "conscious" software agent that was developed for the US Navy (Franklin et al. 1998). At the end of each sailor's tour of duty, the sailor is assigned to a new billet. This assignment process is called distribution. The Navy employs some 280 people, called detailers, to effect these new assignments. IDA's task is to facilitate this process by completely automating the role of detailer. IDA must communicate with sailors via email in natural language, by understanding the content and

producing life-like responses. Sometimes she will initiate conversations. She must access several databases, again understanding the content. She must see that the Navy's needs are satisfied by adhering to some ninety policies and seeing that job requirements are fulfilled. She must hold down moving costs, but also cater to the needs and desires of the sailor as well as is possible. This includes negotiating with the sailor via an email correspondence in natural language. Finally, she must write the orders and start them on the way to the sailor. At this writing an almost complete version of IDA is up and running and had been demonstrated to the satisfaction of the Navy.

In this article we will discuss IDA both as a conceptual model and as a computational model of global workspace theory. The conceptual model includes modules that have been designed, including mechanisms, but not implemented. Some unimplemented modules are important to our cognitive modeling but not to the Navy's application. Others, also not critical to the application, have been designed since the demonstration. In the subsections that follow, the modules of the conceptual model will be briefly described with references given to full descriptions. Unimplemented modules will be so indicated, as will the concordance between concepts in our model and those in global workspace theory. For science, the conceptual model us useful as a source of, hopefully testable, hypotheses (Franklin 1997). Each design decision taken gives rise to the hypothesis that humans "do it" that same way. The conceptual model, with its specified mechanisms is sufficient for this process of hypothesizing. Implementation is fine, but not required.

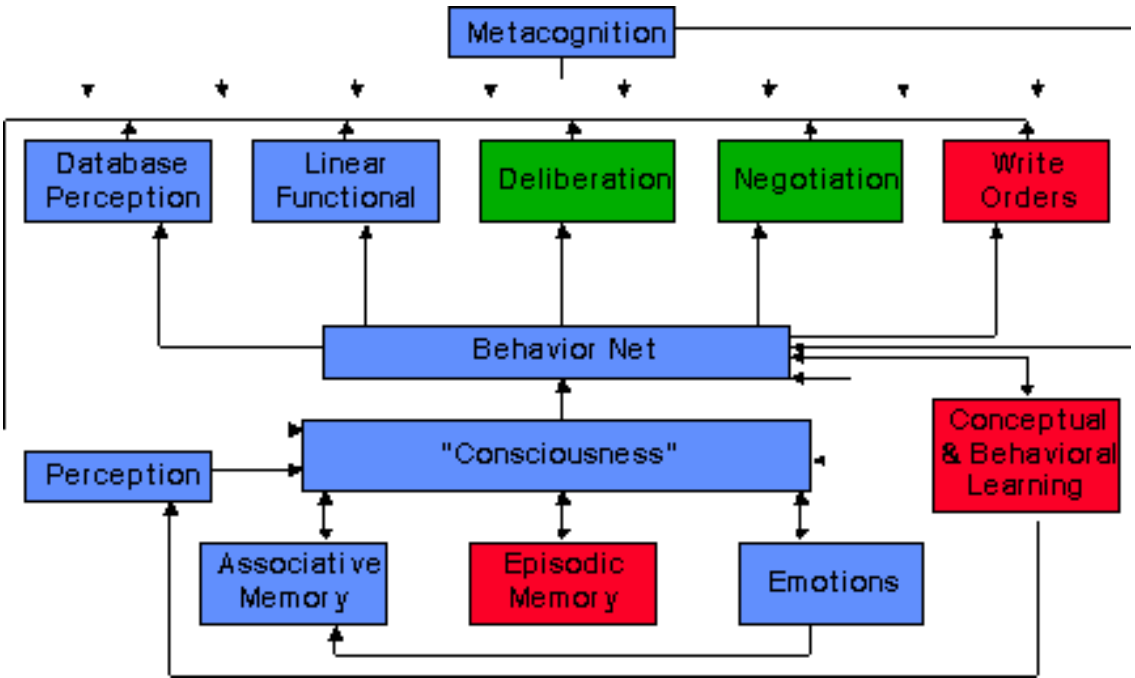
Developed during 1996-2001, the IDA computational model is the product of the "Conscious" Software Research Group at the University of Memphis (http://csrg.cs.memphis.edu/CSRG/). More than twenty-five researchers contributed to its development. Their names can be found on the web site. (See the Acknowledgement section below for an abbreviated list.) Built on a Java platform, the running IDA consists of approximately a quarter-million lines of code, and almost completely consumes the resources of a 2001 high-end workstations. Her various modules are briefly described below.

Following Hofstadter's terminology (see below) a codelet is a special purpose, relatively independent, mini-agent typically implemented as a small piece of code running as a separate thread. IDA depends heavily on such codelets for in almost every module. This is as it should be since codelets in our model correspond with processors in global workspace theory, which are postulated to form the basis of human cognition. In what follows we will encounter several different types of codelets such as perceptual codelets, attention codelets, information codelets, behavior codelets and language generation codelets. Many codelets play the role of demons (as in an operating system) waiting patiently for the conditions under which they can act. Some codelets subserve some higher-level construct, while others act completely independent.

The following sections contain brief description of the various modules that make up the IDA architecture. The names of several of these modules, perception, emotion, deliberation, metacognition, consciousness, etc, are borrowed from folk psychology. This choice of terminology has the advantage of calling to mind the functionality of the various modules. But, there are tradeoffs. Sometimes the meanings are different in more or less subtle ways. For instance, perception to the psychologists includes conscious awareness, while IDA's perception module is preconscious. Also, IDA's consciousness module conforms functionally to Global Workspace theory, but whether it includes phenomenal consciousness is precisely the major question to be considered.

***Perception***

Perception in IDA consists mostly of processing incoming email messages in natural language (Zhang et al. 1998b). In sufficiently narrow domains, natural language understanding may be achieved via an analysis of surface features without the use of a traditional symbolic parser (Jurafsky & Martin 2000), Allen describes this approach to natural language understanding as complex, template-based matching (1995). Ida's relatively limited domain requires her to deal with only a



few dozen or so distinct message types, each with relatively predictable content. This allows for surface level natural language processing. : We hypothesize that much of human language understanding results from a combined bottom up/top down passing of activation through a hierarchical conceptual net with the most abstract concepts in the middle. Thus

**Figure 2. The IDA Architecture**

IDA's language-processing module has been implemented as a Copycat-like architecture with perceptual codelets that are triggered by surface features and a slipnet (Hofstadter & Mitchell 1994), a semantic net that passes activation. The slipnet stores domain knowledge. In addition there's a pool of perceptual codelets (processors) specialized for recognizing particular pieces of text, and production templates used by codelets for building and verifying understanding. Together they constitute an integrated sensing system for IDA, allowing her to recognize, categorize and understand.

For the purposes of this chapter it's important to be clear about what is claimed by the work "understand" as used in the previous sentence. An example may help. A secretary sending out an email announcement of an upcoming seminar on Compact Operators on Banach Spaces can be said to have understood the organizer's request that she do so even though she has no idea of what a Banach pace is much less what compact operators on them are. In most cases it would likely require person years of diligent effort to impart such knowledge. Nonetheless, the secretary understands the request at a level sufficient for her to get out the announcement. In the same way IDA understands incoming email messages well enough to do all the things she needs to with them. An expanded form of this argument can be found in my Artificial Minds (1995). Glenberg also makes a similar argument (1997).

IDA must also perceive the contents read from databases, a much easier task. An underlying assumption motivates our design decisions about perception.

Suppose, for example, that IDA receives a message from a sailor saying that his projected rotation date (PRD) is approaching and asking that a job be found for him. The perception module would recognize the sailor's name and social security number, and that the message is of the please-find-job type. This information would then be written to the workspace. The general principle here is that the contents of perception are written to working memory before becoming conscious.

***Workspace***

IDA solves routine problems with novel content. This novel content goes into her workspace, which is roughly plays the same role as human working memory. IDA's workspace is less dynamic, consisting of registers set aside for particular categories of information. Perceptual codelets write to the workspace as do other, more internal codelets. Quite a number of codelets, including attention codelet (see below) watch what's written in the workspace in order to react to it. IDA's workspace also contains registers dedicated to providing an interface with her long-term associative memory. (Modeling working memory as distinct from long-term memory puts us at odds with the best current thinking of cognitive scientists. We're working on the problem.) Items in the workspace decay over time, and may be overwritten. Not all of the contents of the workspace eventually make their way into consciousness.

***Associative memory***

IDA employs sparse distributed memory (SDM) as her major associative memory (Kanerva 1988, Anwar et al. 1999, Anwar & Franklin to appear). SDM is a content addressable memory that, in many ways, is an ideal computational mechanism for use as a long-term associative memory (LTM). Any item written to the workspace cues a retrieval from LTM, returning prior activity associated with the current entry. In our example, LTM will be accessed as soon as the message information reaches the workspace, and the retrieved associations will be also written to the workspace.

At a given moment IDA's workspace may contain, ready for use, a current entry from perception or elsewhere, prior entries in various states of decay, and associations instigated by the current entry, i.e. activated elements of LTM.. IDA's workspace thus consists of both short-term working memory (STM) and something very similar to the long-term working memory (LT-WM) of Ericsson and Kintsch (1995).

Part, but not all, the workspace, called the *focus*[1], by Kanerva (1988)) is set aside as an interface with long-term LTM. Retrievals from LTM are made with cues taken from the focus and the resulting associations are written to other registers in the focus. The contents of still other registers in the focus are stored in (written to) associative memory as we will see below.

***Emotions***

IDA's design includes mechanisms for emotions (McCauley & Franklin 1998). She may "experience" such emotions as guilt at not getting a sailor's orders out on time, frustration at not understanding a message, and anxiety at not being able to

[1] Not to be confused with focus as in focus of attention, an entirely different concept.

convince a sailor to accept a suitable job. Action selection will be influenced by emotions via their effect on drives (see below), modeling work on human action selection (Damasio 1994). Emotions also influence attention, the means of bringing certain content to consciousness. They also influence the strength with which content is written to LTM. The emotion module has not been implemented in IDA, though it has in earlier "conscious" software agents (McCauley, L. et al. 2000).

***Consciousness mechanism***

The apparatus for "consciousness" consists of a coalition manager, a spotlight controller, a broadcast manager, and a collection of attention codelets whose job it is to bring appropriate contents to "consciousness" (Bogner et al. 2000). Each attention codelet keeps a watchful eye out for some particular occurrence that might call for "conscious" intervention. In most cases the attention codelet is watching the workspace, which will likely contain both perceptual information and data created internally, the products of "thoughts." Upon encountering such a situation, the appropriate attention codelet will form a coalition with the small number of information codelets that carry the information describing the situation. In the example above of our message, these codelets would carry the sailor's name, his or her social security number, and the message type. This association should lead to the collection of this small number of information codelets, together with the attention codelet that collected them, becoming a coalition. Codelets also have activations. The attention codelet increases its activation in order that the coalition, if one is formed, might compete for the spotlight of "consciousness". Upon winning the competition, the contents of the coalition is then broadcast to all codelets. Continuing our example, an attention codelet will note the please-find-job message type, gather information codelets carrying name, social security number and message type, be formed into a coalition, and will compete for "consciousness." If or when successful, its contents will be broadcast. Broadcast contents are also stored in (written to) associative memory as the contents of "consciousness" should be. (This is a correction to earlier descriptions of the model.)

***Action selection (decision making)***

IDA depends on a behavior net (Maes 1989, Negatu & Franklin 2002) for high-level action selection in the service of built-in drives. She has several distinct drives operating in parallel. These drives vary in urgency as time passes and her environment changes. Behaviors are typically mid-level actions, many depending on several behavior codelets for their execution. A behavior net is composed of behaviors, corresponding to goal contexts in GW theory, and their various links. A behavior looks very much like a production rule, having preconditions as well as additions and deletions. It's typically at a higher level of abstraction often requiring the efforts of several codelets to effect its action. A behavior can be thought of as the collection of its codelets (processors) in accordance with global workspace theory. Each behavior occupies a node in a digraph. The three types of links, successor, predecessor and conflictor, of the digraph are completely determined by the pre- and post-condition of its behaviors (Maes 1989).

As in connectionist models (McClelland et al. 1986), this digraph spreads activation. The activation comes from that stored in the behaviors themselves, from the environment, from drives, and from internal states. The more relevant a behavior is to the current situation, the more activation it is going to receive from the environment. Each drive awards activation to those behavior that will satisfy it. Certain internal states of the agent can also send activation to the behavior net. One example might be activation from a coalition of codelets responding to a "conscious" broadcast. Activation spreads from behavior to behavior along both excitatory and inhibitory links and a behavior is chosen to execute based on activation. Her behavior net produces flexible, tunable action selection for IDA. As is widely recognized in humans the hierarchy of goal contexts is fueled at the top by drives, that is, by primitive motivators, and at the bottom by input from the environment, both external and internal.

Returning to our example, the broadcast is received by appropriate behavior codelets who know to instantiate a behavior stream in the behavior net for reading the sailor's personnel record. They also bind appropriate variables with name and social security number, and send activation to a behavior that knows how to initiate the accessing of the database. If or when that behavior is chosen to be executed, behavior codelets associated with it begin to read data from the sailor's file. This data is written to the workspace. Each such write results in another round of associations, the triggering of an attention codelets, the resulting information coming to "consciousness," additional binding of variables and passing of activation and the execution of the next behavior. As long as it's the most important activity going, this process is continued until all the relevant personnel data is written to the workspace. In a similar fashion, repeated runs through "consciousness" and the behavior net result in a coarse selection of possible suitable jobs being made from the job requisition database.

The process just described leads us to speculate that in humans, like in IDA, processors (neuronal groups) bring perceptions and thoughts to consciousness. Other processors, aware of the contents of consciousness, instantiate an appropriate goal context hierarchy, which in turn, motivates yet other processors to perform internal or external actions.

***Constraint satisfaction***

IDA is provided with a constraint satisfaction module designed around a linear functional. It provides a numerical measure of the suitability, or fitness, of a specific job for a given sailor (Kelemen et al. 2002 in press). For each preference (say for San Diego) or issue (say moving costs) or policy (say sea duty following shore duty) there's a function that measures suitability in that respect. Coefficients indicate the relative importance assigned to each preference, issue or policy. The weighted sum measures the job's fitness for this sailor at this time. The same process, beginning with an attention codelet and ending with behavior codelets, brings each function value to "consciousness" and writes the next into the workspace. At last, the job's fitness value is written to the workspace.

***Deliberation***

Since IDA's domain is fairly complex, she requires *deliberation* in the sense of creating possible scenarios, partial plans of actions, and choosing between them (Sloman 1999). In our example, IDA now has a list of a number of possible jobs in her workspace, together with their fitness values. She must construct a temporal scenario for at least a few of these possible billets to see if the timing will work out (say if the sailor can be aboard ship before the departure date). In each scenario the sailor leaves his or her current post during a approved time interval, spends a specified length of time on leave, possibly reports to a training facility on a certain date, uses travel time, and arrives at the new billet with in a given time frame. Such scenarios are valued on how well they fit the temporal constraints (the gap) and on moving and training costs. These scenarios are composed of scenes organized around events, and are constructed in the workspace by the same attention codelet to "consciousness" to behavior net to behavior codelets as described previously.

***Voluntary action***

We humans most often select actions subconsciously, that is, without conscious thought. But we also make voluntary choices of action, often as a result of the kind of deliberation described above. Baars argues that such voluntary choice is the same a conscious choice (1997, p. 131) We must carefully distinguish between being conscious of the results of an action and consciously deciding to take that action, that is, of consciously deliberating on the decision. It's the latter case that constitutes voluntary action. William James proposed the *ideomotor theory* of voluntary action (James 1890). James suggests that any idea (internal proposal) for an action that comes to mind (to consciousness) is acted upon unless it provokes some opposing idea or some counter proposal. GW theory adopts James' ideomotor theory as is (Baars 1988, Chapter 7) and provides a functional architecture for it. The IDA model furnishes an underlying mechanism that implements that theory of volition and its architecture in a software agent.

Suppose that in our example at least one scenario has been successfully constructed in the workspace. The players in this decision making process include several proposing attention codelets and a timekeeper codelet. A proposing attention codelet's task is to propose that a certain job be offered to the sailor. Choosing a job to propose on the basis of the codelet's particular pattern of preferences, it brings information about itself and the proposed job to "consciousness" so that the timekeeper codelet and others can know of it. Its preference pattern may include several different issues (say priority, moving cost, gap, etc) with differing weights assigned to each. For example, our proposing attention codelet may place great weight on low moving cost, some weight on fitness value, and little weight on the others. This codelet may propose the second job on the scenario list because of its low cost and high fitness, in spite of low priority and a sizable gap. If no other proposing attention codelet objects (by bringing itself to "consciousness" with an objecting message) and no other such codelet proposes a different job within a given span of time, the timekeeper codelet will mark the proposed job as being one to be offered. If an objection or a new proposal is made in a timely fashion, it will not do so.

Two proposing attention codelets may alternatively propose the same two jobs several times. Several mechanisms tend to prevent continuing oscillation. Each time a codelet proposes the same job it does so with less activation and, so, has less chance of coming to "consciousness." Also, the timekeeper loses patience as the process continues, thereby diminishing the time span required for a decision. Finally, the metacognitive module watches the whole process and intervenes if things get too bad (Zhang et al. 1998a). A job proposal may also alternate with an objection, rather than with another proposal, with the same kinds of consequences. These occurrences may also be interspersed with the creation of new scenarios. If a job is proposed but objected to, and no other is proposed, the scenario building may be expected to continue yielding the possibility of finding a job that can be agreed upon.

***Negotiation***

After IDA has selected one or more jobs to be offered to a given sailor, her next chore is to negotiate with the sailor until one job is decided upon. The US Navy is quite concerned about retention of sailors in the service. This depends heavily on the sailor's job satisfaction. Thus the Navy gives a high priority to the assignment of a job that both satisfies the sailor's preferences and offers opportunity for advancement, sometimes including additional training. Whenever possible the final job assignment is made with the sailor's agreement. IDA must negotiate this agreement with the sailor.

When the initial job offerings are made the sailor may respond in several different ways. He may accept one of the jobs offered. He may decline all of them and request some different job assignment. He may ask for a particular job not among those offered. He may ask that the process be postponed until a new requisition list appears, hoping to find something more to his liking. IDA may accede to or deny any of these requests, the decision often dependent on time constraints and/or the needs of the service. The continuing negotiations offer many possible paths. It ends with one job being assigned to the sailor, most often with his agreement, but sometimes without.

IDA must be able to carry out such negotiations. This requires making decisions and responding to the sailor's messages. We've already had a close look at her decision making processes involving both "consciousness," the behavior net and voluntary action.

### *Metacognition*

Metacognition guides people to select, evaluate, revise, and abandon cognitive tasks, goals, and strategies (Hacker 1997). Most researchers agree that metacognition should include knowledge of one's own knowledge and cognitive processes, and the ability to actively monitor and consciously regulate them.. While metacognition isn't implemented in the current running version of IDA, it's very much a part of the conceptual IDA model, and have been implemented previously in an earlier "conscious" software agent, CMattie (Zhang et al. 1998a). This, admittedly impoverished, metacognition module was able to interrupt oscillations during voluntary action selection, and also to tune the behavior net by adjusting it's parameters so that it ran more smoothly. Metacognition is not yet conscious in our model. We've experimented with several mechanisms for metacognition including a classifier system, a fuzzy logic controller and a fuzzy classifier system. All worked about equally well. The latter also learned.

### *Learning*

The IDA model incorporates several different learning mechanisms. The simplest is the associative learning that occurs as the contents of "consciousness" is stored in associative memory with every "conscious" broadcast. It's the associations between the various items comprising the "conscious" contents that's learned. Also, associations with other similar items are learned by means of the Sparse Distributed Memory mechanism.

Learning into associative memory can be considered a type of declarative learning. IDA is also capable of a more procedural learning. Codelets in IDA participate in a pandemonium theory style organization (Jackson 1987). Those codelets who share time in the spotlight of "consciousness" have associations between them formed or strengthened, or perhaps diminished if things are going badly. Those codelets sharing time in the playing field also change associations, but at a much lesser rate. This dynamic association of codelets allows a Hebbian type of learning. As codelets become more or less associated, the likelihood of their coming to "consciousness" together changes, and with it the likelihood of their together helping to initiate some behavior. These changes in likelihood constitute a kind of procedural learning.

When the same coalition of codelets, acting in parallel or in sequence, often produce a useful result, this coalition can be expected to merge into a higher-level concept codelet. This merging constitutes a second form of temporal proximity learning. The concept codelet, when active, performs the same actions as do its member codelets combined, but without requiring successive executions of the behavior net. This process is comparable to chunking in SOAR (Laird et al. 1987).

In the conceptual model of IDA we include mechanisms for emotions (McCauley, T. L. & Franklin 1998). Action selection will be influenced by emotions via their effect on drives, modeling recent work on human action selection (Damasio 1994). IDA's emotional mechanism includes an activation passing network with weighted links that allows learning. Weights will decay with disuse. Each use tends to decrease the decay rate. Weights will increase according to a sigmoidal function of any activation carried over the link, allowing for Hebbian style learning. Thus IDA is, in principle, able to learn to use emotions to affect various cognitive functions as we humans do.

Also, IDA is, in principle, able to learn metacognitively using the built-in fuzzy classifier mechanism. In particular, IDA should be able to learn to tune her action selection by means of adjusting the parameters of her behavior net as CMattie did before her.

Plans are afoot for IDA to learn by being told by a human (Ramamurthy et al. 1998a, 1998b, 2001), but we don't claim this form of learning as part of the IDA conceptual model since the required mechanisms are not yet completely designed.

A final form of learning in IDA allows the automatization of repeated actions as we humans do. Here, the mechanisms are in place and constitute part of the IDA conceptual model. This is new work that will be the subject of another report, and cannot be briefly described here.

## "Conscious" and unconscious decision making in IDA

As humans do, IDA uses "consciousness" during her decision making (action selection) process is several different ways. Also, she makes some decisions unconsciously. In this section we'll describe these various processes.

If I feel hungry and choose to make myself a Nathan's hotdog, I've engaged in *internally generated voluntary action selection* as described by William James' ideomotor theory. Of course, the process may have been much more complex including consciously considering the possibilities of going to Bozo's for barbecue or too Mikasa for sushi instead, or of simply settling for a piece of fruit, or of eating later after finishing the section I'm writing. Once that decision is made a sequence of actions typically follow. Some of these may involve other instances of internally generated voluntary actions, such as the decision of whether to cook the hotdog on the stove or in the microwave, or whether to wear a coat. Such decisions are characterized by a *conscious consideration of alternatives motivated by some internal drive or goal*. The conscious consideration may be of a single alternative (Nathan's hotdog) where the choice is whether to do it.

James' theory, and our implementation of it in IDA, is described in the section on voluntary action above. IDA engages in internally generated voluntary action selection when deciding which jobs to offer the sailor. This action is motivated by her drive to find a job suitable to both the sailor and the US Navy. Since alternatives are considered without being acted upon, the process is a deliberative one in the sense of Sloman (1999).

Voluntary action selection can also be occasioned by external environmental situations. I want to get into a trailer. The door is locked. Shall I try to find the hidden key or shall I look for Shelby who has a key? This is a example *of externally generated voluntary action selection* which is characterized by voluntary action selection motivated by a drive or goal, but occasioned by an obstacle or an opportunity presented by the external environment. IDA's current incarnation doesn't display externally generated voluntary action selection, but she easily could. Suppose IDA was unable to find a sailor's personnel record. She could consider writing to the sailor to check the social security number, the index of the personnel file. Or, she might think of trying a query based on the sailor's name, or she might refer the matter to a human. To implement this capability would require the addition of a few attention codelets and a couple of new behavior streams with their behavior codelets. Ideally, the IDA conceptual model should allow for internal construction of these codelets and streams as we humans do. This illustrates one of the many gaps in the model. We're working on this one.

Many of our actions, though not voluntary, are nonetheless consciously mediated. Having retrieved the key for the trailer I must be conscious of at least of the steps leading up to the door. I must be conscious of the position of the keyhole. These should not, however, be construed as conscious decision points requiring voluntary action selection. I do not consider whether or not to take the first step, nor whether to insert the key into the keyhole. These actions are parts of several preplanned and executed action sequences, *consciously mediated* in order to fill in required data from the environment, in this case location of step and keyhole. Consciously mediated action selection is involuntary in that a sequence of actions is selected, perhaps voluntarily, and then proceeds via a sequence of action decisions that may involve consciousness but that involvement is not a matter of deciding whether to take the action in question. Rather consciousness provides information needed as a precondition to taking the action, such as a location. Consciousness mediates in helping to align the key with the keyhole.

Consciousness can also mediate internal action selections. Here's an example thoughtfully provided by my daughter Sunny. She and her friend, Christina, were playing the game of guessing people's star signs. Christina made a guess for Sunny's brother, Bruce. As part of the process of seeing if Christina guessed right, Sunny had to recall Bruce's birthday, an internal action, but not voluntarily selected. Sunny wasn't considering whether to remember Bruce's birthday. She simply had to have the birthday consciously available in order to tell it to Christina.

The process of "consciously" mediated action selection in IDA is implemented as described in the section on action selection above. The preplanned actions sequences mentioned above are implemented as behavior streams. Behavior codelets instantiate such behavior streams, perhaps motivated by voluntary action selection, perhaps not. The selection of a behavior in a stream for execution is accomplished by the behavior net quite unconsciously, as described above. For example, suppose the executed behavior reads a field from the sailor's personnel record. The contents of that field, brought to "consciousness" by an attention codelet, will serve to trigger the reading of the next field, provided the contents aren't unexpected. Behavior codelets responsible for obtaining the current duty station, seeing the contents of the name field in "consciousness," know that it's their turn to act. They send (environmental) activation to the next behavior in the stream, their behavior, helping it to be executed next. In this case the "conscious" contents serve as a cue for the next action. We hypothesize that well rehearsed action sequences in humans work the same way.

IDA also performs actions "consciously" mediated internally. For instance, when composing an email message to a sailor, IDA "consciously" fills in the blanks in scripts carried by codelets. The intended content of a field is "consciously" retrieved from working memory by an executing behavior. Behavior codelets responding to the resulting "conscious" broadcast not only send activation to the next behavior in the stream (fill in this particular field) but bind the variable containing the intended content of the field. Just as one can't insert a key into a key into a keyhole without seeing to align it properly, one can't fill in a blank in a script without knowing the intended content of the blank.

Finally, both humans and IDA, at least in her conceptual model, perform actions that are not consciously mediated. Once the key is aligned with the keyhole, the sequence of actions involving inserting the key and turning it can be done unconsciously, as long as it proceeds as expected. Such action sequences involving primitive actions by a number of distinct

muscle groups are typically first learned consciously and become automatized with use. The same process is available to IDA in her conceptual model. Transferring needed data from an external personnel file to her workspace (working memory) is at first done "consciously" mediated for each piece of data, as described above. Later association between adjacent behaviors become strong enough for one to call the next behavior in pandemonium style (Jackson 1987) without the benefit of "conscious" intervention. Thus we have a machine version of the automatization process. This version also includes the capability of bringing a behavior back to "consciousness" when it doesn't work as expected. This work is as yet unpublished.

## Is IDA conscious?

Clearly IDA is quite a complex software agent that models a broad swath of human cognition including "consciousness" in the sense of implementing global workspace theory (Franklin & Graesser 1999). But, is there any sense in which she can be said to be conscious without the quotes. Well, there's the ill-defined cluster concept of functional consciousness. IDA exhibits both external and internal voluntary actions selection, as well as consciously mediated action selection of both the internal and external variety. Though IDA does not, as yet, engage in non-routine problem solving, work on adding that capability is in progress. She uses her "consciousness" module to handle routine problems with novel content. All this together makes a strong case, in my view, for functional consciousness.

But what about phenomenal consciousness? Can we claim it for IDA? Is she *really* a conscious artifact? I can see no convincing arguments for such a claim. One might argue that anything that senses must have some sort of subjective consciousness as a consequence. Would we thus attribute consciousness to a thermostat? Not lightly. On the other hand, I can see no convincing arguments against a claim for phenomenal consciousness in IDA. Yes, she's not biological, much less the possessor of neuropil. Does that really matter? Perhaps. Or, perhaps neuropil is only necessary for phenomenal consciousness in biological agents. IDA is a piece of running software and is thus immaterial. What is there to be conscious? (Tom Ziempke, personal communication) On the other hand, perhaps human consciousness runs as much on a virtual machine as IDA does. These requirements that conscious beings be material or biological or have neuropil may be no more valid than those prompting the denial of phenomenal consciousness to other apes on the grounds that they are not human.

But, can IDA report her "conscious experiences"? No, but it's only because we haven't built in that capability for her. To have her report the contents of her "consciousness" would require no new ideas or mechanisms. It would only be a matter of the time and effort to do the work. We know how, and it would not be in principle different from what IDA does now. Reportability is NOT the issue. It's currently part of the IDA conceptual model.

What about self-consciousness? Though a self is postulated as part of global workspace theory, it has not been implemented in IDA, partially because we just didn't understand the concept well enough to know how to implement it. Damasio's formulation of a self (Damasio 1999) has given us ideas so that adding a self to IDA by way of a proto-self, now seems possible. However the mechanisms are not in place, meaning that a self is not yet part of the IDA conceptual model. It does seem at least possible that a self is necessary for phenomenal consciousness though self-consciousness isn't. The work of Humphrey, speculating on how phenomenal consciousness might have evolved in relatively simple organisms, makes this idea seem quite plausible (2000).

## Conclusion

So, what are we to conclude from all this? Suppose the neuroscientists succeed in finding neural correlates of phenomenal consciousness. This may yield evidence as to the consciousness or lack thereof of various animals. It seems unlikely to help us at all with the question of machine consciousness. Nor does anything else seem likely to be of help. Subjective consciousness seems to be such an inherently first person phenomena as to be impervious to any form of proof. A solipsistic position seems as defensible as any other. The best we can do is to offer evidence in favor such as reportablity that can be externally confirmed. Such evidence, whether based on behavior or on internal structure and/or activity, is unlikely to make much of an impression on the philosophers who are willing to take seriously the possibility of zombies, thus giving rise to the *hard problem* of consciousness.

Following Chalmers (Chalmers 1996) I'm beginning to view phenomenal consciousness as a fundamental process of nature comparable to mass or energy. Both mass and energy can only be measured indirectly by their effects though inferences can be drawn from their structure and activity (again effects). I suspect that the same will prove true of phenomenal consciousness. How can we measure its effects? Will this become the *important problem* of consciousness? Likely, I would think. In the realm of machine consciousness this problem seems even harder. Deducing structure from behavior is notoriously difficult to which both the psychologists and the neuroscientists will attest. In software agents such deductions are both theoretically impossible and practically problematic. Still, measuring mass and energy weren't easy problems until someone solved them.

Where does that leave us? At least for now, the attribution of phenomenal consciousness will depend on the level of evidential support the person doing the attributing will accept. Were I faced with Commander Data from Star Trek: The Next

Generation I would undoubted attribute consciousness. One the other had I do not attribute phenomenal consciousness to my own "conscious" software agent, IDA, in spite of her many human-like behaviors. This in spite of watching several US Navy detailers repeatedly nodding their heads saying "Yes, that's how I do it" while watching IDA's internal and external actions as she performs her task. I suspect they might be more willing to attribute phenomenal consciousness than I.

Speaking of the "additional magic ingredient" that produces phenomenal consciousness out of functional consciousness, Harvey says "…the additional magic ingredient for [phenomenal consciousness] is merely a change in attitude in us, the observers." (Harvey 2002) The claim is that the presence or absence of phenomenal consciousness can never be more than a matter of attribution. Baring that claim in mind, let me close with a prediction. I expect there to be software agents and robots sufficiently intelligent, sufficiently capable, and sufficiently communicative that people with simply assume that they are conscious artifacts (Moravec 1988). The issue of machine consciousness will no longer be relevant.

## Acknowledgements

This work has been supported in part by ONR grant N00014-98-1-0332. Much of it emerged from the combined effort of members of the Conscious Software Research Group at the University of Memphis. These include, or have included, Art Graesser, , Lee McCauley, Hongjun Song, Zhaohua Zhang, Satish Ambati, Ashraf Anwar, Myles Bogner, Arpad Kelemen, Ravikumar Kondadadi, Irina Makkaveeva, Aregahegn Negatu, Alexei Stoliartchouk, Uma Ramamurthy.

## References

Aleksander, I. 2000. *How to Build a Mind*. London: Weidenfeld and Nicolson.

Allen, J. J. 1995. *Natural Language Understanding*. Redwood City CA: Benjamin/Cummings; Benjamin; Cummings.

Anwar, A., and S. Franklin. to appear. Sparse Distributed Memory for "Conscious" Software Agents. *Cognitive Systems Research*.

Anwar, A., D. Dasgupta, and S. Franklin; 1999. Using Genetic Algorithms for Sparse Distributed Memory Initialization. International Conference Genetic and Evolutionary Computation(GECCO). July, 1999.

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.

Baars, B.; 2001. Surrounded by consciousness: The scientific evidence for animal consciousness since the Mesozoic. Consciousness and its Place in Nature: Toward a Science of Consciousness; ; Skovde, Sweden; August 7-11, 2001. .

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

Chalmers, D. J. 1996. *The Conscious Mind*. Oxford: Oxford University Press.

Cotterill, R.; 2001; ; ; Consciousness in Theory - and in the Computer?. Consciousness and its Place in Nature: Toward a Science of Consciousness. Skovde, Sweden; August 7-11, 2001.

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

Damasio, A. R. 1999. *The Feeling of What Happens*. New York: Harcourt Brace.

Edelman, G. M. 1987. *Neural Darwinism*. New York: Basic Books.

Edelman, G. M., and G. Tononi. 2000. *A Universe of Consciousness*. New York: Basic Books.

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

Franklin, S. 1997. Autonomous Agents as Embodied AI. *Cybernetics and Systems* 28:499–520.

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.

Franklin, S., and A. Graesser. 1999. A Software Agent Model of Consciousness. *Consciousness and Cognition* 8:285–305.

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

Freeman, W. J. 1995. *Societies of Brains*. Hillsdale NJ: Lawrence Erlbaum.

Freeman, W. J. 1999. *How Brains Make Up Their Minds*. London: Weidenfeld & Nicolson General.

Gallup, G. 1982. Self-awareness and the emergence of mind in primates. *American Journal of Primatology* 2:237–246.

Glenberg, A. M. 1997. What memory is for. *Behavioral and Brain Sciences* 20:1–19.

Hacker, D. 1997. Metacognitive: Definitions and Empirical Foundations. In *Metacognition in Educational Theory and Practice*, ed. D. Hacker, J. Dunlosky, and A. Graesser. Hillsdale, NJ: Erlbaum. (Hillsdale)

Harvey, I. 2002. Evolving Robot Consciousness: The Easy Problems and the Rest. In *Evolving Consciousness*, ed. J. H. Fetzer. Amsterdam: John Benjamins.

Hofstadter, D. R., and M. Mitchell. 1994. 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.

Humphrey, N. 2000. *How to Solve the Mind-Body Problem*. Bowling Green, OH: Imprint Academic.

Jackson, J. V. 1987. Idea for a Mind. *Siggart* Newsletter, 181:23–26.
James, W. 1890. *The Principles of Psychology*. Cambridge, MA: Harvard University Press.
Jurafsky, D., and J. H. Martin. 2000. *Speech and Language Processing*. Englewood Cliffs, NJ: Prentice-Hall.
Kanerva, P. 1988. *Sparse Distributed Memory*. Cambridge MA: The MIT Press.
Kelemen, A., Y. Liang, R. Kozma, and S. Franklin. 2002 in press. Optimizing Intelligent Agent's Constraint Satisfaction with Neural Networks. In *Innovations in Intelligent Systems*, ed. A. Abraham, and B. Nath. Heidelberg, Germany: Springer-Verlag,.
Laird, E. J., A. Newell, and Rosenbloom P. S. 1987. SOAR: An Architecture for General Intelligence. *Artificial Intelligence* 33:1–64.
Maes, P. 1989. How to do the right thing. *Connection Science* 1:291–323.
Maturana, R. H., and F. J. Varela. 1980. *Autopoiesis and Cognition: The Realization of the Living, Dordrecht*. Netherlands: Reidel.
Maturana, H. R. 1975. The Organization of the Living: A Theory of the Living Organization. *International Journal of Man-Machine Studies* 7:313–332.
McCauley, L., S. Franklin, and M. Bogner. 2000. An Emotion-Based "Conscious" Software Agent Architecture. In *Affective Interactions*, Lecture Notes on Artificial Intelligence ed., vol. 1814, ed. A. Paiva. Berlin: Springer.
McCauley, T. L., and S. Franklin. 1998. An Architecture for Emotion. In *AAAI Fall Symposium Emotional and Intelligent: The Tangled Knot of Cognition*. Menlo Park, CA: AAAI Press.
Minsky, M. 1985. *The Society of Mind*. New York: Simon and Schuster.
Moravec, H. 1988. *Mind Children*. Cambridge MA: Harvard University Press.
Negatu, A., and S. Franklin. 2002. An action selection mechanism for 'conscious' software agents. *Cognitive Science Quarterly* 2:363–386.
Ornstein, R. 1986. *Multimind*. Boston: Houghton Mifflin.
Ramamurthy, U., M. Bogner, and S. Franklin. 1998a. "Conscious" Learning in an Adaptive Software Agent. In *Proceedings of the Second Asia Pacific Conference on Simulated Evolution and Learning*, ed. X. Yao, R. I. McKay, C. S. Newton, J. H. Kim, and T. Furuhashi. Canberra, Australia: University of New South Wales.
Ramamurthy, U., S. Franklin, and A. Negatu. 1998b. Learning Concepts in Software Agents. In *From animals to animats 5: Proceedings of The Fifth International Conference on Simulation of Adaptive Behavior*, ed. R. Pfeifer, B. Blumberg, J-A. Meyer, and S. W. Wilson. Cambridge,Mass: MIT Press.
Ramamurthy, U., A. Negatu, and S. Franklin; 2001. Learning Mechanisms for Intelligent Systems. SSGRR-2001 International Conference on Advances in Infrastructure for e-Business, e-Education and e-Science on the Internet. L'Aquila, Italy; August 6-12, 2001.
Sloman, A. 1987. Motives Mechanisms Emotions. *Cognition and Emotion* 1:217–234.
Sloman, A. 1995. A philosophical encounter. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*. : Morgan Kaufman.
Sloman, A. 1999. What Sort of Architecture is Required for a Human-like Agent? In *Foundations of Rational Agency*, ed. M. Wooldridge, and A. Rao. Dordrecht, Netherlands: Kluwer Academic Publishers.
Zhang, Z., D. Dasgupta, and S. Franklin. 1998a. Metacognition in Software Agents using Classifier Systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*. Madison, Wisconsin: MIT Press.
Zhang, Z., S. Franklin, B. Olde, Y. Wan, and A. Graesser. 1998b. Natural Language Sensing for Autonomous Agents. In *Proceedings of IEEE International Joint Symposia on Intellgence Systems 98*.