1999

A Software Agent Model of Consciousness

S. Franklin

A. Graesser

A Software Agent Model of Consciousness

By Stan Franklin and Art Graesser[1, 2]

"conscious" Software Research Group

Institute for Intelligent Systems

The University of Memphis

[1] Authors supported in part by NSF grant SBR-9720314 and by ONR grants N00014-98-1-0332 and N00014-98-1-0331

[2] With essential cooperation from the "conscious" Software Research Group including Myles Bogner, Derek Harter, Arpad Kellerman, Lee McCauley, Aregahegn Negatu, Fergus Nolan, Hongjun Song, Uma Ramamurthy, Zhaohua Zhang

A Software Agent Model of Consciousness

Stan Franklin

Math Sciences Dept

University of Memphis

Memphis, TN 38152

(901) 678-3142

(901) 678-2480 fax

Abstract

Baars has proposed a psychological theory of consciousness, *called global workspace* theory. The present study describes a software agent implementation of that theory, called "Conscious" Mattie (CMattie). CMattie operates in a clerical domain from within a UNIX operating system, sending messages and interpreting messages in natural language that organize seminars at a university. CMattie fleshes out global workspace theory with a detailed computational model that integrates contemporary architectures in cognitive science and artificial intelligence. Baars lists the psychological "facts that any complete theory of consciousness must explain" in his appendix to *In the Theater of Consciousness* (1997); global workspace theory was designed to explain these "facts." The present article discusses how the design of CMattie accounts for these facts and thereby the extent to which it implements global workspace theory.

Global Workspace Theory

Baars' *global workspace* theory (1988, 1997) postulates that human cognition is implemented by a multitude of relatively small, special-purpose processes, almost always

unconscious. Coalitions of such processes find their way into a global workspace and thus into "consciousness". From this limited capacity workspace, the message of the coalition is broadcast 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. All this takes place under the auspices of various contexts, including goal contexts, perceptual contexts, conceptual contexts, and cultural contexts. Each context is itself a coalition of processes. There is much more to the theory, including attention, learning, action selection, and problem solving.

Baars' global workspace theory is a comprehensive theory of both consciousness and general cognition. The theory is appropriately articulated at an abstract, functional, verbal level because of its attempt to account for many cognitive phenomena and empirical findings. Baars' description goes a long way in capturing the cognitive principles and mechanisms at a functional level, but it does not specify all of the mechanisms at the levels of computation and neuroscience. It is the computational level that is the primary focus of the present article. Our goal is to flesh out the computational architectures and mechanisms that are left unspecified in his articulation of global workspace theory. At the very least, this article will identify some of the computational issues that must be addressed in any computer implementation of global workspace theory.

Autonomous Agents

CMattie is an autonomous software agent. An autonomous agent (Franklin & Graesser, 1997) is a system "situated in" 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. In other words, it is structurally coupled to its environment (Maturana, 1975; Maturana & Varela, 1980; Varela, Thompson and Rosch 1991). 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. CMattie is an autonomous software agent, designed for a specific clerical task, that 'lives' in a real world UNIX operating system. The autonomous software agent that we are developing is equipped with computational versions of cognitive features, such as multiple senses, perception, short and long term memory, attention, planning, reasoning, problem solving, learning, emotions, multiple drives, and so forth. In this sense, our software agents are cognitive agents (Franklin, 1997).

We believe that cognitive software agents have the potential to play a synergistic role in both cognitive theory and intelligent software. Minds can be viewed as control structures for autonomous agents (Franklin, 1995). A theory of mind constrains the design of a cognitive agent that implements that theory. While a theory is typically abstract and only broadly sketches an architecture, an implemented computational design provides a fully articulated architecture and complete set of mechanisms. This architecture and set of mechanisms provides a richer, more concrete, and more decisive theory. Moreover, every design decision taken during an implementation furnishes a hypothesis about how human minds work. These hypotheses motivate experiments with humans and other forms of empirical tests. Conversely, the results of such experiments motivate corresponding modifications of the architecture and mechanisms of the cognitive agent. In this way, the concepts and methodologies of cognitive science and of computer science will work synergistically to enhance our understanding of mechanisms of mind (Franklin, 1997). CMattie was designed with this research strategy in mind.

CMattie is designed for the explicit purpose of implementing global workspace theory. Some of the components have already been programmed whereas others are at the design stage. However, she is far enough along to allow us to address the key question of this paper: how well does CMattie, as a conceptual model, account for the psychological facts that global workspace theory was constructed to explain?

The CM-Architecture

Conceptually, CMattie is an autonomous software agent that 'lives' in a UNIX system. CMattie communicates in natural language with seminar organizers and attendees via email, "comprehends" email messages, composes messages, and sends seminar announcements, all without human direction. CMattie is an extension of Virtutal Mattie (VMattie) (Franklin, Graesser, Olde, Song & Negatu, 1996; Song & Franklin, forthcoming; Zhang, Franklin, Olde, Wan & Graesser, 1998). VMattie, which is currently running in a beta testing stage, implements an initial set of components of global workspace theory. VMattie performs all of the functions of CMattie, as listed above, but does so "unconsciously" and without the ability to learn and to flexibly handle novel situations. CMattie adds the missing pieces of global workspace theory, including computational versions of attention, associative and episodic memories, emotions, learning and metacognition.

The computational mechanisms of CMattie incorporate some of the mechanisms of mind discussed at length in *Artificial Minds* (Franklin 1995). Each of the mechanisms mentioned required considerable modification and, often, extension in order that they be suitable for use in CMattie. The high-level action selection uses an extended form of Maes' behavior net (1990). The net is comprised of behaviors, drives and links between them. Activation spreads in one direction from the drives, and in the other from CMattie's percepts. The currently active behavior

is chosen from those whose preconditions are met and whose activations are over threshold. Lower level actions are taken by codelets in the manner of the Copycat architecture (Hofstadter &Mitchell, 1994; Mitchell, 1993). Each codelet is a small piece of code, a little program, that does one thing. Our implementation of Baars' global workspace, discussed in more detail below, relys heavily on the playing field in Jackson's pandemonium theory (1987). All active codelets inhabit the playing field, and those in "consciousness" occupy the global workspace. Kanerva's sparse distributed memory (1988; Anwar & Franklin, forthcoming**)** provides a human-like associative memory for the agent whereas episodic memory (case-based) follows Kolodner's (1993) model. CMattie's emotion mechanism uses pandemonium theory (McCauley & Franklin, in press). Her metacognition module is based on a fuzzy version of Holland's classifier system (Holland, 1986; Zhang, Franklin & Dasgupta, in press). Learning by CMattie is accomplished by a number of mechanisms. Behavior nets can learn by adjusting the weights on links as in artificial neural networks (Maes, 1992). The demons in pandemonium theory become (more) associated as they occur together in the arena (Jackson 1987). The associations that occur automatically in sparse distribute memory constitute learning (Kanerva 1988). CMattie also employs one-trial learning using case-based reasoning (Bogner, Ramamurthy and Franklin, forthcoming; Kolodner,1993; Ramamurthy, Franklin & Negatu, in press).

We next turn to a brief account of how the CM-architecture uses these mechanisms to model global workspace theory. The CM-architecture can be conveniently partitioned into more abstract, high-level constructs and lower level, less abstract codelets. Higher-level constructs such as behaviors and some slipnet nodes overlie collections of codelets that actually do their work. In CMattie, Baars' "vast collection of unconscious processes" are implemented as codelets much in the manner of the Copycat architecture, or almost equivalently as Jackson's demons. His

limited capacity global workspace is a portion of Jackson's playing field, which holds the active codelets. Working memory consists of several distinct workspaces, one for perception, one for composing announcements, two for one-trial learning, and others.

Baars speaks of contexts as "…the great array of unconsciious mental sources that shape our conscious experiences and beliefs." (1997 p. 115) He distinguishes several types, including perceptual contexts, conceptual contexts and goal contexts. The perceptual context provided by a large body of water might help me interpret a white, rectangular cloth as a sail rather than as a bed sheet. The conceptual context of a discussion of money might point me at interpreting "Let's go down by the bank?" as something other that an invitation for a walk, a picnic or a swim. Hunger might well give rise to a goal context. Contexts in global workspace theory are coalitions of codelets. In the CM-architecture high-level constructs are often identified with their underlying collections of codelets and, thus, can be thought of as contexts. Perceptual contexts include particular nodes from a slipnet type associative memory à la Copycat (similar to a semantic net), and particular templates in workspaces. For example, a message-type node is a perceptual context. A node type perceptual context becomes active via spreading activation in the slipnet when the node reaches a threshold. Several nodes can be active at once, producing composite perceptual contexts. These mechanisms allow "conscious" experiences to trigger "unconscious" contexts that help to interpret later "conscious" events. Conceptual contexts also reside in the slipnet, as well as in associative memory. Goal contexts are implemented as instantiated behaviors in a much more dynamic version of Maes' behavior nets. They become active by having preconditions met and by exceeding a time variable threshold. Goal hierarchies are implemented as instantiated behaviors and their associated drives. (My hunger drive might give rise to the goal of eating sushi. The first behavior toward that goal might be walking to my

car.) The dominant goal context is determined by the currently active instantiated behavior. The dominant goal hierarchy is one rooted at the drive associated with the currently active instantiated behavior.

Recruitment of coalitions of "unconscious" processors is accomplished by "consciousness" codelets, as well as by the associations among the occupants of the global workspace, via pandemonium theory. Always active, "consciousness" codelets jump into action when problematic situations occur. An example is described below. Attention is what goes into the global workspace from perception, and from internal monitoring. It also uses pandemonium theory, but requires an extension of it. Both recruitment and attention are modulated by the various context hierarchies. Learning occurs via several mechanisms: as in pandemonium theory, as in sparse distributed memory, as in behavior nets, by extensions of these, and by other mechanisms. High-level action selection is provided by the instantiated behavior net. At a low level, CMattie follows the Copycat architecture procedure of temperature controlled (here emotionally controlled), parallel terraced scanning. Problem solving is accomplished via "conscious" recruitment of coalitions of "unconscious" codelets.

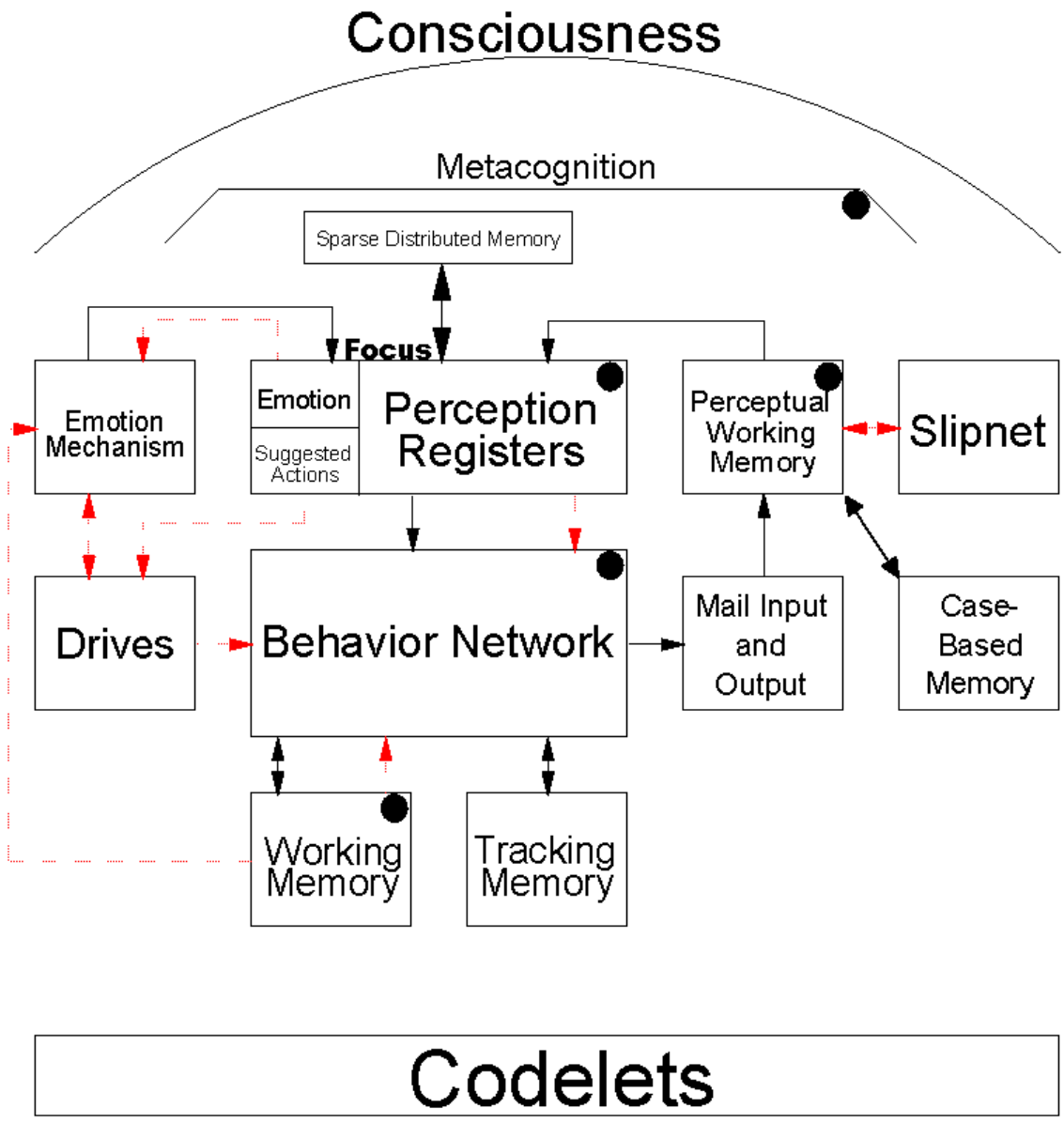

Figure 1. Most of the "conscious" Mattie architecture

Figure 1 gives a functional overview of most of the CMattie architecture. Several important functions, for example conceptual and behavioral learning, are omitted from the diagram, but not from our discussion. Detailed descriptions of the architecture and mechanisms are given in a series of papers by members of the "Conscious" Software Research group (Anwar & Franklin, forthcoming**;** Bogner, Ramamurthy, and Franklin (in press). "Consciousness" and Conceptual Learning in a Socially Situated Agent. in Kerstin Dautenhahn ed. Human Cognition and Social Agent Technology; Franklin, 1997a; Franklin, Graesser, Olde, Song, & Negatu, 1996; McCauley & Franklin, 1998; Ramamurthy, Franklin & Negatu, in press; Song & Franklin, forthcoming; Zhang, Franklin, Olde, Wan & Graesser, 1998; Zhang, Franklin & Dasgupta, 1998).

CMattie in Action

This section described walks through a typical incoming message through the CM-architecture. Suppose CMattie receives the following message from Stan Franklin:

> "Dear CMattie, Hope your day hasn't been as busy as mine. Next week Art Graesser will speak to the Cognitive Science Seminar on Kintsch's Comprehension Theory. Have a good one. Stan"

After the message arrives in CMattie's inbox, codelets transfer it to the perceptual workspace where other codelets begin the process of making sense of it. Some codelet will recognize "Cognitive Science Seminar" as a known seminar and activate its node in the slipnet. Another will identify "Art Graesser" as a name, possibly a speaker, again activating a node. Other such recognitions will occur. Eventually the slipnet's spreading activation will push some message-type node over threshold, in this case probably the Speaker-Topic message type. Once a tentative

message type is identified, a template containing the fields, both mandatory and optional, for that type is placed in the perceptual workspace, and codelets begin to fill in the blanks. If successful, the message is deemed "understood" and its content (meaning) is placed in perception registers inside the focus (see Figure 2), for use by the rest of the system. That is what is known as the "initial percept."

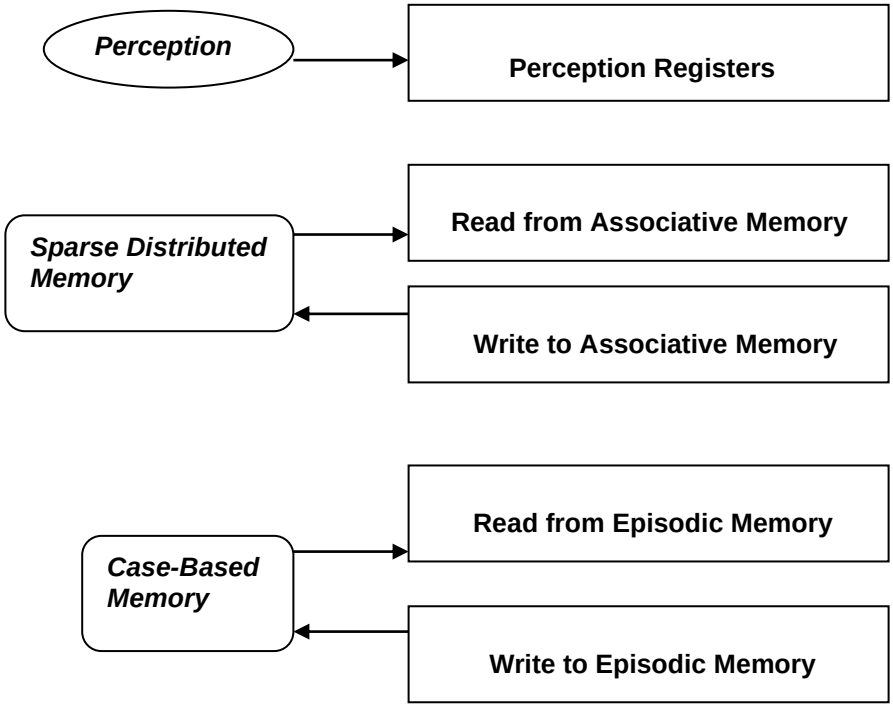

Figure 2. The Focus

CMattie will have ignored the first and last sentences of the messages, since their contents would have triggered no recognition by her codelets. She would have recognized and filled in the following fields: Organizer-Stan Franklin; Seminar Name-Cognitive Science Seminar; Speaker Name-Art Graesser; Title-Kintsch's Comprehension Theory. Each would occupy the appropriate perception register in the focus. Other optional fields, including Date,

Day-of-week, Building-Room and Time would result in empty registers. VMattie, CMattie's predecessor, does all of this with almost 100% success.

CMattie's understanding of incoming messages is based on her knowledge of surface features that are stored in her slipnet and in her perceptual codelets. It is possible for us to implement this on a computer in real time because of the narrow domain of knowledge in our application. CMattie need only be concerned with a few message types, for example seminar-initiation, speaker-topic, seminar-conclusion, change-of-time (or day or room), on-mailing-list, off-mailing-list, negative-response and a few others.

At this point the perception registers already begin to activate the emotion mechanism. Simultaneously, the associative and episodic memories are read into sets of registers of their own within the focus (see Figure 2), with the initial percept used as address to these content-addressable memories. The new contents of the associative memory registers should contain correct default information about the Date, Time, etc, as well as an associated emotion and an associated behavior. Together with the perception register contents, these constitute CMattie's "percept" from the "sensory stimulus" of the incoming message. All this again activates the emotion mechanism, and through "consciousness" the behavior net. Typically, a new behavior will be selected by the net and a new emotion by its mechanism as well. The behavior, the emotion, and the initial percept augmented by default values are then written from the write registers to both memories. This ends one perceptual cycle.

"Consciousness" codelets carrying the information of this percept typically form a coalition that falls within the "spotlight of consciousness." This corresponds to the global workspace of Baars' theory. The resulting broadcast results in streams of behaviors being instantiated, eventually leading to an "acknowledge message" being sent to the sender of the

message. It would also result in this information being written into the announcement template in the composition workspace. This activation for these behavior streams comes from the drives, is influenced by emotions, and is driven by the percept as extended by memory. A selected behavior will, in turn, activate the codelets in its service. These codelets do the actual work.

CMattie's high-level actions, as produced by streams of several behaviors, include sending an acknowledgement, sending a seminar announcement, sending a reminder, adding to or removing from the mailing list and a very few others. One of these others sends a warning to seminar organizers whose sessions for the week overlap in time and place. This situation might be discovered by a "consciousness" codelet comparing the incoming percept as fleshed out from associative memory to the contents returned from the case-based episodic memory. The latter might describe another seminar with overlapping time and place. In this case the "consciousness" codelet making the discovery would become highly activated, would be gathered into a coalition with other codelets carrying the pertinent information, and this coalition would find its way into the spotlight of "consciousness." The resulting global broadcast would awaken a stream of behaviors to send the appropriate warnings.

Accounting for the Facts

When Baars published his recent book describing his global workspace theory (1997), he included an appendix that summarized the "major bodies of evidence about conscious experience." That is, he listed a couple dozen psychological findings that "any complete theory [of consciousness] must explain." Global workspace theory was designed by Baars to explain these "facts." This section reviews these facts and discusses the extent to which CMattie models them.

It is important to acknowledge one of the central arguments that we are making in this article with respect to software and consciousness. We assume that we are simulating consciousness on a computer to the extent that we can simulate the psychological and empirical phenomena that Baars articulates as being the bodies of evidence about conscious experience. He has articulated the facts that have constrained his global workspace theory. Similarly, we have developed a conceptual model and, hopefully, a computational system that models these

facts and implements the global workspace theory. We argue that this is a justifiable method of approaching the problem scientifically. We are not making any strong claims about whether the software is truly aware, or whether the awareness in software is equivalent to awareness in humans. We regard such questions as unanswerable or in the provinces of philosophy (ontology, not epistomology). Our present approach will presumably advance the scientific study of consciousness and may be useful to the philosophical debates.

(1) Below-threshold or masked stimulation (page 170). An external stimulus can have an impact on unconscious processes without having a direct impact on consciousness. This phenomenon essentially addresses the disassociation between conscious and unconscious experience. For example, suppose that we briefly present the word MONEY for 5 milliseconds and then immediately mask the word with letters XXXXX to cut off stimulus-driven sensing of the word. The brief presentation is sufficient to activate unconscious processors, but not long enough for a coalition of processors to evolve into the conscious spotlight. The brief activation of MONEY can still prime the activation of words that are semantically related to it (e.g.,. dollars, bank, mint). The activation of semantic associates is accomplished without any participation of consciousness.

CMattie is capable of simulating the unconscious activation of processors. One subset of the codelets in working memory participates in the coalition of codelets that are in the spotlight of "consciousness". However, another subset (most of them, in fact) never makes it into the spotlight of "consciousness". The strictly "unconscious" codelets are activated and in turn activate other codelets, but they never become part of a coalition that reaches "consciousness". The pertinent example her concerns perceptual codelets working with the slipnet. These

"unconscious" codelets often activate a message type node that isn't chosen for the message in question. The residual activation gives this node a leg up when the next message arrives.

(2). Preperceptual processes (page 170). The unconscious preperceptual processes tap the meaning representation in addition to the surface code and sensations. While people are sleeping, they are not directly attending to the environment. However, the meaning of words spoken by others and the lyrics on the radio can have an impact on the cognitive system. When a word in a language has multiple meanings (e.g., bank is both a region by a body of water and a building that houses money), the relevant meaning of the words is resolved without the need for conscious processing (Kintsch, 1998). When the environment is severely degraded (e.g., the fuzzy and warped image when you put on the wrong pair of spectacles), consciousness is needed to generate hypotheses about the objects and features in the environment.

Preperceptual processing in CMattie includes all processing prior to the appearance of information in the perception registers. These processes include meanings of words in addition to the more surface characteristics, such as letters, punctuation marks, and quotes. Local ambiguities are often resolved with respect to the current perceptual context, namely the message type. "unconscious" preperceptual hypotheses occur both before and after the choice of a message type. Issues that cannot be resolved unconsciously eventually make their way into the spotlight of attention.

(3) Postperceptual representations (p. 170). Humans habituate to stimulus events that frequently occur in their environment. A new piece of jewelry can be attention demanding for a few minutes after first put on, but we fail to notice it after we have worn it for a few days. We can become habituated to the noisy television or lawnmower. Indeed, sometimes it is the silence that

can be distracting whenever we have successfully become habituated to a noisy environment. The deviations from norms and habituated processing command our attention and consciousness.

CMattie has a slipnet of typical knowledge to which the system has been habituated. For example, particular seminars meet at particular times and locations and have a particular organizer. When an email message contains content that is compatible with the slipnet content, then the message can be perperceptually processed unconsciously, that is without the services of a "consciousness" codelet and without a global broadcast. However, when the content of the message clashes with the slipnet content, then "consciousness" needs to be recruited to resolve the discrepancy. These mechanisms are handled naturally by CMattie. One component that is currently being implemented is the process of learning during habituation. We are currently adding this feature as part of CMattie's conceptual learning apparatus. This will be accomplished using episodic memory, implemented by case-based memory (Kolodner, 1993). Whenever a particular seminar is held at a particular room, for example, this invariance is induced and the slipnet's content gets updated. The deviations from habituation will be handled by an explanation-based learning mechanism (Mooney, 1990; Schank, 1986) that induces new concepts, dimensions, and indexes from a set of deviant cases.

(4) Unaccessed interpretations of ambiguous stimuli (p. 171). Most stimulus events are ambiguous when considered out of context, but are rarely ambiguous when considered in context. For example, the word BANK can refer to a building with money, a region of land by the river, or a type of shot in a basketball game. An adequate cognitive model accurately accounts for the process of resolving the ambiguity when the stimulus is processed in context. When an ambiguous word is being processed during the first 200 milliseconds, all senses of the word are initially activated although only one of the senses is relevant to the context (Kintsch,

1998). Later in the processing stream (after 400 milliseconds) the constraints of context prevail and there is convergence on a single sense (except for the rare occasions when two senses equally fit the context). This convergence onto a single sense of a word is normally accomplished "unconsciously," but occasionally "consciousness" does play a role in the disambiguation.

Once again, "unconscious" interpretations are very much part of the CM-architecture. Slipnet nodes retain their increased activation even when they fail to make it into the spotlight of "consciousness". One would expect the understanding of ambiguous data to take longer in VMattie, even when the processing is accomplished in parallel. Extra time is of course needed when "consciousness" is recruited to resolve clashes and contradictions. However, even when "consciousness" is not involved, extra time is needed to sort out which of the alternative meanings is most compatible with the message type and the remaining content. CMattie will eventually be tested to verify whether ambiguous data actually take longer to process.

The matter of emotional priming (p. 172) is also accounted for. As seen in Figure 1, CMattie's emotions are influenced by associated memories, drives, and internal perceptual states. They activate drives and these drives in turn activate appropriate streams of behaviors. CMattie will become anxious as the time for an announcement approaches without all the information being in. She will be annoyed at organizers who fail to respond to her duns. She may be fearful at the possibility of an impending shutdown of the system. The annoyance and the fear result from the actions of emotion codelets. All these emotions serve to direct her attention and to spur her actions. She may, or may not, become "conscious" of the emotions. But, as in Baars' example, her actions can be primed by prior emotions.

(5) Contextual constraints on perception (p. 173). We often perceive what we expect to see, and many of these expectations are formulated at an unconscious level. Similarly, perceptual experiences in CMattie are often constrained by "unconscious" factors. The currently chosen message type, most often "unconscious", constrains the fields available for perception. For example, with a change-of-time message type as a perceptual context, the "30" in "1:30" would be recognized as part of a time rather than as part of a date (as in "April 30"). Earlier in the process, the "unconscious" recognition of a day of the week serves to constrain the recognition of a message type. CMattie has no difficulty accounting for these contextual constraints on perception.

(6) Expectations of specific stimuli (p. 173). We frequently have expectations of what will happen next at an abstract level, but we rarely predict what specific stimulus events will occur (Graesser, Singer, & Trabasso, 1994). We expect that we will have dinner the next day, but we rarely have precise expectations of what we will eat and how much. Consequently, clashes with expectations can readily be spotted whereas it is difficult to get humans to be specific in articulating what expectations they are having.

The understanding of messages by CMattie occurs in two stages. First a candidate message type is selected on the basis of what types of fields (day of week, time, various keywords) are recognized by codelets. Then a template for that type of message is moved into the workspace and the codelets attempt to fill in the "blanks." This template implements expectations as to what is to be found, expectations expressed in terms of which classes of codelets are looking for what. The expectations are at an abstract level, awaiting classes of input rather than specific input. For example, speaker name is expected, but there is no expectation about who the speaker will be. These expectations influence lower level "unconscious" actions.

However, when the content of the input clashes with the expectations, “consciousness” needs to be recruited to help resolve the discrepancies. This is accomplished by “consciousness” codelets and a resulting global broadcast as described above.

(7) The conscious side of imagery (p. 174). According to Baars, images are "quasi-perceptual events" that occur when there is no external stimulation in any modality. Imagery includes inner speech and emotional feelings in addition to the prototypical case of the mental visual image. Baars claims that the construction of these images is conscious. Consciousness is recruited when we construct in our minds an argument that we hope to have with a member of the family, when we imagine the feelings of revenge after losing a basketball game, and when we imagine the perfect dessert to have after a frustrating day. One of Baars’ goals is to explain which of the images are conscious and which are unconscious.

CMattie accommodates “conscious” imagery in the form of “conscious” goals, “conscious” messages under construction, and items from episodic memory. For example, a codelet might note that information on the Complex Systems Seminar is missing and brings this lack to consciousness. This filling of a gap is a constructive process that requires consciousness. An impending shutdown may engender fear in CMattie, which can become “conscious” by being part of a coalition that makes it into the spotlight.

(8) Memory images before retrieval (p. 174). Episodic memory is represented in some fashion and this representation is tapped during the process of actively retrieving and reconstructing the memory. For example, suppose that you want to remember what you wore at the last New Year's Eve party. At the initial stage of memory retrieval there is some form of code or representation that is accessed unconsciously. In CMattie, this initial representation typically arises from an incoming message. The resulting percept addresses both sparse distributed

memory (Kanerva, 1988) and case-based memory (Kolodner, 1993). Reconstructive processes subsequently embellish this initial percept and it becomes "conscious". On some occasions, humans replay or mentally simulate the fleshed-out image. These are the instances when consciousness must be recruited to play out the image. Thus, features of the representation must be mapped onto a rudimentary spatial coordinate system when visual mental images are reconstructed in the mind's eye (Baddeley, 1992; Kosslyn, 1995). Similarly, in CMattie a stream of behaviors must be mapped onto a projected temporal coordinate system when a series of episodes are retrieved and envisioned in real time. In these cases, memories can be unconscious for a time and then conscious. The content of the representations is accessible to "consciousness", but not the process of retrieval. The spotlight of "consciousness" may or may not shine on them, depending on the grain-size of the spatial or temporal coordinate systems.

CMattie does not currently have the spatial and temporal coordinate systems that are essential for planning and fleshing out mental images. Since she has no spatial senses, a spatial coordiante system, such as the "visual-spatial sketchpad" that is known to exist in the working memory of humans (Baddeley, 1992; Shah & Miyake, in press), is not applicable. However, CMattie must learn through dialogue with human organizers. This process will require a temporal coordinate system to track the order of messages in an interchange. Such a system is being designed at this writing (see Ramamurthy, Franklin & Negatu, in press), for a preliminary report).

(9) Currently unrehearsed items in working memory (p. 175). Information is preserved in working memory for a short period of time while it is not being rehearsed. The duration of this unrehearsed information in memory varies from 30 seconds in short-term memory (e.g., remembering a telephone number while you reach for a phone and dial it) to several minutes in

working memory. In the latter case, humans are expected to actively monitor two or more tasks simultaneously (e.g., driving and holding a conversation). According to research by Baddeley (1986, 1992) the contents of working memory can be actively maintained or recycled either through a "visual-spatial sketchpad", through an "articulatory loop", or through an "executive", so working memory does have a number of separate modalities. In the absence of the active recycling of the information in working memory (much of which involves consciousness), there is content that passively resides in working memory for a few seconds or minutes.

The CMattie architecture allows for several different working memories. In each case the spotlight of attention ("consciousness") can shine on individual items in a working memory while not shining on the others. For example suppose that the seminar announcement template occupies one working memory and the Complex Systems Seminar is under scrutiny. Attention could shine on missing information or on an anomaly, but never the missing information and an anomaly at the same time. Both the missing information and the anomaly would occupy working memory, but only one of these would be in "consciousness" at the same time. The contents of "consciousness" are in working memory, but the working memory contains additional content and is, therefore, not equivalent to "consciousness".

(10) Automatic mental images (p. 175). A mental image can fade from consciousness, yet continue to function subsequently in the processing stream. At one point in time, a mental image is constructed on where a restaurant is located and how to get to it. The image guides the path to the restaurant, even though the image has left consciousness while the person is engaged in conversation with a passenger in the car. CMattie allows an image to linger on well after it has exited consciousness. One example was discussed previously. A reminder being sent to the organizer would likely result in consciousness shifting elsewhere, with the template still in place

and the codelet who noted the gap still somewhat activated. At the same time, codelets recruited to deal with the missing situation would likely be both active and uncouscious. Another example is when consciousness shines on a perceptual gap, such as a missing seminar location. When the default place is found and inserted, processing would continue unconsciously.

(11) Contrasts that recruit attention (page 175). Attention is known to be captured by contrasts in the environment, such as light versus dark, loud versus silent, and motion versus rest. Our attention is captured by contrasts between our knowledge and what appears in the environment (such as an anomalous object or event). These contrasts in the environment automatically capture our attention when the contrasts are extreme, such as an explosion that occurs in the midst of silence. We have an "orientation reflex" that automatically turns to the source of extremely loud blasts; this is prewired in the organism, not a learned response. However, we can also voluntarily control our attention and this control can supercede attention being controlled by data-driven contrasts in the environment. Factory workers can voluntarily monitor their attention to ignore loud blasts.

CMattie can accommodate attention being controlled by the environment and by its goals. When there is an important event, such as a shutdown message, this data-driven input would capture attention and drive out the old contents of consciousness. However, for this to happen, these involuntary controls over attention would need to be in a class of high priority events. Regarding the voluntary control over attention, CMattie's attention will focus on a goal (behavior) as it becomes active as a result of the action of the behavior net. Actions that immediately follow are voluntary. CMattie's metacognitive mechanism can also send activation to a behavior, trying to cause a voluntary action, or to send activation to a particular coalition of codelets, bring it to "consciousness" and constituting voluntary control of attention.

(12) Attended versus unattended messages (page 175). Baars described attention experiments by Don Norman that involved dichotic listening. A person is presented a different message in each ear and is asked to attend to the message on only one channel (e.g., the left ear). The person can identify the voice quality of the unattended channel, but not the individual words even though they are repeated up to 35 times. The situation described in Norman's study can only occur in an agent with more than one sensory input channel. CMattie currently does not have multiple sensory channels, but it in principle could be expanded to have more than one channel. Other than the matter of there being different channels, this phenomenon is exactly the same as below threshold stimulation (see number 1). As discussed earlier, CMattie is capable of simulating the unconscious activation of processors.

Consider the case when two simultaneous messages are sent to CMattie. CMattie attends to and processes one message at a time, so CMattie will attend to message 1 without deeply processing message 2, and vice versa. However, there could be a residue of the "unconscious" activations of a privileged set of the codelets from the unattended message while the focal message is being processed. This is in fact necessary for processing a critical interrupting message, such as an impending shutdown from the systems operator. The metacognitive component is capable of reconstructing whether this residue of "unconscious" activation of privileged codelets is different from its base rate profile of activations. Any discrepancies will allow CMattie to reconstruct particular characteristics of the unattended message, such as the person who sent the message.

(13) Interruption of, and influence on, the attended stream (page 176). As discussed above, a critical message from the system operator can interrupt the process of CMattie's attending to an incoming email message. In this case, the interrupted messages can be attended to later, though

as in humans some information may be lost. A more critical incoming message, such as an announcement of an imminent shut down of the system, can jump the queue to be processed next. This is based on the "unconscious" perception of a privileged set of features, as discussed in 12. Once the urgent message is perceived, it interrupts (and takes precedence over) the further processing of the earlier message.

(14) Voluntary versus involuntary attention (page 176). The spotlight of consciousness may be constructed voluntarily, following an agenda of goals and drives. This occurs when a person drives his automobile along a dangerous mountain pass. Consciousness is directed and explores information that is relevant to the goals and drives. Alternatively, the spotlight of consciousness may be unexpectedly captured by an intense stimulus, such as the load roar of a nearby train. Thus, consciousness fluctuates between the continuum of being goal-driven and stimulus-driven.

CMattie also has its attention being driven by either goals or stimuli. When CMattie is trying to fill in a missing seminar location, these activities are goal-driven. In contrast, when an unexpected shut-down message occurs, there is a stimulus-driven recruitment of "consciousness".

(15) Dishabituation of the orienting response (page 176). Predictable stimuli are accommodated by unconscious mechanisms whereas unpredictable stimuli require consciousness. When a shutdown message first occurs, the spotlight of "consciousness" will be recruited by CMattie. However, when a shutdown message routinely occurs at the same time and same place, then this invariant feature will be acquired through CMattie's learning mechanisms. The templates and codelets will eventually be updated. At that point, the shutdown message will be handled by "unconscious" mechanisms.

The CM-architecture permits a mechanism for habituation as part of the controller of the spotlight of attention. If a speaker-topic message is perceived that lists the time of the Cognitive Science Seminar as 1:30 on Wednesday, the traditional time, the controller will likely ignore it entirely. The message meaning will routinely come to "consciousness", but the time will be ignored. If it lists 3:30 on Wednesday, a codelet will likely pick up the difference and bring the issue to "consciousness". A coalition with high activation will be formed of that codelet and other codelets carrying the information concerning that seminar. A message questioning the accuracy of the new time will also be sent to the organizer, to verify the unexpected deviation. Dishabituation will eventually be completed when the memories are updated.

(16) Most thinking during problem solving is inexplicit (page 176). Humans are not conscious of all stages and representations in problem solving. We are conscious of the beginning state (the representation of what the problem is), the goal state (what we hope to achieve by solving the problem), the landmark keys to the solution, and some of the salient intermediate states (Ericsson & Simon, 1980). However, we are not aware of the massive blur of incubation processes, of searches of large spaces, and of the hundreds of intermediate knowledge states in route to the solution.

In CMattie, the spotlight of "consciousness" drifts to the content that is affiliated with missing parts, contradictions, obstacles to goals, contrasts, anomalies, and other violations of expectations. "consciousness" basks in these challenges of the atypical input. These also are precisely the situations when problem solving occurs. However, the process of activating codelets and behaviors in the behavior network are "unconscious" and therefore not in the spotlight of "consciousness".

(17) Word retrieval and question answering (page 177). The answering of a difficult question is a special case of the problem solving scenario above. CMattie may well be "conscious" that the seminar name is missing from a particular speaker-topic message. That is, the spotlight is shining on a part of a template in the perceptual work space. Later, CMattie may be again "conscious" of the completely understood message without having been "conscious" of the process of retrieving the seminar name. Research on human question answering has revealed that the search of information is unconscious, fast, and executed in parallel, whereas the process of verifying that a fetched answer is correct is conscious, slow, and serial (Graesser, Lang, & Roberts, 1991).

(18) Recall from long-term memory (page 177). The opacity of memory retrieval is built in to the CM-architecture. The spotlight does not shine inside of sparse distributed memory, nor inside of the episodic memory. It may occasionally shine on some types of slipnet nodes that constitute a conceptual context that could become "conscious" during metacognitive reflection.

(19) Action planning and control (page 177). CMattie plans implicitly and unconsciously as a result of the structure and activity flow of her behavior net. For the most part, she does not plan explicitly or consciously. The spotlight of "consciousness" drifts to the content affiliated with the missing information, the contradictions, the obstacles to goals, the anomalies, and other violations of expectation that arise throughout the course of planning. CMattie is "conscious" of (a) what the problem or difficulty is, (b) occasionally of internal activities such as of an oscillation, (c) of her external actions and (d) of the eventual solution if one occurs. She's typically "unconscious" of intermediate steps.

(20) Perceptual reorganization (page 178). CMattie can be expected to exhibit precisely the "conscious"-"unconscious"-"conscious" pattern that occurs when we perceive a Necker cube and other ambiguous stimuli. The "consciousness"-"unconscious"-"consciousness"

stream of processing occurs when the spotlight recruits additional codelets or behaviors to help in the recognition and categorization of input. CMattie might oscillate between Dr. Garzon being the organizer of the seminar or the speaker on a particular day. CMattie might oscillate when interpreting 430 as a room number and the time of a seminar.

(21) Developing automaticity with practice in predictable tasks (page 178).

Though CMattie is capable of learning by adding particular slipnet nodes, codelets, behaviors, and solutions to particular problems (stored in episodic memory), these are one-trail learning episodes rather than learning with practice. CMattie does overlearn and automatize as a particular collection of codelets increases in their mutual associations to the extent that they become one of Jackson's concept codelets and are called to action as a single unit.

(22) Loss of conscious access to visual information that nonetheless continues to inform problem solving (page 179). There is nothing analogous to this in CMattie's architecture. Her two senses "see" only incoming email messages and operating system messages. Her mental images consist of parts of templates in working memory or in perceptual registers. These cannot be manipulated spatially, though they will be manipulated temporally. But, "unconscious" acts do inform problem solving in many ways (see item 4 above for an example). Also, it should not be hard to design and build a "conscious" software agent, capable of rotating visual images in this way, so that the rotation become "unconscious" after automaticity occurs, but its problem solving effect continues. The CM-architecture allows this in principle, but it is not currently implemented.

(23) Implicit learning of miniature grammars (page 179). Again, CMattie's domain does not allow for this phenomenon, but nothing in principle obstructs a "conscious" software agent from learning grammars implicitly.

(24) Capability contrasts (page 179). This is perforce an empirical matter to be decided when CMattie is up and running. We suspect all of the capabilities listed in the table will be found as advertised, but this remains to be seen.

Conclusion

So how well does CMattie's architecture account for Baars' collection of psychological facts that serve to constrain consciousness? How well does she implement global workspace theory? For current purposes we take these questions to be synonymous. Our conclusion is: quite well, but not perfectly. Almost all the psychological facts are accounted for. Of those that are not, most fail as a consequence of the choice of domain, for example, because CMattie has no visual sense. They do not, in principle, present difficulty for the architecture. The weakest link seems to be a not completely adequate performance in habituation and in acquiring automaticity. Still, our experience with CMattie as a conceptual model shows her to be a useful tool in thinking through cognitive issues.

Suppose we agree that the CMattie architecture implements global workspace theory as intended. Can we then conclude that the eventually computational implementation of CMattie, running on a UNIX system, will be "conscious" in the sense of being in some way sentient? We don't know, and know of no way of telling. However, we do believe that should any piece of software become sentient, it must be based on some such architecture that provides mechanisms for consciousness, and cannot simply depend on complexity to do the trick.

References

Anwar, Ashraf and Stan Franklin (forthcoming). Sparse Distributed Memory as a tool for "conscious" Cognitive Software Agents.

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

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

Baddeley, A.D. (1986). *Working memory.* New York: Oxford University Press.

Baddeley, A.D. (1992). Working memory. *Science,* **255**, 556-559.

Bogner, Myles, Uma Ramamurthy, and Stan Franklin (in press). "Consciousness" and Conceptual Learning in a Socially Situated Agent. in Kerstin Dautenhahn ed. Human Cognition and Social Agent Technology

Ercsson, K.A., & Simon, H.A. (1980). Verbal reports as data. *Psychological Review*. **87**, 215-251.

Franklin, Stan (1995). *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. (1997a). Global Workspace Agents. *Journal of Consciousness Studies.* **4** (4), 322-234.

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, Graesser, Art, Olde, B., Song, H., and Negatu, A. (1996). Virtual Mattie-an Intelligent Clerical Agent. AAAI Fall Symposium on Embodied AI.

Graesser, A.C., Lang, K.L., & Roberts, R.M. (1991). Question answering in the context of stories. *Journal of Experimental Psychology: General*. **120**, 254-277.

Graesser, A.C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative comperehension. *Psychological Review.* **101**, 371-395.

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

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.

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

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

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.

Kolodner, J.L. (1993). *Case-based reasoning*. Hillsdale, NJ: Erlbaum.

Kosslyn, S.M. (1994). *Image and brain*. Cambridge, MA: MIT Press.

Leung, K.S., and C. T. Lin (1988). Fuzzy concepts in expert systems. *Computer*. **21**(9):43-56

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

Maes, Pattie (1992). Learning Behavior Networks from Experience. *Proceedings of the First European Conference on Artificial Life*, Paris, December 1991, 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 Franklin, Stan (1998). An Architecture For Emotion. AAAI Fall Symposium on Emotional and Intelligent: The Tangled Knot of Cognition.

Mitchell, Melanie (1993). *Analogy-Making as Perception.* Cambridge MA: The MIT Press.

Mooney, R.J.(1990). *A general explanation-based learning mechanism and its application in narrative understanding.* San Mateo, CA: Morgan-Kaufman.

Ramamurthy, Uma, Franklin, Stan and Negatu, Aregahegn (in press) Learning Concepts in Software Agents. *From Animals to Animats IV* (Proc. SAB'98) Cambridge, MA: MIT Press.

Schank, R.C. (1986). *Explanation patterns: Understanding mechanically and creatively.* Hillsdale, NJ: Erlbaum.

Shah, P., & Miyake, A. (in press)(Eds.). *Working memory.* New York: Cambridge University Press.

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

Varela, F. J., Thompson, E., and Rosch, E. (1991). *The Embodied Mind.* Cambridge, MA: MIT Press.

Zhang, Zhaohua, Stan Franklin, Brent Olde, Yun Wan and Art Graesser (1998). Natural Language Sensing for Autonomous Agents. Proceedings IEEE International Joint Symposium on Intelligence and Systems, 374-381.

Zhang, Zhaohua, Stan Franklin and Dipankar Dasgupta (1998), Metacognition in Software Agents using Classifer Systems, Proc AAAI 98, 82-88