CHAPTER 7

Computational Models of Consciousness: A Taxonomy and Some Examples

Ron Sun and Stan Franklin

Abstract

This chapter aims to provide an overview of existing computational (mechanistic) models of cognition in relation to the study of consciousness, on the basis of psychological and philosophical theories and data. It examines various mechanistic explanations of consciousness in existing computational cognitive models. Serving as an example for the discussions, a computational model of the conscious/unconscious interaction, utilizing the representational difference explanation of consciousness, is described briefly. As a further example, a software agent model that captures another explanation of consciousness (the access explanation of consciousness) is also described. The discussions serve to highlight various possibilities in developing computational models of consciousness and in providing computational explanations of conscious and unconscious cognitive processes.

Introduction

In this chapter, we aim to present a short survey and a brief evaluation of existing computational (mechanistic) models of cognition in relation to the study of consciousness. The survey focuses on their explanations of the difference between conscious and unconscious cognitive processes on the basis of psychological and philosophical theories and data, as well as potential practical applications.

Given the plethora of models, theories, and data, we try to provide in this chapter an overall (and thus necessarily sketchy) examination of computational models of consciousness in relation to the available psychological data and theories, as well as the existing philosophical accounts. We come to some tentative conclusions as to what a plausible computational account should be like, synthesizing various operationalized psychological notions related to consciousness.

We begin by examining some foundational issues concerning computational approaches toward consciousness. Then, various existing models and their explanations of the conscious/unconscious distinction are presented. After examining a particular model embodying a two-system approach, we look at one embodying a unified (one-system) approach and then at a few additional models.

Computational Explanations of Consciousness

Work in the area of computational modeling of consciousness generally assumes the sufficiency and the necessity of mechanistic explanations. By mechanistic explanation, we mean any concrete computational processes, in the broadest sense of the term "computation." In general, computation is a broad term that can be used to denote any process that can be realized on generic computing devices, such as Turing machines (or even beyond if there is such a possibility). Thus, mechanistic explanations may utilize, in addition to standard computational notions, a variety of other conceptual constructs ranging, for example, from chaotic dynamics (Freeman, 1995), to "Darwinian" competition (Edelman, 1989), and to quantum mechanics (Penrose, 1994). (We leave out the issue of complexity for now.)

In terms of the *sufficiency* of mechanistic explanations, a general working hypothesis is succinctly expressed by the following statement (Jackendoff, 1987):

> *Hypothesis of computational sufficiency: every phenomenological distinction is caused by/supported by/projected from a corresponding computational distinction.*

For the lack of a clearly better alternative, this hypothesis remains a viable working hypothesis in the area of computational models of consciousness, despite various criticisms (e.g., Damasio, 1994; Edelman, 1989; Freeman, 1995; Penrose, 1994; Searle, 1980).

On the other hand, the necessity of mechanistic explanations, according to the foregoing definition of mechanistic processes, should be intuitively obvious to anyone who is not a dualist. If one accepts the universality of computation, then computation, in its broadest sense, can be expected to include the necessary conditions for consciousness.

On the basis of such intuition, we need to provide an explanation of the computational/mechanistic basis of consciousness that answers the following questions. What kind of mechanism leads to conscious processes, and what kind of mechanism leads to unconscious processes? What is the functional role of conscious processes (Baars, 1988, 2002; Sun, 1999a, b)? What is the functional role of unconscious processes? There have been many such explanations in computational or mechanistic terms. These computational or mechanistic explanations are highly relevant to the science of consciousness as they provide useful theoretical frameworks for further empirical work.

Another issue we need to address before we move on to details of computational work is the relation between biological/physiological models and computational models in general. The problem with biologically centered studies of consciousness in general is that the gap between phenomenology and physiology/biology is so great that something else may be needed to bridge it. Otherwise, if we rush directly into complex neurophysiological thickets (Edelman, 1989; Crick & Koch, 1990; Damasio et al., 1990; LeDoux, 1992,), we may lose sight of the forests. Computation, in its broadest sense, can serve to bridge the gap. It provides an intermediate level of explanation in terms of processes, mechanisms, and functions and helps determine how various aspects of conscious and unconscious processes should figure into the architecture of the mind (Anderson & Lebiere, 1998; Sun, 2002). It is possible that an intermediate level between phenomenology and physiology/neurobiology might be more apt to capture fundamental characteristics of consciousness (Coward & Sun, 2004). This notion of an intermediate level of explanation

has been variously expounded recently; for example, in terms of virtual machines by Sloman and Chrisley (2003).

Different Computational Accounts of Consciousness

Existing computational explanations of the conscious/unconscious distinction may be categorized based on the following different emphases: (1) differences in knowledge organization (e.g., the SN+PS view, to be detailed later), (2) differences in knowledge-processing mechanisms (e.g., the PS+SN view), (3) differences in knowledge content (e.g., the episode+activation view), (4) differences in knowledge representation (e.g., the localist+distributed view), or (5) different processing modes of the same system (e.g., the attractor view or the threshold view).

Contrary to some critics, the debate among these differing views is not analogous to a debate between algebraists and geometers in physics (which would be irrelevant). It is more analogous to the wave vs. particle debate in physics concerning the nature of light, which was truly substantive. Let us discuss some of the better known views concerning computational accounts of the conscious/unconscious distinction one by one.

First of all, some explanations are based on recognizing that there are two separate systems in the mind. The difference between the two systems can be explained in terms of differences in either knowledge organization, knowledge-processing mechanisms, knowledge content, or knowledge representation:

- *The SN+PS view*: an instance of the explanations based on differences in knowledge organization. As originally proposed by Anderson (1983) in his ACT* model, there are two types of knowledge: Declarative knowledge is represented by semantic networks (SN), and it is consciously accessible, whereas procedural knowledge is represented by rules in a production system (PS), and it is inaccessible. The difference lies in the two different ways of organizing knowledge – whether in an action-centered way (procedural knowledge) or in an action-independent way (declarative knowledge). Computationally, both types of knowledge are represented symbolically (using either symbolic semantic networks or symbolic production rules).[1] The semantic networks use parallel spreading activation (Collins & Loftus, 1975) to activate relevant nodes, and the production rules compete for control through parallel matching and firing. The models embodying this view have been used for modeling a variety of psychological tasks, especially skill learning tasks (Anderson, 1983, Anderson & Lebiere, 1998).
- *The PS+SN view*: an instance of the explanations based on differences in knowledge-processing mechanisms. As proposed by Hunt and Lansman (1986), the "deliberate" computational process of production matching and firing in a production system (PS), which is serial in this case, is assumed to be a conscious process, whereas the spreading activation computation (Collins & Loftus, 1975) in semantic networks (SN), which is massively parallel, is assumed to be an unconscious process. The model based on this view has been used to model controlled and automatic processing data in the attention-performance literature (Hunt & Lansman, 1986). Note that this view is the exact opposite of the view advocated by Anderson (1983), in terms of the roles of the two computational mechanisms involved. Note also that the emphasis in this view is on the processing difference of the two mechanisms, serial vs. parallel, and not on knowledge organization.
- *The algorithm + instance view*: another instance of the explanations based on differences in knowledge-processing mechanisms. As proposed by Logan (1988) and also by Stanley et al. (1989), the computation involved in retrieval and use of instances of past experience is considered to be unconscious (Stanley

et al., 1989) or automatic (Logan 1988), whereas the use of "algorithms" involves conscious awareness. Here the term "algorithm" is not clearly defined and apparently refers to computation more complex than instance retrieval/use. Computationally, it was suggested that the use of an algorithm is under tight control and carried out in a serial, step-by-step way, whereas instances can be retrieved in parallel and effortlessly (Logan, 1988). The emphasis here is again on the differences in processing mechanisms. This view is also similar to the view advocated by Neal and Hesketh (1997), which emphasizes the unconscious influence of what they called episodic memory. Note that the views by Logan (1988), Stanley et al. (1989), and Neal and Hesketh (1997) are the exact opposite of the view advocated by Anderson (1983) and Bower (1996), in which instances/episodes are consciously accessed rather than unconsciously accessed.

- *The episode+activation view*: an instance of the explanations based on differences in knowledge content. As proposed by Bower (1996), unconscious processes are based on activation propagation through strengths or weights (e.g., in a connectionist fashion) between different nodes representing perceptual or conceptual primitives, whereas conscious processes are based on explicit episodic memory of past episodes. What is emphasized in this view is the rich spatial-temporal context in episodic memory (i.e., the ad hoc associations with contextual information, acquired on an one-shot basis), which is termed type-2 associations as opposed to regular type-1 associations (which are based on semantic relatedness). This emphasis somewhat distinguishes this view from other views concerning instances/episodes (Logan, 1988; Neal & Hesketh, 1997; Stanley et al. 1989).[2] The reliance on memory of specific events in this view bears some resemblance to some neurobiologically motivated views that rely on the interplay of various memory systems, such as that advocated by Taylor (1997) and McClelland et al. (1995).
- *The localist+distributed representation view*: an instance of the explanations based on differences in knowledge representation. As proposed in Sun (1994, 2002), different representational forms used in different components may be used to explain the qualitative difference between conscious and unconscious processes. One type of representation is symbolic or localist, in which one distinct entity (e.g., a node in a connectionist model) represents a concept. The other type of representation is distributed, in which a non-exclusive set of entities (e.g., a set of nodes in a connectionist model) are used for representing one concept, and the representations of different concepts overlap each other; in other words, a concept is represented as a pattern of activations over a set of entities (e.g., a set of nodes). Conceptual structures (e.g., rules) can be implemented in the localist/symbolic system in a straightforward way by connections between relevant entities. In distributed representations, such structures (including rules) are diffusely duplicated in a way consistent with the meanings of the structures (Sun, 1994), which captures unconscious performance. There may be various connections between corresponding representations across the two systems. (A system embodying this view, CLARION, is described later.)

In contrast to these two-systems views, there exist some theoretical views that insist on the unitary nature of the conscious and the unconscious. That is, they hold that conscious and unconscious processes are different manifestations of the same underlying system. The difference between conscious and unconscious processes lies in the different processing modes for conscious versus unconscious information within the same

system. There are several possibilities in this regard:

- *The threshold view*: As proposed by various researchers, including Bowers et al. (1990), the difference between conscious and unconscious processes can be explained by the difference between activations of mental representations above a certain threshold and activations of such representations below that threshold. When activations reach the threshold level, an individual becomes aware of the content of the activated representations; otherwise, although the activated representations may influence behavior, they will not be accessible consciously.
- *The chunking view*: As in the models described by Servan-Schreiber and Anderson (1987) and by Rosenbloom et al. (1993), a chunk is considered a unitary representation and its internal working is oblique (although its input/output are accessible). A chunk can be a production rule (as in Rosenbloom et al., 1993) or a short sequence of perceptual-motor elements (as in Servan-Schreiber & Anderson, 1987). Because of the lack of transparency of the internal working of a chunk, it is equated with implicit learning (Servan-Schreiber & Anderson, 1987) or automaticity (Rosenbloom et al., 1993). According to this view, the difference between conscious and unconscious processes is the difference between using multiple (simple) chunks (involving some consciousness) and using one (complex) chunk (involving no consciousness).
- *The attractor view*: As suggested by the model of Mathis and Mozer (1996), being in a stable attractor of a dynamic system (a neural network in particular) leads to consciousness. The distinction between conscious and unconscious processes is reduced to the distinction of being in a stable attractor and being in a transient state. O'Brien and Opie (1998) proposed an essentially similar view. This view may be generalized to a general coherence view – the emphasis may be placed on the role of internal consistency in producing consciousness. There has been support for this possibility from neuroscience, for example, in terms of a coherent "thalamocortical core" (Edelman & Tononi, 2000).
- *The access view*: As suggested by Baars (1988), consciousness is believed to help mobilize and integrate mental functions that are otherwise disparate and independent. Thus, consciousness is aimed at solving the relevance problem – finding the exact internal resources needed to deal with the current situation. Some evidence has been accumulated for this view (Baars, 2002). A computational implementation of Baars' theory in the form of IDA (a running software agent system; Franklin et al., 1998) is described in detail later. See also Coward and Sun (2004).

The coexistence of these various views of consciousness seems quite analogous to the parable of the Blind Men and the Elephant. Each of them captures some aspect of the truth about consciousness, but the portion of the truth captured is limited by the view itself. None seems to capture the whole picture.

In the next two sections, we look into some details of two representative computational models, exemplifying either two-system or one-system views. The models illustrate what a plausible computational model of consciousness should be like, synthesizing various psychological notions and relating to various available psychological theories.

A Model Adopting the Representational Difference View

Let us look into the representational difference view as embodied in the cognitive architecture CLARION (which stands for Connectionist Learning with Rule Induction ON-line; Sun 1997, 2002, 2003), as an example of the two-system views for explaining consciousness.

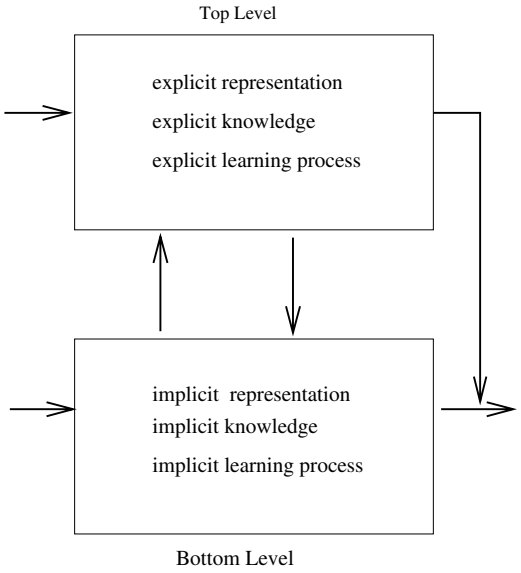

Figure 7.1. The CLARION model.

The important premises of subsequent discussions are the *direct accessibility* of conscious processes and the *direct inaccessibility* of unconscious processes. Conscious processes should be *directly* accessible – that is, directly verbally expressible – without involving intermediate interpretive or transformational steps, which is a requirement prescribed and/or accepted by many theoreticians (see, e.g., Clark, 1992; Hadley, 1995).[3] Unconscious processes should be, in contrast, inaccessible directly (but they might be accessed indirectly through some interpretive processes), thus exhibiting different psychological properties (see, e.g., Berry & Broadbent, 1988; Reber, 1989; more discussions later).

An example model in this regard is CLARION, which is a two-level model that uses the localist and distributed representations in the two levels, respectively, and learns using two different methods in the two levels, respectively. In developing the model, four criteria were hypothesized (see Sun, 1994), on the basis of the aforementioned considerations: (1) direct accessibility of conscious processes; (2) direct inaccessibility of unconscious processes; and furthermore, (3) linkages from localist concepts to distributed features: once a localist concept is activated, its corresponding distributed representations (features) are also activated, as assumed in most cognitive models, ranging from Tversky (1977) to Sun (1995);[4] and (4) linkages from distributed features to localist concepts: under appropriate circumstances, once some or most of the distributed features of a concept are activated, the localist concept itself can be activated to "cover" these features (roughly corresponding to categorization; Smith & Medin, 1981).

The direct inaccessibility of unconscious knowledge can be best captured by a "subsymbolic" distributed representation such as that provided by a backpropagation network (Rumelhart et al., 1986), because representational units in a distributed representation

Dimensions	bottom	top
Cognitive phenomena	implicit learning	explicit learning
	implicit memory	explicit memory
	automatic processing	controlled processing
	intuition	explicit reasoning
Source of knowledge	trial-and-error	external sources
	assimilation of explicit knowledge	extraction from the bottom level
Representation	distributed (micro)features	localist conceptual units
Operation	similarity-based	explicit symbol manipulation
Characteristics	more context sensitive, fuzzy	more crisp, precise
	less selective	more selective
	more complex	simpler

Figure 7.2. Comparisons of the two levels of the CLARION architecture.

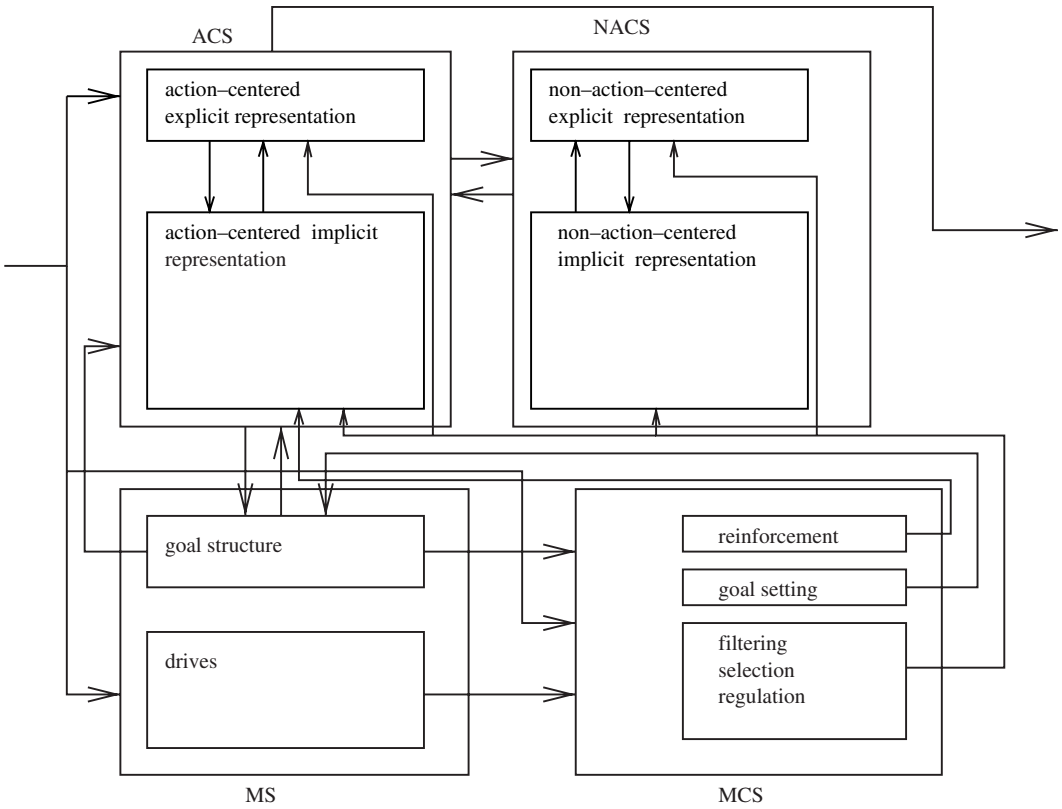

Figure 7.3. The implementation of CLARION. ACS denotes the action-centered subsystem, NACS the non-action-centered subsystem, MS the motivational subsystem, and MCS the metacognitive subsystem. The top level contains localist encoding of concepts and rules. The bottom level contains multiple (modular) connectionist networks for capturing unconscious processes. The interaction of the two levels and the information flows are indicated with arrows.

are capable of accomplishing tasks but are generally uninterpretable directly (see Rumelhart et al., 1986; Sun, 1994). In contrast, conscious knowledge can be captured in computational modeling by a symbolic or localist representation (Clark & Karmiloff-Smith, 1993; Sun & Bookman 1994), in which each unit has a clear conceptual meaning/interpretation (i.e., a semantic label). This captures the property of conscious processes being directly accessible and manipulable (Smolensky, 1988; Sun, 1994). This difference in representation leads to a two-level structure whereby each level uses one type of representation (Sun, 1994, 1995, 1997; Sun et al., 1996, 1998, 2001). The bottom level is based on distributed representation, whereas the top level is based on localist/symbolic representation. For learning, the bottom level uses gradual weight tuning, whereas the top level uses explicit, one-shot hypothesis testing learning, in correspondence with the representational characteristics of the two levels. There are various connections across the two levels for exerting mutual influences. See Figure 7.1 for an abstract sketch of the model. The different characteristics of the two levels are summarized in Figure 7.2.

Let us look into some implementational details of CLARION. Note that the details of the model have been described extensively in a series of previous papers, including Sun (1997, 2002, 2003), Sun and Peterson (1998), and Sun et al. (1998, 2001). It has a dual representational structure – implicit and explicit representations being in two separate "levels" (Hadley, 1995; Seger,

1994). Essentially it is a dual-process theory of mind (Chaiken & Trope, 1999). It also consists of a number of functional subsystems, including the action-centered subsystem, the non-action-centered subsystem, the metacognitive subsystem, and the motivational subsystem (see Figure 7.3).

Let us first focus on the action-centered subsystem of CLARION. In this subsystem, the two levels interact by cooperating in actions, through a combination of the action recommendations from the two levels, respectively, as well as cooperating in learning through a bottom-up and a top-down process (to be discussed below). Actions and learning of the action-centered subsystem may be described as follows:

1. Observe the current state x.
2. Compute in the bottom level the "values" of x associated with each of all the possible actions a_i's: $Q(x, a_1)$, $Q(x, a_2)$, ... , $Q(x, a_n)$ (to be explained below).
3. Find out all the possible actions (b_1, b_2, ... , b_m) at the top level, based on the input x (sent up from the bottom level) and the rules in place.
4. Compare or combine the values of the a_is with those of b_js (sent down from the top level), and choose an appropriate action b.
5. Perform the action b, and observe the next state y and (possibly) the reinforcement r.
6. Update Q-values at the bottom level in accordance with the *Q-Learning-Backpropagation* algorithm (to be explained later).
7. Update the rule network at the top level using the *Rule-Extraction-Refinement* algorithm (to be explained later).
8. Go back to Step 1.

In the bottom level of the action-centered subsystem, implicit reactive routines are learned: A Q-value is an evaluation of the "quality" of an action in a given state: $Q(x, a)$ indicates how desirable action a is in state x (which consists of some sensory input). The agent may choose an action in any state based on Q-values (for example, by choosing the action with the highest Q-value). To acquire the Q-values, one may use the *Q-learning* algorithm (Watkins 1989), a reinforcement learning algorithm. It basically compares the values of successive actions and adjusts an evaluation function on that basis. It thereby develops reactive sequential behaviors.

The bottom level of the action-centered subsystem is modular; that is, a number of small neural networks coexist, each of which is adapted to specific modalities, tasks, or groups of input stimuli. This coincides with the modularity claim (Baars, 1988; Cosmides & Tooby, 1994; Edelman, 1987; Fodor, 1983; Hirschfield & Gelman, 1994; Karmiloff-Smith, 1986) that much processing in the human mind is done by limited, encapsulated (to some extent), specialized processors that are highly effcient. Some of these modules are formed evolutionarily; that is, given a priori to agents, reflecting their hardwired instincts and propensities (Hirschfield & Gelman, 1994). Some of them can be learned through interacting with the world (computationally through various decomposition methods; e.g., Sun & Peterson, 1999).

In the top level of the action-centered subsystem, explicit conceptual knowledge is captured in the form of rules. Symbolic/localist representations are used. See Sun (2003) for further details of encoding (they are not directly relevant here).

Humans are clearly able to learn implicit knowledge through trial and error, without necessarily utilizing a priori explicit knowledge (Seger, 1994). On top of that, explicit knowledge can be acquired, also from ongoing experience in the world, and possibly through the mediation of implicit knowledge (i.e., bottom-up learning; see Karmiloff-Smith, 1986; Stanley et al., 1989; Sun, 1997, 2002; Willingham et al., 1989). The basic process of bottom-up learning is as follows (Sun, 2002). If an action decided by the bottom level is successful, then the agent extracts a rule that corresponds to the action selected by the bottom level and adds the rule to the top level. Then, in subsequent interaction with the world, the agent verifies the extracted rule by considering the

outcome of applying the rule: If the outcome is not successful, then the rule should be made more specific and exclusive of the current case, and if the outcome is successful, the agent may try to generalize the rule to make it more universal (e.g., Michalski, 1983). The details of the bottom-up learning algorithm (the Rule-Extraction-Refinement algorithm) can be found in Sun and Peterson (1998). After rules have been learned, a variety of explicit reasoning methods may be used. Learning explicit conceptual representation at the top level can also be useful in enhancing learning of implicit reactive routines (reinforcement learning) at the bottom level.

Although CLARION can learn even when no a priori or externally provided knowledge is available, it can make use of it when such knowledge is available (cf. Anderson, 1983; Schneider & Oliver, 1991). To deal with instructed learning, externally provided knowledge (in the forms of explicit conceptual structures, such as rules, plans, routines, categories, and so on) should (1) be combined with autonomously generated conceptual structures at the top level (i.e., internalization) and (2) be assimilated into implicit reactive routines at the bottom level (i.e., assimilation). This process is known as top-down learning. See Sun (2003) for further details.

The non-action-centered subsystem represents general knowledge about the world, which is equivalent to the notion of semantic memory (as in, e.g., Quillian, 1968). It may be used for performing various kinds of retrievals and inferences. It is under the control of the action-centered subsystem (through the actions of the action-centered subsystem). At the bottom level, associative memory networks encode non-action-centered implicit knowledge. Associations are formed by mapping an input to an output. The regular backpropagation learning algorithm can be used to establish such associations between pairs of input and output (Rumelhart et al., 1986).

On the other hand, at the top level of the non-action-centered subsystem, a general knowledge store encodes explicit non-action-centered knowledge (Sun, 1994). In this network, chunks are specified through dimensional values. A node is set up at the top level to represent a chunk. The chunk node (a symbolic representation) connects to its corresponding features (dimension-value pairs) represented as nodes in the bottom level (which form a distributed representation). Additionally, links between chunks at the top level encode explicit associations between pairs of chunks, known as associative rules. Explicit associative rules may be formed (i.e., learned) in a variety of ways (Sun, 2003).

On top of associative rules, similarity-based reasoning may be employed in the non-action-centered subsystem. During reasoning, a known (given or inferred) chunk may be automatically compared with another chunk. If the similarity between them is sufficiently high, then the latter chunk is inferred (see Sun, 2003, for details). Similarity-based and rule-based reasoning can be intermixed. As a result of mixing similarity-based and rule-based reasoning, complex patterns of reasoning emerge. As shown by Sun (1994), different sequences of mixed similarity-based and rule-based reasoning capture essential patterns of human everyday (mundane, common-sense) reasoning.

As in the action-centered subsystem, top-down or bottom-up learning may take place in the non-action-centered subsystem, either to extract explicit knowledge in the top level from the implicit knowledge in the bottom level or to assimilate explicit knowledge of the top level into implicit knowledge in the bottom level.

The motivational subsystem is concerned with drives and their interactions (Toates, 1986). It is concerned with why an agent does what it does. Simply saying that an agent chooses actions to maximizes gains, rewards, or payoffs leaves open the question of what determines these things. The relevance of the motivational subsystem to the action-centered subsystem lies primarily in the fact that it provides the context in which the goal and the payoff of the action-centered subsystem are set. It thereby influences

the working of the action-centered subsystem and, by extension, the working of the non-action-centered subsystem.

A bipartite system of motivational representation is again in place in CLARION. The explicit goals (such as "finding food") of an agent (which is tied to the working of the action-centered subsystem) may be generated based on internal drive states (for example, "being hungry"). See Sun (2003) for details.

Beyond low-level drives concerning physiological needs, there are also higher-level drives. Some of them are primary, in the sense of being "hardwired." For example, Maslow (1987) developed a set of these drives in the form of a "need hierarchy." Whereas primary drives are built-in and relatively unalterable, there are also "derived" drives, which are secondary, changeable, and acquired mostly in the process of satisfying primary drives.

The metacognitive subsystem is closely tied to the motivational subsystem. The metacognitive subsystem monitors, controls, and regulates cognitive processes for the sake of improving cognitive performance (Nelson, 1993; Sloman & Chrisley, 2003; Smith et al., 2003). Control and regulation may be in the forms of setting goals for the action-centered subsystem, setting essential parameters of the action-centered and the non-action-centered subsystem, interrupting and changing ongoing processes in the action-centered and the non-action-centered subsystem, and so on. Control and regulation may also be carried out through setting reinforcement functions for the action-centered subsystem on the basis of drive states. The metacognitive subsystem is also made up of two levels: the top level (explicit) and the bottom level (implicit).

Note that in CLARION, there are thus a variety of memories: procedural memory (in the action-centered subsystem) in both implicit and explicit forms, general "semantic" memory (in the non-action-centered subsystem) in both implicit and explicit forms, episodic memory (in the non-action-centered subsystem), working memory (in the action-centered subsystem), goal structures (in the action-centered subsystem), and so on. See Sun (2003) for further details of these memories. As touched upon before, these memories are important for accounting for various forms of conscious and unconscious processes (also see, e.g., McClelland et al., 1995; Schacter, 1990; Taylor, 1997).

CLARION has been successful in accounting for a variety of psychological data. A number of well-known skill learning tasks have been simulated using CLARION;these span the spectrum ranging from simple reactive skills to complex cognitive skills. The tasks include serial reaction time (SRT) tasks, artificial grammar learning (AGL) tasks, process control (PC) tasks, the categorical inference (CI) task, the alphabetical arithmetic (AA) task, and the Tower of Hanoi (TOH) task (see Sun, 2002). Among them, SRT, AGL, and PC are typical implicit learning tasks, very much relevant to the issue of consciousness as they operationalize the notion of consciousness in the context of psychological experiments (Coward & Sun, 2004; Reber, 1989; Seger, 1994; Sun et al., 2005), whereas TOH and AA are typical high-level cognitive skill acquisition tasks. In addition, extensive work have been done on a complex minefield navigation task (see Sun & Peterson, 1998; Sun et al., 2001). Metacognitive and motivational simulations have also been undertaken, as have social simulation tasks (e.g., Sun & Naveh, 2004).

In evaluating the contribution of CLARION to our understanding of consciousness, we note that the simulations using CLARION provide detailed, process-based interpretations of experimental data related to consciousness, in the context of a broadly scoped cognitive architecture and a unified theory of cognition. Such interpretations are important for a precise, process-based understanding of consciousness and other aspects of cognition, leading to better appreciations of the role of consciousness in human cognition (Sun, 1999a). CLARION also makes quantitative and qualitative predictions regarding cognition in the areas of memory, learning, motivation, metacognition, and so on. These predictions either

have been experimentally tested already or are in the process of being tested (see, e.g., Sun, 2002; Sun et al., 2001, 2005). Because of the complex structures and their complex interactions specified within the framework of CLARION, it has a lot to say about the roles that different types of processes, conscious or unconscious, play in human cognition, as well as their synergy (Sun et al., 2005).

Comparing CLARION with Bower (1996), the latter may be viewed as a special case of CLARION for dealing specifically with implicit memory phenomena. The type-1 and type-2 connections, hypothesized by Bower (1996) as the main explanatory constructs, can be equated roughly to top-level representations and bottom-level representations, respectively. In addition to making the distinction between type-1 and type-2 connections, Bower (1996) also endeavored to specify the details of multiple pathways of spreading activation in the bottom level. These pathways were phonological, orthographical, semantic, and other connections that store long-term implicit knowledge. In the top level, associated with type-2 connections, it was claimed on the other hand that rich contextual information was stored. These details nicely complement the specification of CLARION and can thus be incorporated into the model.

The proposal by McClelland et al. (1995) that there are complementary learning systems in the hippocampus and neocortex is also relevant here. According to their account, cortical systems learn slowly, and the learning of new information destroys the old, unless the learning of new information is interleaved with ongoing exposure to the old information. To resolve these two problems, new information is initially stored in the hippocampus, an explicit memory system, in which crisp, explicit representations are used to minimize interference of information (so that catastrophic interference is avoided there). It allows rapid learning of new material. Then, the new information stored in the hippocampus is assimilated into cortical systems. The assimilation is interleaved with the assimilation of all other information in the hippocampus and with the ongoing events. Weights are adjusted by a small amount after each experience, so that the overall direction of weight change is governed by the structure present in the ensemble of events and experiences, using distributed representations (with weights). Therefore, catastrophic interference is avoided in cortical systems. This model is very similar to the two-level idea of CLARION, in that it not only adopts a two-system view but also utilizes representational differences between the two systems. However, in contrast to this model, which captures only what may be termed top-down learning (that is, learning that proceeds from the conscious to the unconscious), CLARION can capture both top-down learning (from the top level to the bottom level) and bottom-up learning (from the bottom level to the top level). See Sun et al. (2001) and Sun (2002) for details of bottom-up learning.

Turning to the declarative/procedural knowledge models, ACT* (Anderson, 1983) is made up of a semantic network (for declarative knowledge) and a production system (for procedural knowledge). ACT-R is a descendant of ACT*, in which procedural learning is limited to production formation through mimicking, and production firing is based on log odds of success. CLARION succeeds in explaining two issues that ACT did not address. First, whereas ACT takes a mostly top-down approach toward learning (i.e, from given declarative knowledge to procedural knowledge), CLARION can proceed bottom-up. Thus, Clarion can account for implicit learning better than ACT (see Sun, 2002, for details). Second, in ACT both types of knowledge are represented in explicit, symbolic forms (i.e., semantic networks and productions), and thus it does not explain, from a representational viewpoint, the differences in conscious accessibility (Sun, 1999b). CLARION accounts for this difference based on the use of two different forms of representation. Top-level knowledge is represented explicitly and thus consciously accessible, whereas bottom-level knowledge is represented implicitly and

thus inaccessible. Thus, this distinction in CLARION is intrinsic, instead of assumed as in ACT (Sun, 1999b).

Comparing CLARION with Hunt and Lansman's (1986) model, there are similarities. The production system in Hunt and Lansman's model clearly resembles the top level in CLARION, in that both use explicit manipulations in much the same way. Likewise, the spreading activation in the semantic network in Hunt and Lansman's model resembles the connectionist network in the bottom level of CLARION, because the same kind of spreading activation was used in both models, although the representation in Hunt and Lansman's model was symbolic, not distributed. Because of the uniformly symbolic representations used in Hunt and Lansman's model, it does not explain convincingly the qualitative difference between conscious and unconscious processes (see Sun, 1999b).

An Application of the Access View

Let us now examine an application of the access view on consciousness in building a practically useful system. The access view is a rather popular approach in computational accounts of consciousness (Baars, 2002), and therefore it deserves some attention. It is also presented here as an example of various one-system views.

Most computational models of cognitive processes are designed to predict experimental data. IDA (Intelligent Distribution Agent), in contrast, models consciousness in the form of an autonomous software agent (Franklin & Graesser, 1997). Specifically, IDA was developed for Navy applications (Franklin et al., 1998). At the end of each sailor's tour of duty, he or she is assigned to a new billet in a process called distribution. The Navy employs almost 300 people (called detailers) to effect these new assignments. IDA's task is to play the role of a detailer.

Designing IDA presents both communication problems and action selection problems involving constraint satisfaction. It must communicate with sailors via e-mail and in English, understanding the content and producing human-like responses. It must access a number of existing Navy databases, again understanding the content. It must see that the Navy's needs are satisfied while adhering to Navy policies. For example, a particular ship may require a certain number of sonar technicians with the requisite types of training. It must hold down moving costs. And it must cater to the needs and desires of the sailor as well as possible. This includes negotiating with the sailor via an e-mail correspondence in natural language. Finally, it must authorize the finally selected new billet and start the writing of the sailor's orders.

Although the IDA model was not initially developed to reproduce experimental data, it is nonetheless based on psychological and neurobiological theories of consciousness and does generate hypotheses and qualitative predictions (Baars & Franklin, 2003: Franklin et al., 2005). IDA successfully implements much of the global workspace theory (Baars, 1988), and there is a growing body of empirical evidence supporting that theory (Baars, 2002). IDA's flexible cognitive cycle has also been used to analyze the relation of consciousness to working memory at a fine level of detail, offering explanations of such classical working memory tasks as visual imagery to gain information and the rehearsal of a telephone number (Baars & Franklin, 2003: Franklin et al., 2005).

In his global workspace theory (see Figure 7.4), Baars (1988) postulates that human cognition is implemented by a multitude of relatively small, special-purpose processors, which are almost always unconscious (i.e., the modularity hypothesis as discussed earlier). Communication between them is rare and over a narrow bandwidth. Coalitions of such processes find their way into a global workspace (and thereby 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.

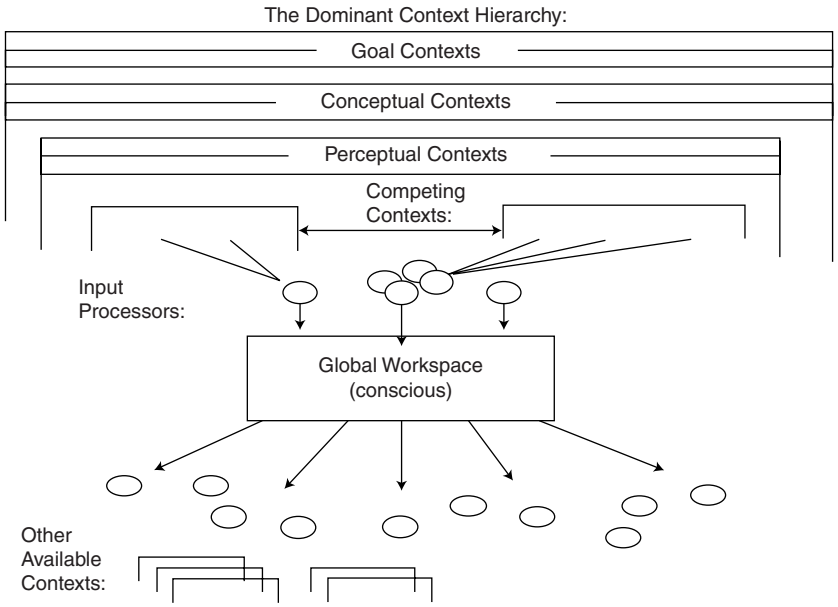

Figure 7.4. Baars' global workspace theory.

Thus consciousness, in this theory, allows us to deal with novel or problematic situations that cannot be dealt with effciently, or at all, by habituated unconscious processes. In particular, it provides access to appropriately useful resources. Global workspace theory offers an explanation for the limited capacity of consciousness. Large messages would be overwhelming to tiny processors. In addition, all activities of these processors take place under the auspices of contexts: goal contexts, perceptual contexts, conceptual contexts, and/or cultural contexts. Though contexts are typically unconscious, they strongly influence conscious processes.

Let us look into some details of the IDA architecture and its main mechanisms. At the higher level, the IDA architecture is modular with module names borrowed from psychology (see Figure 7.5). There are modules for Perception, Working Memory, Autobiographical Memory, Transient Episodic Memory, Consciousness, Action Selection, Constraint Satisfaction, Language Generation, and Deliberation.

In the lower level of IDA, the processors postulated by the global workspace theory are implemented by "codelets." Codelets are small pieces of code running as independent threads, each of which is specialized for some relatively simple task. They often play the role of "demons,"[5] waiting for a particular situation to occur in response to which they should act. Codelets also correspond more or less to Edelman's neuronal groups (Edelman, 1987) or Minsky's agents (Minsky, 1985). Codelets come in a number of varieties, each with different functions to perform. Most of these codelets subserve some high-level entity, such as a behavior. However, some codelets work on their own, performing such tasks as watching for incoming e-mail and instantiating goal structures. An important type of codelet that works on its own is the attention codelets that serve to bring information to "consciousness."

IDA senses only strings of characters, which are not imbued with meaning but which correspond to primitive sensations, like, for example, the patterns of activity on the rods and cones of the retina. These strings may come from e-mail messages, an operating system message, or from a database record.

The perception module employs analysis of surface features for natural-language understanding. It partially implements perceptual symbol system theory (Barsalou, 1999); perceptual symbols serve as a uniform

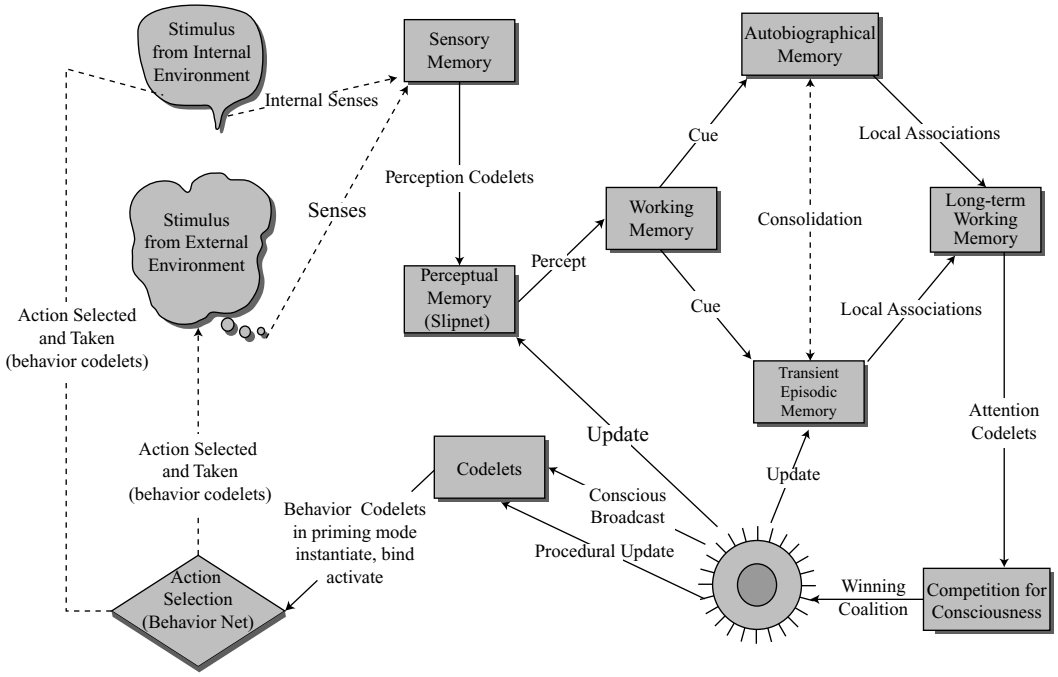

Figure 7.5. IDA's cognitive cycle.

system of representations throughout the system. Its underlying mechanism constitutes a portion of the Copycat architecture (Hofstadter & Mitchell, 1994). IDA's perceptual memory takes the form of a semantic net with activation passing, called the slipnet (see Figure 7.6). The slipnet embodies the perceptual contexts and some conceptual contexts from the global workspace theory. Nodes of the slipnet constitute the agent's perceptual symbols. Perceptual codelets recognize various features of the incoming stimulus; that is, various concepts. Perceptual codelets descend on an incoming message, looking for words or phrases they recognize. When such are found, appropriate nodes in the slipnet are activated. This activation passes around the net until it settles. A node (or several) is selected by its high activation, and the appropriate template(s) is filled by codelets with selected items from the message. The information thus created from the incoming message is then written to the workspace (working memory, to be described below), making it available to the rest of the system.

The results of this process, information created by the agent for its own use, are written to the workspace (working memory, not to be confused with Baars' global workspace). (Almost all of IDA's modules either write to the workspace, read from it, or both.)

IDA employs sparse distributed memory (SDM) as its major associative memory (Anwar & Franklin, 2003; Kanerva, 1988). SDM is a content-addressable memory. Being content addressable means that items in memory can be retrieved by using part of their contents as a cue, rather than having to know the item's address in memory.

Reads and writes, to and from associative memory, are accomplished through a gateway within the workspace called the focus. When any item is written to the workspace, another copy is written to the read registers of the focus. The contents of these read registers of the focus are then used as an address to query associative memory. The results of this query – that is, whatever IDA associates with this incoming information – are written into their own registers in the focus.

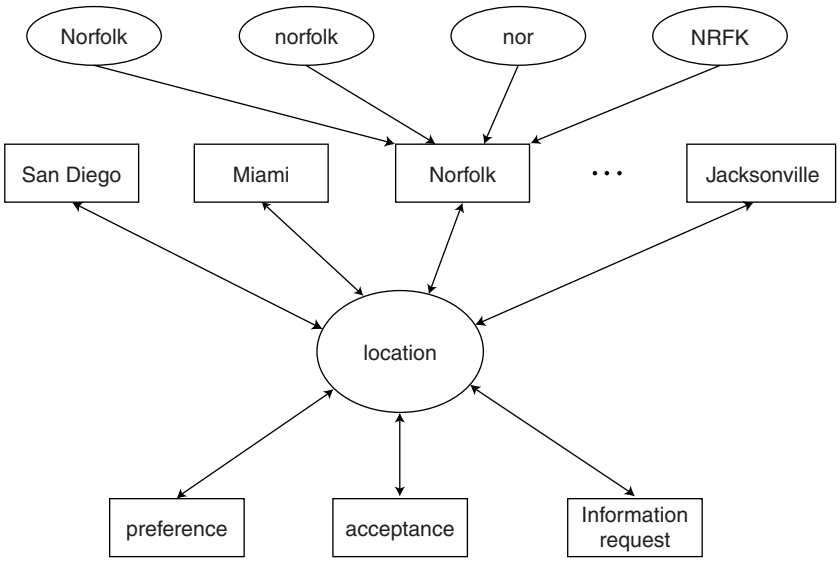

Figure 7.6. A portion of the slipnet in IDA.

This may include some emotion and some action previously taken. Thus associations with any incoming information, either from the outside world or from some part of IDA itself, are immediately available. (Writes to associative memory are made later and are described below.)

In addition to long-term memory, IDA includes a transient episodic memory (Ramamurthy, D'Mello, & Franklin, 2004). Long-term, content-addressable, associative memories are not typically capable of retrieving details of the latest of a long sequence of quite similar events (e.g., where I parked in the parking garage this morning or what I had for lunch yesterday). The distinguishing details of such events tend to blur due to interference from similar events. In IDA, this problem is solved by the addition of a transient episodic memory implemented with a sparse distributed memory. This SDM decays so that past sequences of similar events no longer interfere with the latest such events.

The apparatus for producing "consciousness" consists of a coalition manager, a spotlight controller, a broadcast manager, and a collection of attention codelets that recognize novel or problematic situations. Attention codelets have the task of bringing information to "consciousness." Each attention codelet keeps a watchful eye out for some particular situation to occur that might call for "conscious" intervention. Upon encountering such a situation, the appropriate attention codelet will be associated with the small number of information codelets that carry the information describing the situation. This association should lead to the collection of this small number of codelets, together with the attention codelet that collected them, becoming a coalition. Codelets also have activations. The attention codelet increases its activation in proportion to how well the current situation fits its particular interest, so that the coalition might compete for "consciousness," if one is formed.

In IDA, the coalition manager is responsible for forming and tracking coalitions of codelets. Such coalitions are initiated on the basis of the mutual associations between the member codelets. At any given time, one of these coalitions finds it way to "consciousness," chosen by the spotlight controller, which picks the coalition with the highest average activation among its member codelets. Baars' global workspace theory calls for the contents of "consciousness"

to be broadcast to each of the codelets in the system, and in particular, to the behavior codelets. The broadcast manager accomplishes this task.

IDA depends on the idea of a behavior net (Maes, 1989; Negatu & Franklin, 2002) for high-level action selection in the service of built-in drives. It has several distinct drives operating in parallel, and these drives vary in urgency as time passes and the environment changes. A behavior net is composed of behaviors and their various links. A behavior has preconditions as well as additions and deletions. A behavior also has an activation, a number intended to measure the behavior's relevance to both the current environment (external and internal) and its ability to help satisfy the various drives it serves.

The activation comes from activation stored in the behaviors themselves, from the external environment, from drives, and from internal states. The environment awards activation to a behavior for each of its true preconditions. The more relevant it is to the current situation, the more activation it receives from the environment. (This source of activation tends to make the system opportunistic.) Each drive awards activation to every behavior that, by being active, will help satisfy that drive. This source of activation tends to make the system goal directed. Certain internal states of the agent can also send activation to the behavior net. This activation, for example, might come from a coalition of codelets responding to a "conscious" broadcast. Finally, activation spreads from behavior to behavior along links.

IDA's behavior net acts in consort with its "consciousness" mechanism to select actions (Negatu & Franklin, 2002). Suppose some piece of information is written to the workspace by perception or some other module. Attention codelets watch both it and the resulting associations. One of these attention codelets may decide that this information should be acted upon. This codelet would then attempt to take the information to "consciousness," perhaps along with any discrepancies it may find with the help of associations. If the attempt is successful, the coalition manager makes a coalition of them, the spotlight controller eventually selects that coalition, and the contents of the coalition are broadcast to all the codelets. In response to the broadcast, appropriate behavior-priming codelets perform three tasks: an appropriate goal structure is instantiated in the behavior net, the codelets bind variables in the behaviors of that structure, and the codelets send activation to the currently appropriate behavior of the structure. Eventually that behavior is chosen to be acted upon. At this point, information about the current emotion and the currently executing behavior is written to the focus by the behavior codelets associated with the chosen behavior. The current contents of the write registers in the focus are then written to associative memory. The rest of the behavior codelets associated with the chosen behavior then perform their tasks. Thus, an action has been selected and carried out by means of collaboration between "consciousness" and the behavior net.

This background information on the IDA architecture and mechanisms should enable the reader to understand IDA's cognitive cycle (Baars & Franklin, 2003: Franklin et al., 2005). The cognitive cycle specifies the functional roles of memory, emotions, consciousness, and decision making in cognition, according to the global workspace theory. Below, we sketch the steps of the cognitive cycle; see Figure 7.5 for an overview.

1. *Perception*. Sensory stimuli, external or internal, are received and interpreted by perception. This stage is unconscious.
2. *Percept to Preconscious Buffer*. The percept is stored in preconscious buffers of IDA's working memory.
3. *Local Associations*. Using the incoming percept and the residual contents of the preconscious buffers as cues, local associations are automatically retrieved from transient episodic memory and from long-term memory.
4. *Competition for Consciousness*. Attention codelets, whose job is to bring relevant, urgent, or insistent events to consciousness, gather information, form coalitions, and actively compete against each other.

(The competition may also include attention codelets from a recent previous cycle.)

5. *Conscious Broadcast*. A coalition of codelets, typically an attention codelet and its covey of related information codelets carrying content, gains access to the global workspace and has its contents broadcast. The contents of perceptual memory are updated in light of the current contents of consciousness. Transient episodic memory is updated with the current contents of consciousness as events. (The contents of transient episodic memory are separately consolidated into long-term memory.) Procedural memory (recent actions) is also updated.
6. *Recruitment of Resources*. Relevant behavior codelets respond to the conscious broadcast. These are typically codelets whose variables can be bound from information in the conscious broadcast. If the successful attention codelet was an expectation codelet calling attention to an unexpected result from a previous action, the responding codelets may be those that can help rectify the unexpected situation. (Thus consciousness solves the relevancy problem in recruiting resources.)
7. *Setting Goal Context Hierarchy*. The recruited processors use the contents of consciousness to instantiate new goal context hierarchies, bind their variables, and increase their activation. Emotions directly affect motivation and determine which terminal goal contexts receive activation and how much. Other (environmental) conditions determine which of the earlier goal contexts receive additional activation.
8. *Action Chosen*. The behavior net chooses a single behavior (goal context). This selection is heavily influenced by activation passed to various behaviors influenced by the various emotions. The choice is also affected by the current situation, external and internal conditions, by the relation between the behaviors, and by the residual activation values of various behaviors.
9. *Action Taken*. The execution of a behavior (goal context) results in the behavior codelets performing their specialized tasks, which may have external or internal consequences. The acting codelets also include an expectation codelet (see Step 6) whose task is to monitor the action and to try and bring to consciousness any failure in the expected results.

IDA's elementary cognitive activities occur within a single cognitive cycle. More complex cognitive functions are implemented over multiple cycles. These include deliberation, metacognition, and voluntary action (Franklin, 2000).

The IDA model employs a methodology that is different from that which is currently typical of computational cognitive models. Although the model is based on experimental findings in cognitive psychology and brain science, there is only qualitative consistency with experiments. Rather, there are a number of hypotheses derived from IDA as a unified theory of cognition. The IDA model generates hypotheses about human cognition and the role of consciousness through its design, the mechanisms of its modules, their interaction, and its performance.

Every agent must sample and act on its world through a sense-select-act cycle. The frequent sampling allows for a fine-grained analysis of common cognitive phenomena, such as process dissociation, recognition vs. recall, and the availability heuristic. At a high level of abstraction, the analyses support the commonly held explanations of what occurs in these situations and why. At a finer-grained level, the analyses flesh out common explanations, adding details and functional mechanisms. Therein lies the value of these analyses.

Unfortunately, currently available techniques for studying some phenomena at a fine-grained level, such as PET, fMRI, EEG, implanted electrodes, etc., are lacking either in scope, in spatial resolution, or in temporal resolution. As a result, some of the hypotheses from the IDA model, although testable in principle, seem not to be testable at the

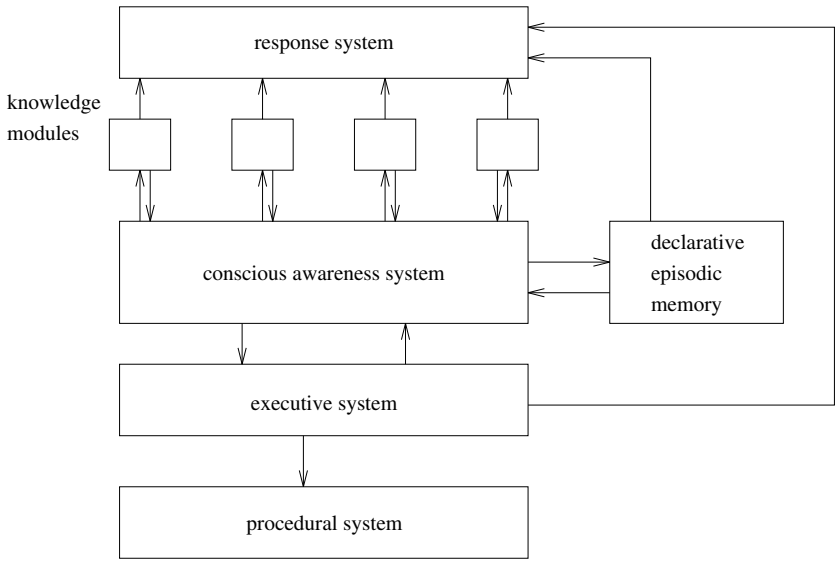

Figure 7.7. Schacter's model of consciousness.

present time for lack of technologies with suitable scope and resolution.

There is also the issue of the breadth of the IDA model, which encompasses perception, working memory, declarative memory, attention, decision making, procedural learning, and more. How can such a broad model produce anything useful? The IDA model suggests that these various aspects of human cognition are highly integrated. A more global view can be expected to add additional understanding to that produced by more specific models. This assertion seems to be borne out by the analyses of various cognitive phenomena (Baars & Franklin, 2003; Franklin et al., 2005).

Sketches of Some Other Views

As we have seen, there are many attempts to explain the difference in conscious accessibility. Various explanations have been advanced in terms of the content of knowledge (e.g., instances vs. rules), the organization of knowledge (e.g., declarative vs. procedural), processing mechanisms (e.g., spreading activation vs. rule matching and firing), the representation of knowledge (e.g., localist/symbolic vs. distributed), and so on. In addition to the two views elaborated on earlier, let us look into some more details of a few other views. Although some of the models that are discussed below are not strictly speaking computational (because they may not have been fully computationally implemented), they are nevertheless important because they point to possible ways of constructing computational explanations of consciousness.

We can examine Schacter's (1990) model as an example. The model is based on neuropsychological findings of the dissociation of different types of knowledge (especially in brain-damaged patients). It includes a number of "knowledge modules" that perform specialized and unconscious processing and may send their outcomes to a "conscious awareness system," which gives rise to conscious awareness (see Figure 7.7). Schacter's explanation of some neuropsychological disorders (e.g., hemisphere neglect, blindsight, aphasia, agnosia, and prosopagnosia) is that brain damages result in the disconnection of some of the modules from the conscious awareness system, which causes their inaccessibility to consciousness. However, as has been pointed out by others, this explanation cannot account for many findings in implicit memory research (e.g., Roediger,

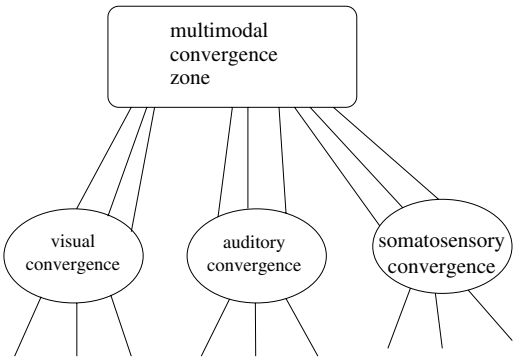

Figure 7.8. Damasio's model of consciousness.

1990). Revonsuo (1993) advocated a similar view, albeit from a philosophical viewpoint, largely on the basis of using Schacter's (1990) data as evidence. Johnson-Laird's (1983) model was somewhat similar to Schacter's model in its overall structure in that there was a hierarchy of processors and consciousness resided in the processes at the top of the hierarchy. Shallice (1972) put forward a model in which a number of "action systems" could be activated by "selector input" and the activated action systems correspond to consciousness. It is not clear, however, what the computational (mechanistic) difference between conscious and unconscious processes is in those models, which did not offer a mechanistic explanation.

We can compare Schacter (1990)'s model with CLARION. It is similar to CLARION, in that it includes a number of "knowledge modules" that perform specialized and unconscious processing (analogous to bottom-level modules in CLARION) and send their outcomes to a "conscious awareness system" (analogous to the top level in CLARION), which gives rise to conscious awareness. Unlike CLARION's explanation of the conscious/unconscious distinction through the difference between localist/symbolic versus distributed representations, however, Schacter's model does not elucidate in computational/mechanistic terms the qualitative distinction between conscious and unconscious processes, in that the "conscious awareness system" lacks any apparent qualitative difference from the unconscious systems.

We can also examine Damasio's neuroanatomically motivated model (Damasio et al., 1990). The model hypothesizes the existence of many "sensory convergence zones" that integrate information from individual sensory modalities through forward and backward synaptic connections and the resulting reverberations of activations, without a central location for information storage and comparisons; it also hypothesizes the global "multimodal convergence zone," which integrates information across modalities also through reverberation (via recurrent connections; see Figure 7.8). Correlated with consistency is global information availability; that is, once "broadcast" or "reverberation" is achieved, all the information about an entity stored in difference places of the brain becomes available. This was believed to have explained the accessibility of consciousness.[6] In terms of CLARION, different sensory convergence zones may be roughly captured by bottom-level modules, each of which takes care of sensory inputs of one modality (at a properly fine level), and the role of the global multi-modal convergence zone (similar to the global workspace in a way) may be played by the top level of CLARION, which has the ultimate responsibility for integrating information (and also serves as the "conscious awareness system"). The widely recognized role of reverberation (Damasio, 1994; Taylor, 1994) may be captured in CLARION through using recurrent connections within modules at the bottom level and through multiple top-down and bottom-up information flows across the two levels, which leads to the unity of consciousness that is the synthesis of all the information present (Baars, 1988; Marcel, 1983).

Similarly, Crick and Koch (1990) hypothesize that synchronous firing at 35–75 Hz in the cerebral cortex is the basis for consciousness – with such synchronous firing, pieces of information regarding different aspects of an entity are brought together, and thus consciousness emerges. Although consciousness has been experimentally observed to be somewhat correlated with synchronous firing at 35–75 Hz, there is no explanation of *why* this is the case and there is

no computational/mechanistic explanation of any *qualitative* difference between 35–75 Hz synchronous firing and other firing patterns.

Cotterill (1997) offers a "master-module" model of consciousness, which asserts that consciousness arises from movement or the planning of movement. The master-module refers to the brain region that is responsible for motor planning. This model sees the conscious system as being profligate with its resources: Perforce it must plan and organize movements, even though it does not always execute them. The model stresses the vital role that movement plays and is quite compatible with the IDA model. This centrality of movement was illustrated by the observation that blind people were able to read braille when allowed to move their fingers, but were unable to do so when the dots were moved against their still fingers (Cotterill, 1997).

Finally, readers interested in the possibility of computational models of consciousness actually producing "conscious" artifacts may consult Holland (2003) and other work along that line.

Concluding Remarks

This chapter has examined general frameworks of computational accounts of consciousness. Various related issues, such as the utility of computational models, explanations of psychological data, and potential applications of machine consciousness, have been touched on in the process. Based on existing psychological and philosophical evidence, existing models were compared and contrasted to some extent. It appears inevitable at this stage that there is the coexistence of various computational accounts of consciousness. Each of them seems to capture some aspect of consciousness, but each also has severe limitations. To capture the whole picture in a unified computational framework, much more work is needed. In this regard, CLARION and IDA provide some hope.

Much more work can be conducted on various issues of consciousness along this computational line. Such work may include further specifications of details of computational models. It may also include reconciliations of existing computational models of consciousness. More importantly, it may, and should, include the validation of computational models through empirical and theoretical means. The last point in particular should be emphasized in future work (see the earlier discussions concerning CLARION and IDA). In addition, we may also attempt to account for consciousness computationally at multiple levels, from phenomenology, via various intermediate levels, all the way down to physiology, which will likely lead to a much more complete computational account and a much better picture of consciousness (Coward & Sun, 2004).

Acknowledgments

Ron Sun acknowledges support in part from Office of Naval Research grant N00014-95-1-0440 and Army Research Institute grants DASW01-00-K-0012 and W74V8H-04-K-0002. Stan Franklin acknowledges support from the Office of Naval Research and other U.S. Navy sources under grants N00014-01-1-0917, N00014-98-1-0332, N00014-00-1-0769, and DAAH04-96-C-0086.

Notes

1. There are also various numerical measures involved, which are not important for the present discussion.
2. Cleeremans and McClelland's (1991) model of artificial grammar learning can be viewed as instantiating half of the system (the unconscious half), in which implicit learning takes place based on gradual weight changes in response to practice on a task and the resulting changes in activation of various representations when performing the task.
3. Note that the accessibility is defined in terms of the surface syntactic structures of the

objects being accessed (at the level of outcomes or processes), not their semantic meanings. Thus, for example, a LISP expression is directly accessible, even though one may not fully understand its meaning. The internal working of a neural network may be inaccessible even though one may know what the network essentially does (through an interpretive process). Note also that objects and processes that are directly accessible at a certain level may not be accessible at a finer level of details.

4. This activation of features is important in subsequent uses of the information associated with the concept and in directing behaviors.
5. This is a term borrowed from computer operating systems that describes a small piece of code that waits and watches for a particular event or condition to occur before it acts.
6. However, consciousness does not necessarily mean accessibility/availability of all the information about an entity; for otherwise, conscious inference, deliberate recollection, and other related processes would be unnecessary.

References

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anderson, J., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Anwar, A., & Franklin, S. (2003). Sparse distributed memory for "conscious" software agents. *Cognitive Systems Research*, *4*, 339–354.

Baars, B. (1988). *A cognitive theory of consciousness*. New York: Cambridge University Press.

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

Baars, B., & Franklin, S. (2003). How conscious experience and working memory interact. *Trends in Cognitive Science*, *7*, 166–172.

Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577–609.

Berry, D., & Broadbent, D. (1988). Interactive tasks and the implicit-explicit distinction. *British Journal of Psychology*, *79*, 251–272.

Bower, G. (1996). Reactivating a reactivation theory of implicit memory. *Consciousness and Cognition*, *5*(1/2), 27–72.

Bowers, K., Regehr, G., Balthazard, C., & Parker, K. (1990). Intuition in the context of discovery. *Cognitive Psychology*, *22*, 72–110.

Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual process theories in social psychology*. New York: Guilford Press.

Clark, A. (1992). The presence of a symbol. *Connection Science*. *4*, 193–205.

Clark, A., & Karmiloff-Smith, A. (1993). The cognizer's innards: A psychological and philosophical perspective on the development of thought. *Mind and Language*, *8*(4), 487–519.

Cleeremans, A., & McClelland, J. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*, 235–253.

Collins, A., & Loftus, J. (1975). Spreading activation theory of semantic processing. *Psychological Review*, *82*, 407–428.

Cosmides, L., & Tooby, J. (1994). Beyond intuition and instinct blindness: Toward an evolutionarily rigorous cognitive science. *Cognition*, *50*, 41–77.

Cotterill, R. (1997). On the mechanism of consciousness. *Journal of Consciousness Studies*, *4*, 231–247.

Coward, L. A., & Sun, R. (2004). Criteria for an effective theory of consciousness and some preliminary attempts. *Consciousness and Cognition*, *13*, 268–301.

Crick, F., & Koch, C. (1990). Toward a neurobiological theory of consciousness. *Seminars in the Neuroscience*, *2*, 263–275.

Damasio, A. (1994). *Descartes' error*. New York: Grosset/Putnam.

Damasio. A., et al. (1990). Neural regionalization of knowledge access. *Cold Spring Harbor Symposium on Quantitative Biology*, *LV*.

Dennett, D. (1991), *Consciousness explained*. Boston: Little Brown.

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

Edelman, G. (1989). *The remembered present: A biological theory of consciousness*. New York: Basic Books.

Edelman, G., & Tononi, G. (2000). *A universe of consciousness*. New York: Basic Books.

Freeman, W. (1995). *Societies of brains*. Hillsdale, NJ: Erlbaum.

Fodor, J. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Franklin, S. (2000). Deliberation and voluntary action in 'conscious' software agents. *Neural Network World*, *10*, 505–521.

Franklin, S., Baars, B. J., Ramamurthy, U., & Ventura, M. 2005. The Role of consciousness in memory. *Brains, Minds and Media*, *1*, 1–38.

Franklin, S., & Graesser, A. C. (1997). Is it an agent, or just a program?: A taxonomy for autonomous agents. *Intelligent agents III*, 21–35.

Franklin, S., Kelemen, A., & McCauley, L. (1998). IDA: A cognitive agent architecture. *IEEE Conference on Systems, Man and Cybernetics*.

Hadley, R. (1995). The explicit-implicit distinction. *Minds and Machines*, *5*, 219–242.

Hirschfeld, L., & Gelman, S. (1994). *Mapping the Mind: Domain Specificity in Cognition and Culture*. New York: Cambridge University Press.

Hofstadter, D., & Mitchell, M. (1994). The copycat project: A model of mental fluidity and analogy-making. In K. J. Holyoak, & J. A. Barnden (Eds.), *Advances in connectionist and neural computation theory, Vol. 2: Logical connections*. Norwood, NJ: Ablex.

Holland, O. (2003). *Machine consciousness*. Exeter, UK: Imprint Academic.

Hunt, E., & Lansman, M. (1986). Unified model of attention and problem solving. *Psychological Review*, *93*(4), 446–461.

Jackendoff, R. (1987). *Consciousness and the computational mind*. Cambridge, MA: MIT Press.

Johnson-Laird, P. (1983). A computational analysis of consciousness. *Cognition and Brain Theory*, *6*, 499–508.

Kanerva, P. (1988). *Sparse distributed memory*. Cambridge MA: MIT Press.

Karmiloff-Smith, A. (1986). From metaprocesses to conscious access: Evidence from children's metalinguistic and repair data. *Cognition*, *23*, 95–147.

LeDoux, J. (1992). Brain mechanisms of emotion and emotional learning. *Current Opinion in Neurobiology*, *2*(2), 191–197.

Logan, G. (1988). Toward a theory of automatization. *Psychological Review*, *95*(4), 492–527.

Maes, P. (1989). How to do the right thing. *Connection Science*, *1*, 291–323.

Marcel, A. (1983). Conscious and unconscious perception: An approach to the relations between phenomenal experience and perceptual processes. *Cognitive Psychology*, *15*, 238–300.

Maslow, A. (1987). *Motivation and personality* (3d ed.). New York: Harper and Row.

D. Mathis D., & Mozer, M., (1996). Conscious and unconscious perception: A computational theory. *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, 324–328.

McClelland, J., McNaughton, B., & O'Reilly, R. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457.

Michalski, R. (1983). A theory and methodology of inductive learning. *Artificial Intelligence*, *20*, 111–161.

Minsky, M. (1985). *The society of mind*. New York: Simon and Schuster.

Moscovitch, M., & Umilta, C. (1991). Conscious and unconscious aspects of memory. In *Perspectives on cognitive neuroscience*. New York: Oxford University Press.

Neal, A., & Hesketh, B. (1997). Episodic knowledge and implicit learning. *Psychonomic Bulletin and Review*, *4*(1), 24–37.

Negatu, A., & Franklin, S. (2002). An action selection mechanism for 'conscious' software agents. *Cognitive Science Quarterly*, *2*, 363–386.

Nelson, T. (Ed.) (1993). *Metacognition: Core Readings*. Boston, MA: Allyn and Bacon.

O'Brien. G., & Opie, J. (1998). A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences*, *22*, 127–148.

Penrose, R. (1994). *Shadows of the mind*. Oxford: Oxford University Press.

Quillian, M. R. (1968). Semantic memory. In M. Minsky (Ed.), *Semantic information processing* (pp. 227–270). Cambridge, MA: MIT Press.

Ramamurthy, U., D'Mello, S., & Franklin, S. (2004). Modified sparse distributed memory as transient episodic memory for cognitive software agents. In *Proceedings of the International Conference on Systems, Man and Cybernetics*. Piscataway, NJ: IEEE.

Reber, A. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, *118*(3), 219–235.

Revonsuo, A. (1993). Cognitive models of consciousness. In M. Kamppinen (Ed.), *Consciousness, cognitive schemata and relativism* (pp. 27–130). Dordrecht, Netheland: Kluwer.

Roediger, H. (1990). Implicit memory: Retention without remembering. *American Psychologist*, *45*(9), 1043–1056.

Rosenbloom, P., Laird, J., & Newell, A. (1993). *The SOAR papers: Research on integrated intelligence*. Cambridge, MA: *MIT* Press.

Rosenthal, D. (Ed.). (1991). *The nature of mind*. Oxford: Oxford University Press.

Rumelhart, D., McClelland, J., & the PDP Research Group, (1986). *Parallel distributed processing: Explorations in the microstructures of cognition*. Cambridge, MA: MIT Press.

Schacter, D. (1990). Toward a cognitive neuropsychology of awareness: Implicit knowledge and anosagnosia. *Journal of Clinical and Experimental Neuropsychology*, *12*(1), 155–178.

Schneider, W., & Oliver, W. (1991). An instructable connectionist/control architecture. In K. VanLehn (Ed.), *Architectures for intelligence*. Hillsdale, NJ: Erlbaum.

Searle, J. (1980). Minds, brains, and programs. *Brain and Behavioral Sciences*, *3*, 417–457.

Seger, C. (1994). Implicit learning. *Psychological Bulletin*, *115*(2), 163–196.

Servan-Schreiber, E., & Anderson, J. (1987). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 592–608.

Shallice, T. (1972). Dual functions of consciousness. *Psychological Review*, *79*(5), 383–393.

Sloman, A., & Chrisley, R. (2003). Virtual machines and consciousness. *Journal of Consciousness Studies*, *10*, 133–172.

Smith, E., & Medin, D. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Smith, J. D., Shields, W. E., & Washburn, D. A. (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences*. *26*, 317–339.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*(1), 1–74.

Stanley, W., Mathews, R., Buss, R., & Kotler-Cope, S. (1989). Insight without awareness: On the interaction of verbalization, instruction and practice in a simulated process control task. *Quarterly Journal of Experimental Psychology*, *41A*(3), 553–577.

Sun, R. (1994). *Integrating rules and connectionism for robust commonsense reasoning*. New York: John Wiley and Sons.

Sun, R. (1995). Robust reasoning: Integrating rule-based and similarity-based reasoning. *Artificial Intelligence*, *75*(2), 241–296.

Sun, R. (1997). Learning, action, and consciousness: A hybrid approach towards modeling consciousness. *Neural Networks*, *10*(7), 1317–1331.

Sun, R. (1999a). Accounting for the computational basis of consciousness: A connectionist approach. *Consciousness and Cognition*, *8*, 529–565.

Sun, R. (1999b). Computational models of consciousness: An evaluation. *Journal of Intelligent Systems [Special Issue on Consciousness]*, *9*(5–6), 507–562.

Sun, R. (2002). *Duality of the mind*. Mahwah, NJ: Erlbaum.

Sun, R. (2003). *A tutorial on CLARION*. Retrieved from http://www.cogsci.rpi.edu/~rsun/sun.tutorial.pdf.

Sun, R., & Bookman, L. (Eds.). (1994). *Computational architectures integrating neural and symbolic processes*. Norwell, MA: Kluwer.

Sun, R., Merrill, E., & Peterson, T. (2001). From implicit skills to explicit knowledge: A bottom-up model of skill learning. *Cognitive Science*, *25*(2), 203–244.

Sun, R., & Naveh, I. (2004, June). Simulating organizational decision making with a cognitive architecture CLARION. *Journal of Artificial Society and Social Simulation*, *7*(3). Retrieved from http://jasss.soc.surrey.ac.uk/7/3/5.html.

Sun, R., & Peterson, T. (1998). Autonomous learning of sequential tasks: Experiments and analyses. *IEEE Transactions on Neural Networks*, *9*(6), 1217–1234.

Sun, R., & Peterson, T. (1999). Multi-agent reinforcement learning: Weighting and partitioning. *Neural Networks*, *12*(4–5). 127–153.

Sun, R., Peterson, T., & Merrill, E. (1996). Bottom-up skill learning in reactive sequential decision tasks. *Proceedings of 18th Cognitive Science Society Conference*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Sun, R., Merrill, E.. & Peterson, T. (1998). A bottom-up model of skill learning. *Proceedings of the 20th Cognitive Science Society Conference*

(pp. 1037–1042). Mahwah, NJ: Lawrence Erlbaum Associates

Sun , R., Slusarz, P., & Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: A dual-process approach. *Psychological Review*, *112* , 159–192.

Taylor, J. (1994). Goal, drives and consciousness. *Neural Networks*, *7* (6/7), 1181–1190.

Taylor, J. (1997).The relational mind. In A. Browne (Ed.), *Neural network perspectives on cognition and adaptive robotics*. Bristol, UK: Institute of Physics.

Toates, F. (1986). *Motivational systems*. Cambridge: Cambridge University Press.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.

Watkins, C. (1989). *Learning with delayed rewards*. PhD Thesis, Cambridge University, Cambridge, UK.

Willingham, D., Nissen, M., & Bullemer, P. (1989). On the development of procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 1047–1060.