Accounting for the Computational Basis of Consciousness: A Connectionist Approach

Ron Sun
NEC Research Institute, 4 Independence Way, Princeton, NJ 08540

EMAIL: rsun@research.nj.nec.com
PHONE: (215) 592-4150

September 6, 1999

To appear in: Consciousness and Cognition, 1999

ACKNOWLEDGEMENT: This work resulted from the project supported (in part) by Office of Naval Research grant N00014-95-1-0440. Thanks to Ed Merrill, Jeff Shrager, Jack Gelfand, and David Roskos-Ewoldsen for discussions or comments. Todd Peterson and Chris Terry helped with related work. The author also wishes to thank Bernard Baars and an anonymous reviewer for their detailed comments.

Abstract

This paper argues for an explanation of the mechanistic (computational) basis of consciousness that is based on the distinction between localist (symbolic) representation and distributed representation, the ideas of which have been put forth in the connectionist literature. A model is developed to substantiate and test this approach. The paper also explores the issue of the functional roles of consciousness, in relation to the proposed mechanistic explanation of consciousness. The model, embodying the representational difference, is able to account for the functional role of consciousness, in the form of the synergy between the conscious and the unconscious. The fit between the model and various cognitive phenomena and data (documented in the psychological literatures) is discussed to accentuate the plausibility of the model and its explanation of consciousness. Comparisons with existing models of consciousness are made in the end.

1 Introduction

Although the study of consciousness has been downplayed in the study of cognition for long, the importance of consciousness to cognitive science cannot be over-estimated. Consciousness is a central question in cognition. Studying only the brain, the computational models of quantitative data, or some combinations thereof will not be sufficient to help us to understand the central issue of the human mind – the consciousness. We need to confront the issue directly.

Recently, there is a resurgence of interest in consciousness per se as a subject for experimental and theoretical study in cognitive science. Various new views have been proposed, including consciousness as an emergent property (e.g., as in a connectionist model that settles into an attractor), consciousness as a system separate from the rest of the mind that works deliberately and serially, consciousness as a supervisory system, or consciousness as a dominant process in a pool of processes running in parallel, and so on. In this paper, I will try to come to some conclusions as to what a plausible computational account of consciousness should be like. I will describe a model (Sun et al 1996, 1997, Sun 1997) that incorporates the distinction between localist vs. distributed representations. I will show how the model accounts for the physical basis and the functional roles of consciousness and shed light on various issues concerning consciousness. In so doing, I will also bring together various operationalized notions and computational accounts of consciousness (or aspects of it), as motivations for, and to be accounted for by, the model.

I assume the sufficiency and necessity of mechanistic explanations. By mechanistic explanation, I mean any concrete physical processes, that is, computational processes in the broadest sense of the term "computational". In terms of the sufficiency of mechanistic explanations, the following hypothesis serves as our working hypothesis (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, despite various assaults on it. [1] In general, "computation" is a broad term that can be used to denote any process that can be realized computationally, ranging from chaotic dynamics (Freeman 1995) and "Darwinian" competition (Edelman 1989), to quantum mechanics (Penrose 1994). To avoid confusion that may arise as to what computation refers to, I will stick to the terms "mechanism" or "mechanistic processes". On the other hand, as to the necessity of mechanistic explanations, it is obvious to anyone who is not a dualist that the foregoing definition of mechanistic processes has to include the necessary condition for consciousness; for the physical basis of mental activities and phenomenal experience cannot be anything else but such mechanistic processes.

We need an explanation of the mechanistic basis of consciousness and its mechanistic roles (or functions) in the human mind: what kind of mechanism leads to the conscious, and what kind of mechanism leads to the unconscious? What is the functional role of the conscious? What is the

[1] These assaults (e.g., Edelman 1989, Freeman 1995, Damasio 1994, Penrose 1994, Searle 1991) failed to show that computation, in general, cannot account for the nature of consciousness, although they had some legitimate complains about specific computational approaches and models (some of these complaints are shared by this author).

functional role of the unconscious? There have been many such explanations in computational or biological terms. Contrary to some critics, the debate among them is *not* analogous to a debate between algebraists and geometers on physics (which would be meaningless and irrelevant). By all measures, it is analogous to some substantive debates, e.g., the wave vs. particle debate in physics concerning the nature of light. It is substantive because it provides necessary theoretical frameworks for further empirical work on consciousness.

In the remainder of this paper, section 2 will provide a review of work in this area, which points to the plausibility of dichotomous structuring of the mind. In section 3, a model will be developed along that line. A discussion will ensue in section 4 concerning how the model accounts for the physical basis of consciousness. In section 5, the model will be further used to account for the functional roles of consciousness, with a sketch of the simulation using the model. The model will then be compared to existing models in section 6. Section 7 concludes the paper.

2 Background

Before taking a stab at the problem of devising a plausible model, I will present some evidence and arguments for the duality, or dichotomy, of the mind, which naturally leads to the *dual-representation hypothesis* (which I put forth earlier in Sun 1992, 1994, 1995) and consequently the two-level models (Sun 1994, 1995, 1997).

First let us look into some of the early ideas concerning dichotomies or dualities of the mind that dated back before the inception of cognitive science. For instance, Heidegger's distinction — the preontological vs. the ontological — is a highly abstract version of such a dichotomy. As a first approximation, his view is that, since the essential way of being is existence in the world, an agent always embodies an understanding of its being through such existence. This embodied understanding is implicit and consists of skills, reactions, and know-hows, without an explicit "ontology", and is thus *preontological*. On that basis, an agent may also achieve an explicit understanding, an *ontological* understanding, especially through making explicit the implicitly held understanding; or in other words, the agent can turn preontological understanding into ontological understanding (Heidegger 1927, Dreyfus 1991). This dichotomy and progression from the concrete to the abstract are the basis of our model (to be explained later).

It is also worthwhile to mention William James's distinction of "empirical thinking" and "true reasoning". According to James, on the one hand, empirical thinking is associative, made up of sequences of "images" that are suggested by one another. It is "reproductive", because it is always replicating in some way past experience, instead of producing new or stand-alone ideas. Empirical thinking relies on overall comparisons and similarity among various concrete situations, and therefore may lose sight of some critical information. On the other hand, "true reasoning" can be arrived at by abstracting particular attributes (i.e., those attributes that are essential and critical) out of a situation. In so doing, we assume a particular way of conceiving things — we see things as a particular aspect of them. It is "productive", because it is capable of producing novel ideas through abstraction. An important function that "true reasoning" serves is to break up the direct link between thought and

action, and to provide means for articulately and theoretically reasoning about consequences of an action without actually performing it.

Dreyfus and Dreyfus (1987) recently proposed the distinction of analytical and intuitive thinking, refining and revising Heidegger's distinction in a contemporary context. They claim that analytical thinking corresponds to what traditional (symbolic) AI models are aimed to capture: deliberate, sequential processing that follows rules and performs symbolic manipulation. According to them, when an agent first learns about a domain (for example, chess), an agent learns explicit rules and follows them one-by-one. After gaining some experience with the domain, one will start to develop certain overall understanding of a situation as a whole, without deliberate rule-following and analytical thinking. That is, one starts to use intuitive thinking, which has the characteristics of being situationally sensitive, "holographic", and non-rule-like. Contrasting the two types of thinking, there is clearly a dichotomy, although the focus here is the reverse progression from the abstract to the concrete.

There are a few arguments from within cognitive science, and from connectionists in particular, that are related to this dichotomy. Smolensky (1988) proposed the distinction of conceptual and subconceptual processing. Conceptual processing involves knowledge that possesses the following characteristics: (1) public access, (2) reliability, and (3) formality. In other words, they are what traditional symbolic AI tries to capture (as similarly identified by Dreyfus and Dreyfus 1987). On the other hand, there are other kinds of capacities, such as skill, intuition, and individual knowledge that are not expressible in linguistic forms and do not conform to the three criteria prescribed above. It has been futile so far to try to model such capacities with conceptual processing based models (traditional AI symbolic processing models). Some of the capacities should be viewed as an entirely different level in cognition and modeled as such, that is, at the subconceptual level. The subconceptual level may be better dealt with by the connectionist subsymbolic models, because the connectionist approach seems to be able to overcome some problems symbolic AI models encountered in modeling subconceptual processing. Smolensky (1989) explicitly tied the distinction of the conscious and the subconscious to that of the conceptual and the subconceptual. Hinton (1990) posited a related hypothesis regarding a similar distinction — the distinction between "intuitive inference" and "rational inference", as a justification for the distinction between the traditional symbolic AI models and the connectionist subsymbolic models.

For many decades up until very recently, in experimental studies of the human mind, the notion of consciousness has been replaced with various operationalized notions, e.g., concerning the explicit vs. the implicit, the controlled vs. the automatic, the intentional vs. the incidental, and so on. For instance, in cognitive psychology, there is the well established distinction of implicit memory vs. explicit memory (Schacter 1990, Roedeger 1990). Implicit memory refers to unconscious retrieval of memories, without explicit awareness. Based on the dissociation of explicit and implicit memory tests, it was suggested that implicit memory and explicit memory involved different memory systems (for example, the episodic memory system vs. the semantic memory system, or the declarative memory vs. the procedural memory; Bower 1996, Squire et al 1993, Schacter 1990). Related, but not identical to this, there is also the distinction of implicit learning and explicit learning. In the research on implicit learning, Berry and Broadbent (1988), Willingham et al (1989), and Reber (1989) expressly demonstrated a dissociation between explicit knowledge and performance in a variety of tasks. The

notion of automaticity (Shiffrin and Schneider 1977, Logan 1988) is related to that of implicit learning. Automatic processing is assumed to be effortless and resource-wise (almost) unbounded, while its opposite, controlled processing, requires the use of limited cognitive resources (Navon and Gopher 1979). There is also the distinction of declarative and procedural knowledge from cognitive psychology. In Anderson (1983), based on psychological data on high-level cognitive skill learning (such as arithmetic and theorem proving), it was suggested that there were these two types of knowledge. As described by Anderson (1983) and many others, while the former type of knowledge is generic and easily accessible, the latter type is embodied and specific. In this theory, there is a clear dichotomy between these two types of knowledge. In all, many different theories in cognitive science are clearly related to the issue of consciousness although under various guises as theories for various operationalized notions. The evidence for these dichotomies lies in experimental data that elucidate various dissociations and differences in performance under different conditions.

Also worth mentioning here is the notions of two types of connectionist representations, used in the connectionist literature. Basically, in *localist* (or symbolic) representation, each representational unit (node) represents a distinct entity to be represented. There is a one-to-one correspondence between units of representation and entities to be represented. In *distributed* representation, each entity is represented by a pattern of activation among a pool of representational units (nodes). Although there is a one-to-one correspondence between an entity to be represented and a pattern of activation, there is no one-to-one correspondence between entities to be represented and representational units (nodes). This distinction is reminiscent of various other distinctions related to the conscious and the unconscious, and actually bears close relationships (Sun 1994, 1995, 1997) to these other distinctions, as will be further discussed later.

There seems to be a consensus with regard to the *qualitative* difference between different types of thinking (although there is no consensus regarding the actual details of the dichotomies). Moreover, most of the aforementioned authors believed in incorporating both sides of the dichotomies in cognitive models, because each side serves a unique cognitive function and is thus indispensable for a complete cognitive model. On this basis, it is natural to hypothesize the existence of two separate components, whereby each component is responsible for one side of a dichotomy, for example, as have been proposed in Anderson (1993), Hunt and Lansman (1986), Logan (1988), Reber (1989), Schacter (1990), and Sun (1994, 1995). The two components were variously described as production systems vs. semantic networks, as algorithmic processes vs. instances retrieval, or as localist representation vs. distributed representations. To help to further narrow down choices in developing the two components, four criteria can be hypothesized (see Sun 1994):

- *Direct accessibility of conscious processes*: Here direct accessibility refers to the direct and immediate availability of mental content for the major operations that are responsible for, or concomitant with, consciousness, such as introspection and forming higher-order thoughts (as well as verbal reporting, and meta-level control and manipulation). To capture such accessibility, concepts (as well as processes operating on concepts) should be *directly* accessible without intermediate interpretive or transformational steps; which is a requirement prescribed and/or accepted by many (see, e.g., Clark 1992, Hadley 1995). This requirement basically rules out connectionist distributed representation, among other things, because a concept represented by

a distributed pattern is not *directly* accessible (to the processes mentioned above; more explanations later). This requirement leaves us with two obvious options: purely symbolic representation and localist connectionist representation.

- *Direct inaccessibility of unconscious processes*: As argued before, unconscious processes are carried out implicitly, for example, in a holistic way. The details of the processes are not accessible. [2] This is important, because implicit and/or holistic operations may entertain a host of properties that other operations lack. A distributed representation in connectionist models seems to be a viable way to go (Sun 1994, 1995, Sloman 1996; more explanations later).

- *Linkages from localist concepts to distributed features*: Once a localist/symbolic concept is activated, its corresponding distributed representations (features) are also activated (as assumed in most cognitive models, ranging from Tversky 1977 to Sun 1995). This (explicit or implicit) activation of features is important in subsequent uses of the information associated with the concept and in directing behaviors.

- *Linkages from distributed features to localist concepts*: Under appropriate circumstances, once some or most of the distributed features of a concept are activated, the concept itself can be activated to "cover" these features (roughly the same as the categorization process; Smith & Medin 1981).

Based on these criteria, the most plausible way of structuring the two components is utilizing the representational difference: the localist (symbolic) vs. distributed representations (as advocated in the connectionist theorizing). This conjecture is supported by much existing work (Reber 1989, Sun 1994, McClelland et al 1994). In recent years, there have been models resulting from the connectionist approach that support the two-component hypothesis and rely on representational differences, inspired by some of the earlier theorizing (by e.g. James, Freud, Heidegger, and Dreyfus). The success of some of these models is an indication that this sort of idea is probably on the right track. For example, Hendler (1989) presented a hybrid system for planning. It utilized connectionist networks for priming relevant concepts through activation propagation, to augment a symbolic planning system in order to pick out right actions, which otherwise might have too many choices. The combination of the two types of mechanisms aided in the effectiveness of the model. Gelfand et al (1989) proposed a model for robot skill learning that included both symbolic and neural network representation of knowledge. It began by encoding all the knowledge in an explicit symbolic form and through practice the knowledge was assimilated into a neural network (using backpropagation). In the end, the network was able to capture in an implicit form the skill for the task. CONSYDERR (Sun 1994) was another example, which consisted of two levels of connectionist networks: one level employed localist representation for modeling conceptual reasoning and the other level employed distributed representation for modeling subconceptual reasoning (see Figure 1). Through the interaction of the two level, many seemingly disparate and difficult patterns of commonsense reasoning were accounted for uniformly.

Note that the questions concerning the adequacy of the symbolic processing capability in either localist or distributed representations have been raised repeatedly. Despite a few attempted theoreti-

[2] It is generally not the case that unconscious processes are not accessible at all but they are definitely less accessible, not as direct and immediate as conscious processes. They may be accessed through indirect, transformational processes.

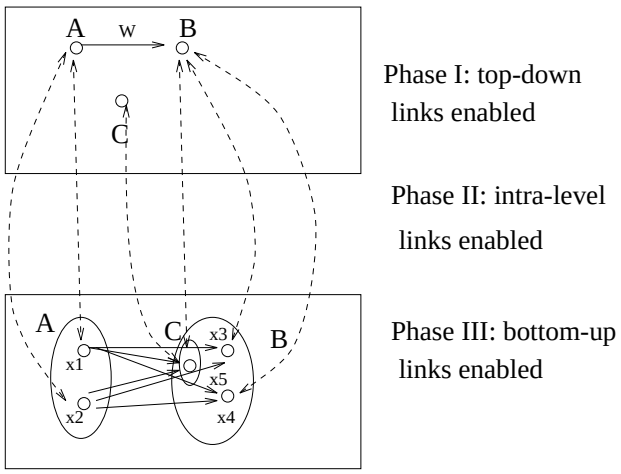

Figure 1: The CONSYDERR Architecture

cal treatments (such as Feldman and Ballard 1982, Smolensky 1988, Clark 1992, Sun 1994), the issue is essentially an empirical one. It is up to experimental work to demonstrate whether a type of representation (especially localist representation) can handle all the symbolic processing tasks and thus correspond to the traditional symbolic models (e.g., Rosenbloom et al 1993, Baars 1988, Anderson 1983, 1993). This issue is not particularly relevant to the present work, since symbolic processing is a dimension that is completely orthogonal to the conscious/unconscious distinction, and the localist/distributed distinction that we employ here.

3 A Model

Based on the four criteria and the earlier discussions, I proposed the model CLARION (which stands for *Connectionist Learning with Adaptive Rule Induction ON-line*). In this model, we use two types of representations, one localist and the other distributed, as discussed before, which embody the "representational difference" view of consciousness. The model was described in detail in Sun et al (1996, 1999), Sun and Peterson (1997, 1998), and Sun (1997). The model consists of two levels: a top level and a bottom level. The essential character of the model is that while the top level of CLARION is localist and thus naturally accessible/explicit, the bottom level contains knowledge embedded in a network with distributed representation and is thus inaccessible/implicit. See Figure 2.

Let us examine the representations. The inaccessible nature of unconscious knowledge can be 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 are capable of accomplishing tasks but are generally uninterpretable and subsymbolic (see Rumelhart et al 1986, Sun 1994). [3]

[3]However, it is generally not the case that distributed representations are not accessible at all but they are definitely less accessible, not as direct and immediate as localist representations. Distributed representations may (or may not) be accessed through indirect, transformational processes. This difference between localist and distributed representations, in my view, accounts for the corresponding difference between the conscious and the unconscious, because the unconscious,

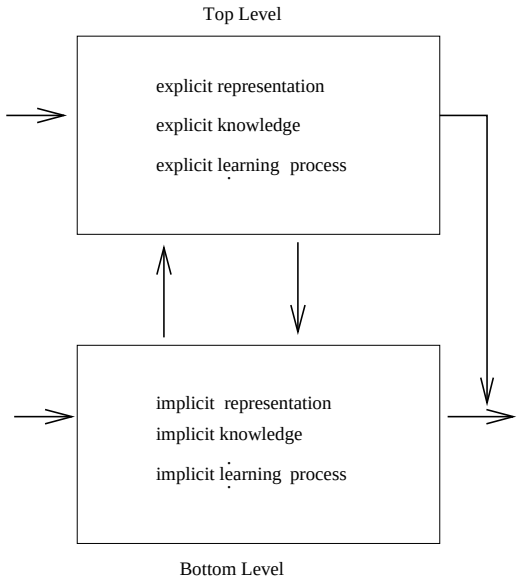

Figure 2: The CLARION Architecture

In contrast, conscious knowledge can be captured in computational modeling by a symbolic or localist representation (Clark and Karmiloff-Smith 1993), in which each unit has a clear conceptual meaning/interpretation (i.e., a semantic label). This captures the property of conscious processes being accessible and manipulable (Smolensky 1988, Sun 1994). [4] This difference in representation leads to the two-level architecture (Sun 1994, 1995, Sun et al 1996, 1997, Sun 1997).

At each level, there are multiple modules (both *action-centered* modules and *non-action-centered* modules; Schacter 1990, Revonsuo 1993, Moscovitch and Umilta 1991). First of all, at the bottom level, action-centered knowledge is highly modular; that is, a number of small backpropagation networks can exist with each adapted to a specific modality, task, or input stimulus type. This is consistent with the well known modularity claim (Fodor 1983; Karmiloff-Smith 1986; Cosmides and Tooby 1994), and is also similar to Shallice's (1972) idea of a multitude of "action systems" competing with each other. Timberlake and Lucas (1993) specified a large set of modules and their inter-relations and transitions (for foraging related activities). Cosmides and Tooby (1994) listed another set of (higher-level) modules for behavior ranging from social contracts (catching cheaters) to syntax of language (parsing sentences). At the top level, the corresponding action-centered knowledge can reside in different modules, in correspondence with the bottom-level structure, or it can reside in more centralized, coarser modules. The knowledge at the top level is more generic, crisp, discrete, and more accessible than its bottom-level counterpart, although, content-wise, it may be similar (or even identical).

On the other hand, the non-action-centered modules of the top level represent more static and more generic type of knowledge. The knowledge there includes what is commonly referred to in psychology as "semantic" memory (i.e., general knowledge about the world in a conceptual, symbolic

although generally inaccessible, may also be brought out in some ways.

[4] Again, accessibility here refers to the direct and immediate availability of mental content for the major operations that are responsible for, or concomitant with, consciousness, such as introspection and forming higher-order thoughts as well as verbal reporting, and meta-level control and manipulation.

form; Tulving 1972). It also includes an episodic memory which stores explicitly recent experiences, with associated spatial-temporal information (Tulving 1972, Bower 1996). Both types of knowledge are represented in a semantic network form (which is a more complex form of the localist network used for representing action-centered knowledge at the top level; Quillian 1968). Likewise, the bottom level contains the corresponding non-action-centered modules that represent the same type of knowledge (but not necessarily the same content), albeit in an implicit (i.e., distributed) form. The encodings in the bottom-level module can be the result of transformation from the corresponding representation at the top level, through turning a localist representation into a distributed representation. [5]

The reason for having both action-centered and non-action-centered modules at each levels is because, as it should be obvious, the action-centered knowledge (roughly, the procedural knowledge) is not necessarily inaccessible (directly), and the non-action-centered knowledge (roughly, the declarative knowledge) is not necessarily accessible (directly). (Although it was argued by some that all procedural knowledge is inaccessible directly and all declarative knowledge is directly accessible, such a clean mapping of the two dichotomies is untenable in my view.) Therefore, there is the need to represent the parts of the action-centered knowledge and the non-action-centered knowledge that are directly accessible at the top level, and to represent the parts of the knowledge (of both types) that are inaccessible directly at the bottom level. Hence, the duplication of modules at both levels is necessary.

The learning of unconscious action-centered knowledge at the bottom level can be done in a variety of ways consistent with the nature of distributed representation. In the learning setting where correct input/output mappings are available, straight backpropagation (a supervised learning algorithm) can be used for each network (Rumelhart et al 1986). Such supervised learning procedures require the a priori determination of a uniquely correct output for each input. In the learning setting where there is no input/output mapping externally provided, reinforcement learning can be used (Sutton 1990, Watkins 1989). Using reinforcement learning, we can measure the goodness of an action through a payoff/reinforcement signal, ranging from, say, 1 to -1 (with 1 being extremely good and -1 being extremely bad and many other possibilities in between). An adjustment can be made to internal weights to increase the chance of selecting the actions that receive positive reinforcement and to reduce the chance of selecting the actions that receive negative reinforcement.

The action-centered conscious knowledge at the top level can also be learned in a variety of ways in accordance with the localist representation used. Because of the representational characteristics, one-shot learning based (mainly) on hypothesis testing (Bruner et al 1956, Nosofsky et al 1993, Sun et al 1996) is needed. We can utilize unconscious knowledge already acquired in the bottom level (i.e., using the *bottom-up* learning; Sun et al 1996). As with unconscious knowledge, we should also be able to dynamically acquire a representation and modify it as needed, to reflect the dynamic on-going nature of the everyday world (especially in skill learning; Heidegger 1927, Sun et al 1996). The basic idea is as follows: if an action chosen (e.g., by the bottom level) is successful (i.e., it satisfies a certain criterion) then the agent constructs a rule (with its action corresponding to the one chosen

[5] For example, if the top-level module contains "I am looking at the car" in a localist fashion, the corresponding bottom-level module contains "I am looking at the car" in a distributed fashion. This re-representation of information in the non-action-centered module at the bottom level from the corresponding non-action-centered module at the top level is useful for achieving self-awareness (when the top-level module extracts from the bottom-level module the following information "I am thinking that I am looking at the car", as will be discussed later).

and with its conditions corresponding to the current sensory state), and adds the rule to the top-level localist network. Then, in subsequent interactions with the world, the agent refines the constructed rule by considering the outcome of applying the rule: if the outcome is successful, the agent may try to generalize the conditions of the rule to make it more universal; if the outcome is not successful, then the conditions of the rule should be made more specific and exclusive of the current case. This is a hypothesis testing process as has been studied (in different contexts) by e.g. Bruner et al (1956) and Nosofsky et al (1993). Other types of learning are also possible (see Appendix).

This hypothesis of the two different learning processes (on top of the hypothesis of the two different types of representations which I have argued for earlier) is consistent with some highly plausible interpretations of relevant findings in the psychological literature (see Sun et al 1996, 1998). Berry and Broadbent (1988) demonstrated this difference using two dynamic control tasks that differed only in the degree to which the pattern of correct responding was salient to the subjects. Results suggested that subjects learned the two tasks in different ways: Subjects in the non-salient condition learned the task implicitly while subjects in the salient condition learned the task more explicitly (as measured by tests of resultant explicit knowledge). Reber (1989) described a similar situation in artificial grammar learning. When complex hierarchical relations were needed in order to judge grammaticality, subjects tended to use implicit learning (improving performance but without generating explicit knowledge). When only pair-wise relations were needed, subjects were more likely to use explicit learning through inducing an explicit rule. Although there may be other possible explanations for these findings, it is highly likely that the differences above reflect the contrast between two separate learning processes.

The combination of the outcomes from the two levels is necessary. The detail of the statistical combination method is described in Appendix. The relative contributions of the two levels in the combination (in learning or performance) may be manipulated to a certain degree, as demonstrated psychologically by e..g. Jacoby et al (1993), and implemented computationally as in Sun (1997) and Sun and Peterson (1998) (see Appendix). See Figure 3 for the whole system. At a particular moment (and in a particular task), whether the top level, the bottom level, or both are used is determined by a number of factors:

- Instructions: Different instructions lead to different combinations of the two levels (that is, with different probabilities, not necessarily deterministically decided). For example, the exclusion/inclusion procedure in Jacoby et al (1993) alters the combination of outcomes from the two levels.
- Situational demands: In circumstances in which explicit explanation is required, subjects can be forced to use the more explicit component, i.e., the top level. See e.g. Sun et al (1998, 1999) for examples.
- Complexity of tasks: If the task is very simple, the top level will likely prevail; if it is sufficiently complex, it is likely that only the bottom level will be employed (as has been discussed in Sun 1997). As mentioned earlier, Reber (1989) and Berry and Broadbent (1988) have shown this phenomenon in their experiments. See Sun (1997) for an explanation of this phenomenon in CLARION based on the model makeup (see also Appendix). [6]

[6]Essentially, the top level, due to its explicit representations and processes, cannot successfully handle complex

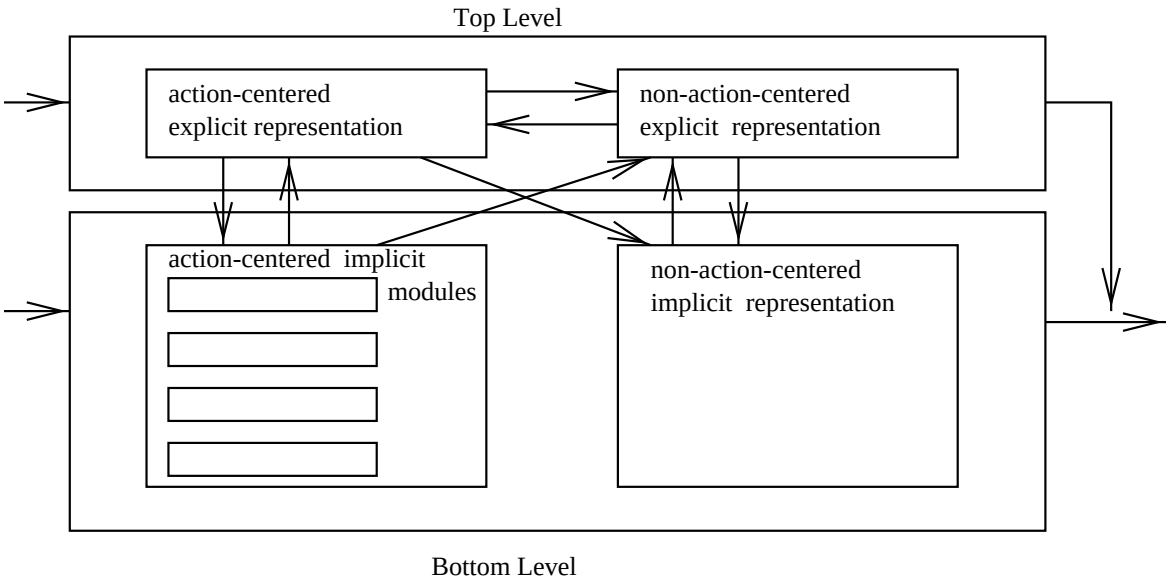

Figure 3: The full CLARION Architecture

- Secondary tasks: When additional tasks are added at the same time when the primary task is being performed, subjects tend to use the bottom level, due to its abilities for handling more complex situations (e.g., see Sun et al 1999).

In terms of learning *directions*, we can distinguish *top-down* learning (assimilation of top-level knowledge into the bottom level; Anderson 1983) and *bottom-up* learning (extraction of explicit knowledge from the bottom level; Sun et al 1996). Which one, or both, or neither, of the two methods will be used is determined by whether the requisite knowledge for performing the task exists first (or is acquired first) in the bottom level, the top level, or both. If it first exists only in the top level, then during the performance of the task, top-down learning will likely occur (Anderson 1983, 1993, Ackerman 1988). On the other hand, if the necessary knowledge first exists only in the bottom level, then during the performance of the task, bottom-up learning will likely occur (Sun et al 1996, Stanley et al 1989). In either of the above two cases, learning within each level can happen separately (Lewicki et al 1987, Bruner et al 1956). If the knowledge coexists or co-develops in both levels, then there might be mixed learning (both top-down and bottom-up), separate learning (learning within each level separately), or no learning at all, depending on the circumstances.

Due to the nature of the distributed representation at the bottom level, it is more sensitive to subtle or complicated forms of information (as demonstrated by, e.g., Elman 1991, Clark 1992, Clark and Karmiloff-Smith 1993, and as observed by Reber 1989, Berry and Broadbent 1988 in human experimental data). Processes operating at this level may be aptly described as associative (James 1890); that is, vague connections between patterns direct the processing (e.g., as described in Sun 1994). Similarity (Tversky 1977) plays a major role here, in forming patterns, connecting patterns, and changing patterns. Thus the processes are more complex and more fuzzy. Because the use of distributed representations, the similarity of two items is the similarity of their representational

situations, but the bottom level can. Thus, the bottom level will prevail in more complex situations.

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

Figure 4: Comparisons of the two levels of the CLARION architecture

patterns (Sun 1995, Sloman 1996), Knowledge or skill at this level is more geared toward specific situations. On the other hand, the top level is more crisp and more discrete, and thus more precise and more reliable (as prescribed by Smolensky 1988 for conceptual processing) and selective (as argued by Hayes and Broadbent 1988). Out of the interrelatedness and the multiplicity of real-world situations, knowledge is extracted as well as abstracted at the top level. Thus it is possible to follow exact rules (as has been argued by Hadley 1995, Sun 1995). The top level can also allow explicit control and manipulation (with its use of explicit representation), including deciding reasoning methods, altering reasoning processes, and controlling reasoning modes. The detail will be discussed further in section 5.2. See Figure 4 for a comparison of the two levels in terms of representations, operations, operational characteristics, and phenomena accounted for. The different characteristics of the two levels makes the combination of the outcomes from the two levels advantageous (Ueda and Nakano 1996, Breiman 1996).

Let us now further clarify the correspondence of consciousness and the two levels in CLARION. Instead of simply equating the top level with the conscious and the bottom level with the unconscious, the full picture will have to be more complex than this. Certainly, the top level, due to its explicitness, facilitates conscious awareness, and the bottom level, due to its implicitness, hinders such awareness (Clark and Karmiloff-Smith 1993). So there is a high degree of correlation between the two dichotomies. The two-level idea provides the representational correlates of the distinction between the conscious and the unconscious. However, it alone, I am afraid, will not be sufficient to completely account for the distinction. First of all, since being accessible does not necessarily mean being accessed, whatever is going on at the top level is not necessarily conscious, but only potentially conscious; that is, it is conscious if it is *being* accessed. Being accessible is a necessary condition of consciousness, but not a sufficient condition. Second, the conscious access at the top level can be either with respect to the process, which is termed *reflective consciousness*, or with respect to the outcome, which is termed *access consciousness*. Third, when we say that the bottom level is not conscious, we mean that the processing details at the bottom level is not directly accessible. However, indirect access can

be obtained through activating corresponding explicit representations at the top level by the bottom level (through some interpretive or transformational processes that turn an implicit representation into an explicit representation; Clark 1992, Hadley 1995).

It should be emphasized that this dichotomous structure of the two levels is minimally necessary for cognitive modeling (in the sense of a *minimal architecture*). The previous arguments show that there are qualitative and fundamental differences between the two types of processing. Given these differences, it is hard to imagine that one level can be derived from the other ontogenetically, without some minimal structure beforehand. Thus, the distinction should be innate somehow. Note also that most of the animal species are not capable of developing an elaborate and complete conceptual system with symbolic processing abilities, while humans rarely fail to develop such a system. The constancy of this inter-species difference points to the innateness of such a difference and the innateness of the two-level structure. It is thus more convincing to hypothesize that the dichotomous structure is a given innate structure for humans, and thus should be incorporated into the architecture.

To sum up, in CLARION, the top level is explicit and conceptual, using localist representation and explicit one-shot hypothesis testing learning, and mostly involved with controlled processing, and thus it is consciously accessible, whereas the bottom level is implicit and subconceptual, using distributed representation, distributed spreading activation, gradual weight tuning for learning, and involved with automated processing, and thus it is inaccessible (Shiffrin and Schneider 1977, Anderson 1983, Hunt and Lansman 1986, Ackerman 1988, Reber 1989, and Sun 1994). Here is a summary of the basic model postulates:

- Representational difference: The two levels employ two different types of representations and thus have different degrees of conscious accessibility.
- Learning difference: Different learning methods are used for the two levels and thus the two levels have different learning characteristics.
- Manipulability of the interaction: The combination of the outcomes from the two levels can be altered based on task situations.
- Action-centered vs. non-action-centered representations: Separate action-centered modules and non-action-centered knowledge representation modules coexist at the top level; similarly, action-centered modules and non-action-centered knowledge representation modules coexist at the bottom level.

These postulates together constitute the essence of CLARION (Sun et al 1996). For further details, see Appendix, as well as Sun et al (1996, 1998, 1999) and Sun and Peterson (1997, 1998).

4 Accounting for the Physical Basis of Consciousness

Reber (1989), based on psychological data, has already hypothesized that the primary difference between the explicit and implicit learning processes (which map onto the conscious and unconscious

processes) lies in the forms of their representations. Lewicki (1992) and Squire et al (1993) had similar views in interpreting their data. My view presented here is an extension of these previous conjectures. The advantage of the representational difference (localist+distributed) explanation of consciousness is that it provides a mechanistic (computational) distinction that explains the phenomenological distinction between the conscious and the unconscious, thus grounding a set of vague notions needing explanation in another set of notions that are much more tangible (i.e., physical and computational) and much more fundamental. Thus, it accounts nicely for the physical basis of consciousness.

To further demonstrate the promise of the representational difference explanation of consciousness as embodied in CLARION in accounting for the physical basis of consciousness. a comparison with alternative views is in order. First of all, let us look into the views that are based on recognizing that there are two separate systems in the mind.

- *The SN+PS view*: as proposed by Anderson (1983) in his ACT* model, there are two types of knowledge. The difference, in this view, lies in the two different ways of organizing knowledge: whether the knowledge is organized in an action-centered way (procedural knowledge) or in an action-independent way (declarative knowledge). However, both types of knowledge are represented symbolically (using either symbolic semantic networks or symbolic production rules). Loftus 1975) to activate

- *The $PS + SN$ view*: as proposed by Hunt and Lansman (1986), the "deliberate" process of production matching and firing, which is serial, is assumed to be a conscious process, while the spreading activation (Collins and Loftus 1975) in semantic networks, which is massively parallel, is assumed to be an unconscious process.

- *The algorithm + instance view*: as proposed by Logan (1988) and Stanley et al (1989), instance retrieval and use are considered to be unconscious (Stanley et al 1989) or automatic (Logan 1988), while the use of "algorithms" involves conscious awareness, either during "algorithmic" processes (Logan 1988) or in terms of the product (i.e., the knowledge resulting from the "algorithmic" processes; Stanley et al 1989).

- *The two-pathway view*: there have been various proposals in neurobiology that there are different pathways in the brain, some of which lead to conscious awareness, while others do not. For example, see Milner and Goodale (1995), Damasio et al (1990), and LeDoux (1992).

- *The connection/disconnection view*: as suggested by e.g. Schacter (1990) and Revonsuo (1993), in the human brain, there are multiple modules each of which performs specialized processing without incurring conscious awareness of the processes, and there is a separate module that is solely responsible for conscious awareness. Each of the specialized modules can send its output to the conscious module and thus makes the output consciously accessible.

The problem with the $SN + PS$ view (as in ACT*; Anderson 1983) is that both types of knowledge (declarative and procedural) are represented in an explicit, symbolic form and thus it did not explain, from a representational viewpoint, the differences in conscious accessibility between the two types of knowledge. Although knowledge organization is apparently different between semantic networks and

production rules (with different degrees of action-centeredness), the difference is insufficient to account for the qualitative difference in conscious accessibility, because both are symbolically represented and thus fundamentally the same. The difference in conscious accessibility is thus simply *assumed* instead of being *intrinsic* in terms of something more fundamental. There is no theoretical reduction of accessibility/inaccessibility to any fundamental mechanistic notions. Another way of viewing the difference between declarative knowledge and procedural knowledge (Anderson 1983) is in terms of processing: one uses spreading activation (Collins and Loftus 1975) and the other uses rule matching and firing (Klahr et al 1986). However, there is no *fundamental* qualitative difference in the mechanisms for accomplishing the two processes, both of which involve similar symbolic manipulation. [7]

The $PS + SN$ view (Hunt and Lansman 1986) is almost the exact opposite of the $SN + PS$ view advocated by Anderson (1983). The problem with the view is also similar: there is no mechanistic difference that is fundamental enough that can account for the qualitative difference between the conscious and the unconscious, since both representations are symbolic and the two processing mechanisms (rules matching/firing and spreading activation) are highly similar (as discussed earlier). The algorithm+instance view (Logan 1988, Stanley et al 1989) is almost identical to the $PS + SN$ view (Hunt and Lansman 1986), and thus suffers the same problem.

On the other hand, the problem with the *connection/disconnection* view is that there is no explanation with regard to why there are two qualitatively different types of modules: one conscious and the other unconscious. There is nothing inherent in the view that can help to shed light on the difference between the conscious and the unconscious in terms of different mechanistic processes underlying them. Similarly, the problem of the two-pathway view is that, although there is also ample biological evidence that indicates the existence of multiple pathways (in visual, language, and other processing modes) and some are correlated with conscious awareness while some others are not (as described in Milner and Goodale 1995 and LeDoux 1992), there is no *explanation* of why some result in consciousness while other do not, beside the fact that they are involved with different neuropathways (which constitutes a verification/confirmation but does not constitute an explanation).

In contrast to the above two-systems views, there are also some views that insist on the unitary nature of the conscious and the unconscious; that is, they share the belief that the conscious and the unconscious are the different manifestations of the same underlying system (or process). The difference is thus based on the difference between processing modes:

- *The threshold view*: as proposed by several researchers, including Bowers et al (1990) and Dienes and Berry (1997), the difference between the conscious and the unconscious can be explained by the difference between activations of mental structures to a level above a certain threshold and the activations of such structures below that threshold. When the activations reach the threshold level, an individual becomes aware of the content of the activated structures; otherwise, although the activated structures may influence behavior, they will not be accessible consciously.

- *The chunking view*: as described by Servan-Schreiber and Anderson (1987) and Rosenbloom et al (1993), a chunk is considered a unitary representation and its internal working is oblique,

[7] The new model (ACT-R) proposed by Anderson (1993) espoused an *instance* + PS *view*, which suffers the same problem.

although its input/output are accessible. Thus, according to this view, the difference between the conscious and the unconscious is the difference between using multiple (simple) chunks (involving some consciousness) and using one (complex) chunks (involving no consciousness).

- *The coherence view*: as suggested by e.g. Baars (1988) and others, some sort of coherence (e.g., the activation of a coherent patterns of representations, coherent firing of neurons, or coherent and stable activations of a neural network) in the dynamics of the mind (or the brain) gives rise to consciousness. The distinction of the conscious and the unconscious is reduced to the distinction between coherent activities and incoherent activities in the mind/brain. Mozer (1996) proposed a special case of this view in which being in a stable attractor in a dynamic system (or a neural network in particular) leads to consciousness. Crick and Koch (1990) proposed another special case in which synchronous firing of neurons leads to conscious awareness.

With the unitary views, no fundamentally different processes or mechanisms are needed to explain the difference between the conscious and the unconscious. Therefore, there seems to be an elegant parsimony of theoretical constructs, which is certainly appealing. The problem with the unitary views in general, however, is that there is the same lack of fundamental, qualitative difference, as with the aforementioned two-system views, between whatever is used to account for the conscious and whatever else is used to account for the unconscious, in contrast to the fundamental phenomenological difference between the conscious and the unconscious. Whether it is coherence/consistency, or synchrony, or reverberation, or above-threshold activations, it is a leap of faith to believe that it can lead to conscious awareness, let alone the phenomenal experience of consciousness. In the chunking view, although the equation of the inaccessibility of the internal working of a chunk with the inaccessibility of the unconscious is appealing, the problem with the chunking view is that the assumption of inaccessibility of the internal working of a chunk is not backed up by any mechanistic explanation of that inaccessibility, in terms of intra-chunk and inter-chunk processes. It is *assumed*, rather than explained by something more fundamental. As a result, there is no theoretical reduction being accomplished.

In addition, the problem with the biologically-motivated views (such as Damasio 1994. Milner and Goodale 1995, Crick and Koch 1990) in general is that the gap between phenomenology and physiology/biology is so great that something else is needed to bridge the gap. Otherwise, if we rush directly into complex neural physiological thickets (Taylor 1994, Edelman 1989, Damasio et al 1990, LeDoux 1992, Crick and Koch 1990), we may lose sight of forests. Computation, in its broadest sense, can serve to bridge the gap. It provides an intermediate level of explanation, in terms of mechanistic processes, and helps to determine how various aspects of the conscious and the unconscious figure into the architecture of the mind (that is, focusing on the essential framework rather than irrelevant details). There is evidence to suggest that a middle level between phenomenology and physiology/neurobiology might be more apt at capturing fundamental characteristics of consciousness.

Judging from the above analysis, all of these afore-critiqued views are flawed or inadequate in some way. In comparison, we found that the representational difference view had a distinct edge. Specifically, the advantage of the view lies in the explanation of consciousness in terms of a mechanistic (computational) distinction, reducing a set of phenomenological notions to a set of mechanistic notions, i.e., the reduction of the dichotomy of the conscious and the unconscious to the more tangible (physical)

dichotomy of the localist (symbolic) representation and the distributed representation. We thereby arrive at a promising framework for accounting for the physical basis of consciousness.

This viewpoint based on the localist/distributed representation may at times seem at odds with existing neurobiological knowledge. For example, it is believed that the ventral visual pathway is responsible for consciousness while it is understood that the area employs rather "distributed" representations (Milner and Goodale 1995). The answer to such apparent paradoxes is two-fold: (1) the localist/distributed distinction applies only at certain levels of abstraction, not necessarily at all possible levels of description; for example, at the molecular level all the representations have to be 'distributed". (2) localist representation has many variations, as discussed by Feldman and Ballard (1982), some of which may seem resembling distributed representation to some extent (e.g., when *replicated* localist nodes are distributed spatially); such seemingly "distributed" representations are actually localist (that is, spatial characteristics are not determining factors of representation as used here; see Feldman and Ballard 1982 for full characterization). Thus a completely dogmatic view on localist representation is unwarranted.

5 Accounting for the Functional Role of Consciousness

5.1 Access Consciousness

"Access consciousness" refers to the availability of the mental content for access (and/or verbal report), while "reflective consciousness" refers to the availability of the *process* of mental activities for access (and/or verbal report). That is, if access consciousness allows one to access the outcome of a reasoning process, reflective consciousness allows one to access the process, or steps, of reasoning. With regard to the functional role of access consciousness, there have been various suggestions:

- The veto view: As suggested by Libet (1985), the function of consciousness is to be able to veto unconsciously initiated actions. In his physiologically derived theory, an action is initiated unconsciously, and 250 ms after the unconscious initiation, there is a window of 100 ms in which consciousness can choose to veto the initiated action.

- The counterbalance view: The function of consciousness is to counterbalance the unconscious processes. [8] This view of consciousness as providing a "counterbalance" to unconsciousness seems to be a generalization of Libet's view on the veto function of consciousness (which is a form of counterbalancing).

- The situational difference view: As shown by Reber (1989), Stanley et al (1989), and others, conscious and unconscious processes are suitable for different situations (especially for different learning situations), so that either a conscious or an unconscious process will be applied to a

[8] Kelley and Jacoby (1993) showed in their experiments, through contrasting "aware" vs. "unaware" conditions, that conscious awareness has a distinct causal role on subsequent behavior. The two different conditions produced different causal attributions by subjects, and consequently, different causal attributions led to different actions on the part of the subjects.

situation depending on which one is most suitable to the situation. [9]

- The language/planning view: As suggested by Crick and Koch (1990), the function of consciousness is to enable the use of language and (explicit) planning, due to its explicitness and accessibility. However, the questions remain: Why should we use language and planning in an explicit and conscious way? Why can't the function of language and planning be achieved without conscious awareness?

- The synergy view: As suggested in Sun (1994, 1995, 1997), the function of the distinction between the conscious and the unconscious lies in the flexibility and synergy that this distinction affords. As shown in Sun (1994, 1997), the interaction of the conscious and the unconscious (as two distinct processes) can (in many common circumstances) lead to a synergy in learning and performance. Let us discuss in more detail this view below.

The available psychological data (e.g., on implicit learning and implicit memory) suggest that conscious processes tend to be more crisp and focused (selective), while unconscious processes tend to be more complex, broadly scoped (unselective), and context-sensitive (see Reber 1989, Berry and Broadbent 1988, and Seger 1994 regarding complexity; see Hayes and Broadbent 1988 regarding selectivity). Similar points have been made by e.g. Baars (1988), Taylor (1997), McClelland et al (1994), and Sun (1997). Due to their vastly different and contrasting characteristics, it should not come as a surprise that the interaction of the conscious and the unconscious leads to synergistic results (Breiman 1996). [10] There is some psychological evidence directly in support of the synergy view. See, e.g., Willingham et al (1989) and Stanley et al (1989).

Furthermore, the synergy view can encompass all the aforementioned views regarding consciousness. According to the synergy view, consciousness can certainly veto or counterbalance unconsciousness, given the right circumstances when such veto or counterbalance improve the overall performance (that is, if they lead to synergy), or if explicit instructions dictate it (Jacoby et al 1993). Likewise, the synergy view can explain the situational difference view, in that in some extreme cases, it may be advantageous to use only conscious or unconscious processes (but in general, both types of processes are present due to their synergistic effect). The synergy view can also encompass the language/planning view, because it explains why one should use conscious language/planning processes on top of unconscious processes: It is because of the possibility of improved overall performance through using both types of processes.

I will review some results obtained from experiments with CLARION that supports the synergy view. The details of the experiments and complete data can be found in Sun and Peterson (1997, 1998) and Sun et al (1996, 1998). Briefly, in a simulated maze running task, in terms of learning speeds, the differences between the bottom level alone (with implicit, unconscious learning) and the

[9] In their psychological experiments, subjects tend to use explicit learning (with consciousness awareness) if the situation was simple (e.g., a sequence in which the next position can be unambiguously predicted by the immediately preceding position; Reber 1989), and use implicit learning (unconsciously) if the situation was more complex.

[10] In statistical literature, it is well known that combining diversified processes (e.g., estimators) can improve performance (Breiman 1996, Ueda and Nakano 1996). (However, clearly, it is not *always* the case that combination leads to improved performance.)

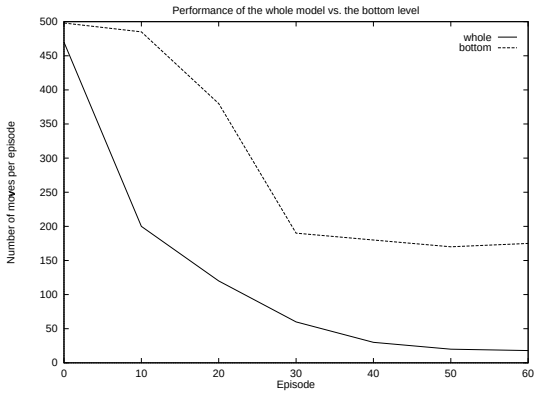

Figure 5: The synergy effect in CLARION in the maze task (Sun et al 1996). The synergy effect is revealed by comparing the whole model with the bottom level alone.

whole CLARION system (which includes both types of processes: implicit/unconscious and explicit/conscious) were very significant. CLARION outperformed the bottom level alone by large margins, which indicated that utilizing both processes helped to speed up learning. In terms of learned performance, CLARION outperformed the bottom level alone by large margins again, indicating utilizing both processes helped to improve learned performance too. We also compared the learned performance of the bottom level after it was trained together with the top level (i.e., the entire system were trained together), with the performance of the bottom level trained alone after an equal number of training episodes, and discovered that training the whole system together not only improved the performance of the whole system, but it also improved the bottom level when it was included as part of CLARION. We also assessed the performance of trained models in a new and larger maze (i.e., testing the transfer ability). We discovered that CLARION transferred much better than the bottom level alone, after the same number of training episodes. This showed that incorporating both processes helped transfer. Furthermore, by comparing the corresponding transfer performance of the top level, the bottom level and the whole CLARION model, after training the whole system together, it was clear that often the top level alone performed better in transfer than the bottom level alone, as well as than the whole CLARION system together. This showed that acquiring explicit (conscious) knowledge facilitated transfer. Taken together, the results showed the combination of the two processes helped performance in various aspects. (Note that all the differences reported were statistically significant, as demonstrated by t tests.) See Figure 5, which compares the performance of the whole CLARION model with that of the bottom level alone.

In another task, a simulated minefield navigation task, as reported in Sun and Peterson (1998), in terms of learning speeds, the superiority of the whole CLARION system over the bottom level alone was statistically significant. This again indicated incorporating both explicit and implicit (conscious and unconscious) processes helped to speed up learning. To assess transfer, after training models on 10-mine minefields, we assessed performance of these models in new minefields that contained 30 mines and 60 mines. CLARION outperformed the bottom level alone. So again, incorporating both processes helped to facilitate transfer of learned skills. Our subsequent experiments with human subjects in this task confirmed that the same synergy effect exists in human performance (Sun et al 1998). See Figure 6, which demonstrates the synergy in CLARION and in the human data by comparing the

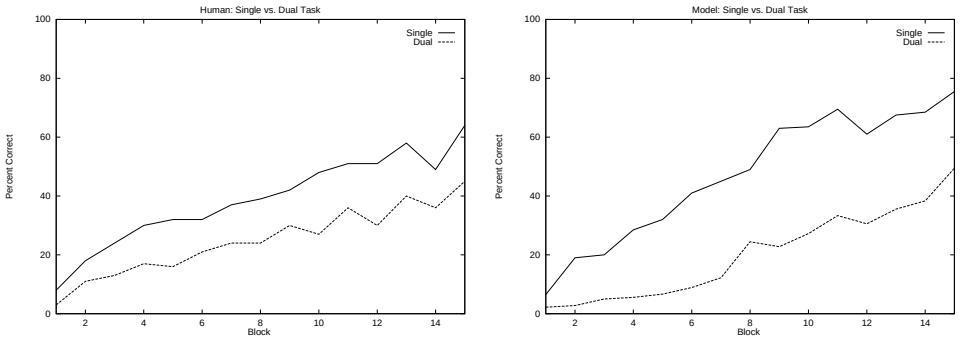

Figure 6: The comparison of the synergy effect in CLARION and in the human performance in the navigation task (Sun et al 1998). The synergy effect is revealed in both cases by comparing the subjects trained in a regular setting (involving both levels) and the subjects trained in a dual task setting (involving mostly the bottom level). The left panel contains averaged human data, and the right averaged model data.

performance of the bottom level alone with the performance of both levels. In all, CLARION was able to demonstrate the synergy effects between conscious and unconscious processes hypothesized earlier.

Quantitative simulation has also been done to capture the synergy effects revealed in the human experiments of Willingham et al (1989) and Stanley et al (1989). (See Sun and Terry (1998) for details.) Willingham et al (1989) found that those subjects who acquired explicit (conscious) knowledge in a serial reaction time task appeared to learn faster than those who did not have explicit knowledge (relying less on conscious processes). Stanley et al (1989) reported that, in a dynamic control task, subjects' learning improved if they were asked to generate verbal instructions for other subjects along the way during learning (because verbalization led to more reliance on conscious processes, which would otherwise less likely be engaged in this task). [11] Furthermore, Willingham et al (1989) found that subjects who verbalized (while performing serial reaction time tasks) were able to attain a higher level of performance than those who did not verbalize. Willingham et al (1989) also reported that subjects who acquired explicit knowledge in a training task tended to have faster response times in a transfer task. See Figures 7 and 8, which compare the data from CLARION with the data of Willingham et al (1989) and Stanley et al (1989), respectively. Both the model and human data demonstrated the synergy effects.

Yet another case of synergy that CLARION captures is the exclusion and inclusion procedure of Jacoby et al (1993). In the exclusion condition, subjects were told to respond if they did not consciously remember a stimulus. In the inclusion condition, subjects were told to respond if they did consciously remember a stimulus. The final result is synergistic of the two types of processes, in the sense that it is dependent on both types of processes and cannot be done without the presence of either. CLARION allows the explicit manipulation of the combination process, through adjusting the parameters governing the combination (see Sun and Peterson 1997, 1998; Appendix), and thus the phenomenon demonstrated by the exclusion/inclusion procedure can be easily accounted for by CLARION.

[11] That is, subjects were able to speed up their own learning through an emphasis on the conscious processes.

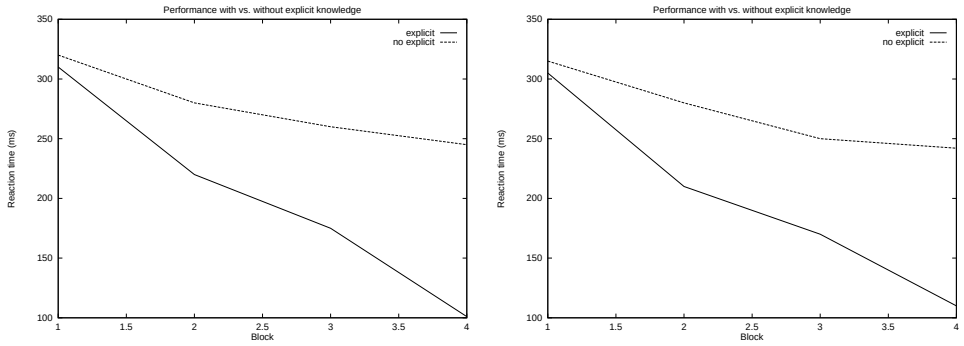

Figure 7: The comparison of the synergy effect in CLARION and in the human data from Willingham et al (1989). The synergy effect is revealed in both cases by comparing the subjects with explicit knowledge (involving both levels) and the subjects without explicit knowledge (involving mainly the bottom level). The left panel contains averaged human data, and the right averaged model data.

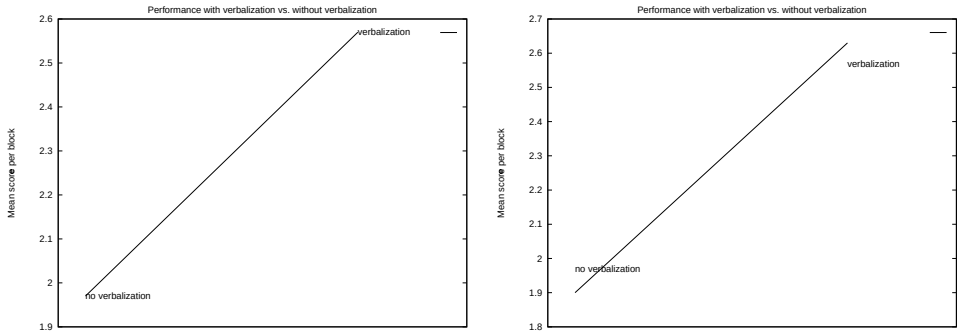

Figure 8: The comparison of the synergy effect in CLARION and in the human data from Stanley et al (1989). The synergy effect is revealed in both cases by comparing the subjects with verbalization (involving both levels) and the subjects without verbalization (involving mainly the bottom level). The left panel contains averaged human data, and the right averaged model data.

The synergy effects revealed in psychological experiments vary with different settings, which may be differentially benefited from the interaction of conscious and unconscious processes. Special cases (and exceptions) abound in the psychological literature, such as implicit learning, automatic processing, unconscious perception, and implicit memory. Although different, these cases are nevertheless consistent with the synergy view of consciousness. CLARION can account for these alternative situations as follows:

- *Automaticity* is roughly the situation in which the execution of actions does not requires explicit processes (Shiffrin and Schneider 1977, Navon and Gopher 1977, Logan 1988, Meyer and Kieras 1997). CLARION accounts for automaticity with a setting in the model in which only bottom-level modules are used for executing a task, leaving the top-level modules for other use, which happens when a task has been well learned and thus has no more need for synergy. In relation to Logan's (1988) account of automaticity (that is, while controlled processes involve explicit processes or "algorithms", automatic processes involve only reactivating past memory), we can equate activation of memory with the activations in the bottom level of CLARION, in which the weights formed from past experience constitute implicit memory of past experience, and equate algorithmic processes in Logan (1988) with the explicit processes at the top level of CLARION.

- Purely implicit learning (Reber 1989, Berry and Broadbent 1988, Lewicki et al 1992) occurs when a situation is so complex that no explicit learning processes at the top level can adequately handle it (as demonstrated by e.g. Reber 1989, Berry and Broadbent 1988 through using tasks of different complexities). In CLARION, this occurs when situations are sufficiently complex so that the top-level learning mechanisms fail to work adequately. The top level does not lend itself easily to the learning of complex structures because of its crisp, individuated representation and the rigid hypothesis testing learning process. On the other hand, in the bottom level, the distributed network representation handles the complex relations better (Sun 1997, Sun et al 1996). This point can be easily seen upon examining the details of the two levels in the model (see Appendix). Our simulation of Lewicki et al (1987), Stanley et al (1989), and Willingham et al (1989) has also demonstrated this point (see Sun and Terry 1998).

- Implicit memory is said to occur when prior information affects the performance of subjects without the subjects being consciously aware of it (Schacter 1987, Jacoby et al 1993). This can be explained with the two-level framework of CLARION. Knowledge in CLARION is separately represented in both the top level and the bottom level. While the top-level information can be lost due to the crisp, individuated representation, the bottom-level information is much more stable due to the distributed representation across a large number of weights and the slow change of weights (requiring multiple updating). Thus, while the conscious processes may lose access to the explicitly represented information at the top level, the unconscious processes retain access to the implicitly represented information at the bottom level, which leads to the implicit memory phenomena. (This account is essentially the same as that conjectured by Bower (1996).)

- Unconscious perception (Merikle 1992) can again be explained within the CLARION framework of two separate levels with localist and distributed representations respectively. While the top level fails to detect weak signals because of its crisp representation and rigid processing, the bottom

level may register the information because of its distributed representation and more flexible processing that can accommodate fuzzy, weak signals due to the inherently similarity-based nature of its distributed representations.

- *Intuition* (Bowers et al 1990) refers to the unconscious reasoning processes. In CLARION, the processes of implicit memory, implicit learning, unconscious perception, and automatization all lead to unconscious information being registered in the bottom level and consequently the (necessarily) unconscious use of that information when the circumstances require the use of it. This leads to what is commonly referred to as intuition.

Some have speculated on the role of *global* access, that is, the simultaneous availability of all relevant mental contents (Baars 1988). In the present model, we do not emphasize this aspect, because we believe that global accessibility is a by-product of the simple accessibility discussed above. There are many possible computational mechanisms for achieving global accessibility. Baars (1988) presented a well-developed model that may be utilized.

5.2 Reflective Consciousness

The function of reflective consciousness lies in explicit control and manipulation of mental processes, which adds meta-level processes on top of regular processes. Such control and manipulation can include, for example, selecting a reasoning method, controlling the direction in which reasoning goes, enable/disable certain inference, or evaluating the progress of reasoning. When meta-level processes get assimilated into regular processes, meta-level processes on top of them can be developed. Thus, potentially, we can have many levels of self-control and self-manipulation of mental processes. Note that I am not claiming that unconscious processes cannot be controlled and/or manipulated, but that it is more difficult to do so due to distributed representations, and hence much less control and manipulation, if any, are applied to the unconscious processes. Therefore, we see the difference between the conscious and the unconscious in terms of meta-level control and manipulation.

CLARION can allow many kinds of explicit control and manipulation at the top level with its use of crisp, discrete, and individuated encoding of separate entities, aspects, or events. Among these meta-level processes are the following:

- Deciding reasoning methods: the reasoning methods that can be adopted include forward reasoning, backward reasoning (that is, deduction and abduction), counterfactual reasoning (which especially requires meta-level control), and other verbally based reasoning methods (potentially rendering the full power of systematicity; Fodor 1975, Fodor and Pylyshyn 1988).

- Altering reasoning processes: reasoning processes at the top level can be altered by meta-level regulations such as blocking some nodes (i.e., blocking the concepts represented by these nodes) by raising their thresholds, enabling some nodes by lowering their thresholds, and consequently directing the reasoning processes into certain regions in certain directions.

- Controlling reasoning modes: we can change the threshold parameters globally, so that the readiness to reason changes accordingly, e.g., from the credulous mode to the suspicious mode.

These different modes require different levels of support for the conclusions to be drawn or the decisions to be made.

The details of these meta-level control processes have been studied in AI, e.g. by Russell and Wefald (1989) and Etzioni (1991). The relevant psychological work includes cognitive studies on executive functions and meta-cognition (such as Metcalfe and Shimamura 1994 and Nelson 1993) and behavioral studies on self control (such as Rachlin 1994). The regulating processes operate on explicit representations at the top level of CLARION, although possibly a combination of top-level and bottom-level modules helps to carry out the control/manipulation processes on the top-level representation.

In addition, the exclusion and inclusion procedure of Jacoby et al (1993) also involves explicit meta-level control (in the psychological experiments described earlier). In this case, the combination between the two types of processes is manipulated through verbal instruction. In CLARION, this is achieved through the setting of the combination parameters.

Summing up the discussion of access and reflective consciousness in CLARION, we see that the conscious processes at the top level of CLARION are characterized by explicit (localist/symbolic) representations, as well as explicit meta-level regulations (i.e., the control and manipulation of the processes operating on the explicit representation). These two aspects together, according to the CLARION model, distinguish conscious processes from unconscious processes. The functional role of access and reflective consciousness follows directly from these two aspects.

6 Relations to Other Models

McClelland et al (1994) propose that there are complementary learning systems in hippocampus and neocortex. The human brain solve the problem of "catastrophic interference" (i.e., later training will destroy previously acquired knowledge) by storing new information in a separate memory system in the hippocampus and later assimilating it into cortical memory systems. In the hippocampus, which is an explicit memory system, 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 information stored in the hippocampus is assimilated into cortical systems. 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. Thus, the assimilation of new information is interleaved with the assimilation of all other information in the hippocampus and with the ongoing events. Using distributed representations, 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. 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, the difference between this model and CLARION is that while this model captures only what I call top-down learning, that is, learning that proceeds from the conscious to the unconscious, CLARION can capture both top-down learning and bottom-up learning (see Sun et al 1998, 1999 for details of bottom-up learning).

Bower's (1996) model can be viewed as a special case, or an instantiation, of CLARION in the context of modeling 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 in CLARION, respectively. In addition to making the distinction between type-1 and type-2 connection, Bower (1996) also endeavored into specifying multiple pathways of spreading activation in the bottom level. These pathways were of phonological, orthographical, semantic, and other types that stored long-term implicit knowledge as weights on links that connected relevant nodes involved. On the other hand, associated with type-2 connections (in the top level), it was claimed that rich contextual information was stored. These details nicely complement the specification of CLARION and can thus be easily incorporated into the model. However, although these details were useful, Bower (1996) did not explain mechanistically the distinction between the two types of processes (i.e., the conscious and unconscious difference between the two types of processes involved).

Let us also look into the declarative/procedural knowledge models (Anderson 1983, 1993). ACT* 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 does not address. First, while ACT takes a top-down approach towards learning (i.e, from given declarative knowledge to procedural knowledge), CLARION can proceed completely bottom-up. Thus, CLARION can account for implicit learning better than ACT (see Sun 1997 for details). Second, in ACT both types of knowledge are represented in an explicit, symbolic form (i.e., semantic networks and productions), and thus it does not explain, from a representational viewpoint, the differences in conscious accessibility. CLARION accounts for this difference based on the use of two different forms of representations.

Comparing CLARION with Hunt and Lansman's (1986) model, there are also similarities. The production system in Hunt and Lansman's model clearly resembles the top level in CLARION, in that explicit manipulations are used in much the same way as in the top level of CLARION. Likewise, spreading activation in the semantic network in Hunt and Lansman's model resembles that in the connectionist network in the bottom level of CLARION, although the representations in Hunt and Lansman's model are symbolic, not distributed. Because of the uniform representation in Hunt and Lansman's model, it does not succeed in explaining the qualitative difference between the conscious and the unconscious, as analyzed earlier.

We can also compare CLARION with Schacter (1990)'s model, which is based on neuropsychological findings of dissociation of different types of knowledge. It is similar to CLARION, in that it includes a number of "knowledge modules" that perform specialized and unconscious processing (analogous to the 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. Schacter's model did not elucidate the qualitative distinction between the conscious and the unconscious in that the "conscious awareness system" lacks any qualitative difference from the unconscious systems. Similarly, in Baars (1988)'s model, a large number of specialist processors perform unconscious processing and a global workspace coordinates their activities through global broadcasting to achieve consistency

and thus conscious experience. The model bears some resemblance to CLARION, in that unconscious specialist processors in that model can be roughly equated to the modules in the bottom level of CLARION, and global workspace may be roughly captured by the top level, which "synthesizes" the bottom-level modules and is essential to conscious processing. Global broadcasting in Baars' model can be viewed as the interaction of the two levels of representations (with the bottom-level representations dispersed within multiple modules) in CLARION, which does produce somewhat consistent outcomes. One difference is that CLARION does not emphasize as much internal consistency (Marcel 1983), which we believe to be limited as a phenomenon in consciousness and may have only limited roles in the emergence of consciousness. The shortcoming of Baars' model is that there is no reduction (explanation) of the fundamental phenomenological distinction of the conscious and the unconscious to some more fundamental distinctions, beside the hypothesis of global information accessibility (due to global broadcasting).

Damasio's neuroanatomically motivated model (Damasio 1990) hypothesized the existence of many "sensory convergence zones" that integrated information from individual sensory modalities through forward and backward synaptic connections and the resulting reverberations of activations, without the need for a central location for information storage and comparisons; it also hypothesized the global "multi-modal convergence zone", which integrated information across modalities also through reverberation via recurrent connections. In relation to CLARION, different sensory convergence zones may be roughly captured by the bottom-level modules in CLARION, each of which takes care of sensory inputs of one modality, 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 explicitly. The process of reverberation (Damasio 1994, Taylor 1994) may be captured in CLARION through recurrent connections within modules and through multiple top-down and bottom-up information flows across the two levels, which lead to some unity of consciousness that is the synthesis of all the information present (Marcel 1983, Baars 1988).

Let us address the issue of the relation between the abstract distinction of localist and distributed representations (posited in CLARION) and the biological brain mechanisms (as we know them currently). First of all, currently, both psychological and biological understanding of the human mind is rudimentary (although biology and brain sciences are making rapid progress). We thus need evidence and ideas from all of these disciplines: psychology, biology, and computational modeling, among some others, in order to make better progress in understanding high-level issues such as consciousness. Moreover, given the preliminary nature of current biological understanding and biology-based theorizing, it is useful to go beyond them when we search for ideas and explanations, even when such ideas and explanations may appear to be at odds with biological theorizing. This is because biology has only explored a small part of the vast space of biological mechanisms and processes and its understanding of even that small part may not be complete and deep enough to reveal all the intricacies. Different approaches can serve complementary roles. Finally, at this time, generally it is not prudent to dismiss any model or theory based on current, limited biological understanding (despite the fact that such understanding is becoming increasingly relevant). Computational modeling on the basis of psychological data, without being mapped completely to current biology, is important, in that it broadens the scope of exploration, and thus it may provide useful ideas and insight and may even constrain and

guide biological studies in the future. The distinction of localist and distributed representations may be useful exactly in this sense.

7 Concluding Remarks

In this paper, I focused on the issue of the physical (mechanistic or computational) basis of consciousness, in the framework of a mechanistic (computational) account of consciousness, and consequently the issue of the functional roles of consciousness in this framework. There have been many different views in cognitive science concerning consciousness, under various guises as theories for implicit memory, implicit learning, automaticity, and so on. Through comparing with various existing views and perspectives, a clear candidate for explaining the physical basis of consciousness and for explaining its functional roles emerged, which was based on the representational difference between the conscious and unconscious processes and embodied in the model CLARION. Analyses and arguments in favor of this view and the model were presented. They showed that the difference between localist (symbolic) representation and distributed representation as employed in CLARION led to a plausible account of consciousness and its functional roles. Subsequently, CLARION was also used to account for other issues related to the function of consciousness. In all, this paper presented a new perspective on consciousness.

A Details of the Model

A pseudo-code algorithm that describes the action-centered part of CLARION, is as follows:

1. Observe the current state x.
2. Compute in the bottom level the Q-value of each of the possible actions (a_i's) associated with the perceptual state x: $Q(x, a_1)$, $Q(x, a_2)$, ..., $Q(x, a_n)$.
3. Find out all the possible actions (b_1, b_2, ..., b_m) at the top level, based on the the perceptual information x and other available information (which goes up from the bottom level) and the rules in place at the top level.
4. Compare the values of a_i's with those of b_j's (which are sent down from the top level), and choose an appropriate action a.
5. Perform the action a, and observe the next state y and (possibly) the reinforcement r.
6. Update the bottom level in accordance with the *Q-Learning-Backpropagation* algorithm, based on the feedback information.
7. Update the top level using the *Rule-Construction-Refinement* algorithm.
8. Go back to Step 1.

In the bottom level, for representing action-centered knowledge, each module calculates Q-values. 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. We can choose an action based on Q-values. At each step, given the input x, we first compute the Q-values for all the possible actions ($Q(x, a)$ for all a's). We then use the Q-values to decide probabilistically on an action to be performed, which is done by a Boltzmann distribution

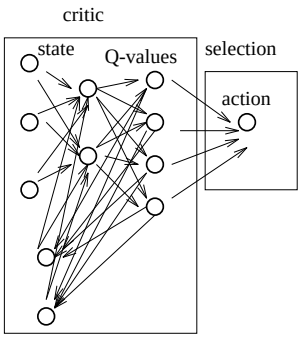

Figure 9: The Q-learning method implemented in a connectionist network. The first three layers constitute a backpropagation network. The output nodes produce Q-values. The fourth layer performs stochastic decision making from all the Q-values.

of Q-values:

$$p(a|x) = \frac{e^{Q(x,a)/\alpha}}{\sum_i e^{Q(x,a_i)/\alpha}} \tag{1}$$

Here α controls the degree of randomness (temperature) of the decision-making process. This method is also known as Luce's choice axiom. This method is cognitively well justified: it is found to match psychological data in a variety of domains.

The calculation of Q-values for the current input with respect to all the possible actions in an action-centered module is done in a connectionist fashion through parallel spreading activation and thus highly efficient (such spreading of activation is assumed to be unconscious/implicit by many, e.g., Hunt and Lansman 1986, Cleeremans and McClelland 1991, Bower 1996). We use a four-layered connectionist network (see Figure 9), in which the first three layers form a (either recurrent or feedforward) backpropagation network for computing Q-values and the fourth layer (with only one node) performs stochastic decision making. The network is internally subsymbolic and uses distributed representation in accordance with our previous considerations. The output of the third layer (i.e., the output layer of the backpropagation network) indicates the Q-value of each action (represented by an individual node), and the node in the fourth layer determines probabilistically the action to be performed based on the Boltzmann distribution.

To acquire the Q-values, supervised and/or reinforcement learning methods may be applied. I will not describe supervised learning here (but see Rumelhart et al 1986). I will examine the *Q-learning* algorithm (Watkins 1989), a reinforcement learning algorithm. In the algorithm, $Q(x, a)$ estimates the maximum discounted cumulative reinforcement that the agent will receive from the current state x on:

$$\max(\sum_{i=0}^{\infty} \gamma^i r_i) \tag{2}$$

where γ is a discount factor that favors reinforcement received sooner relative to that received later, and r_i is the reinforcement received at step i (which may be none). The updating of $Q(x, a)$ is based on minimizing

$$r + \gamma e(y) - Q(x, a) \tag{3}$$

where γ is a discount factor and $e(y) = \max_a Q(y, a)$. Thus, the updating is based on the *temporal difference* in evaluating the current state and the action chosen: In the above formula, $Q(x, a)$ estimates, before action a is performed, the (discounted) cumulative reinforcement to be received if action a is performed, and $r + \gamma e(y)$ estimates the (discounted) cumulative reinforcement that the agent will receive, after action a is performed; so their difference (the temporal difference in evaluating an action) enables the learning of Q-values that approximate the (discounted) cumulative reinforcement. Using Q-learning allows sequential behavior to emerge in an agent. Through successive updates of the Q-values , the agent can learn to take into account future steps in longer and longer sequences.

Applying the Q-learning algorithm, the training of the backpropagation network is based on minimizing the following error at each step:

$$err_i = \begin{cases} r + \gamma e(y) - Q(x, a) & \text{if } a_i = a \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where i is the index for an output node representing the action a_i. Based on the above error measures, the backpropagation algorithm is applied to adjust internal weights (which are randomly initialized before training). (Or, when a correct input/output mapping is available for a step, backpropagation can be directly applied using the error measure of the desired output minus the actual output.) [12]

In the top level, explicit action-centered knowledge is captured in a simple propositional rule form. To facilitate correspondence with the bottom level (and to encourage uniformity and integration; Clark and Karmiloff-Smith 1993), we chose to use a localist connectionist model for implementing these rules (e.g., Sun 1992) in accordance with our previous considerations. Basically, we translate the structure of a set of rules into that of a network. Rules are in the following form: *conditions* $\longrightarrow$ *action*, where the left-hand side is a conjunction of individual conditions each of which refers to a primitive: a value (or a value range) in a dimension of the (sensory) input state x that match the current input (i.e., $x_i = v_i$ or $x_i \in [v_{i1}, v_{i2}]$), and the right-hand side is an action recommendation a. [13] (Alternatively, the rules can be in the forms of *current-state* $\longrightarrow$ *action new-state* or *current-state action* $\longrightarrow$ *new-state*.) Each condition is represented by an individual node. For each rule, a set of links are established, each of which connects a node representing a condition in the left-hand side of a rule to the node representing the conclusion in the right-hand side of the rule. If a condition is in a positive form, the link carries a positive weight w; otherwise, it carries a negative weight $-w$. Sigmoidal functions are used for node activation (which is an obvious choice; other functions are also possible):

$$\frac{1}{1 + e^{-\sum_i i_i w_i + \tau}} \tag{5}$$

The threshold τ of a node is set to be $n * w - \theta$, where n is the number of incoming links (the number of conditions leading to the conclusion represented by this node), and θ is a parameter, selected along with w to make sure that the node has activation above 0.9 when all of its conditions are satisfied, and has activation below 0.1 when some of its conditions are not met. Activations above 0.9 are considered

[12] This learning process performs both structural credit assignment (with backpropagation), so that the agent knows which element in a state should be assigned credit/blame, as well as temporal credit assignment (through temporal difference updating), so that the agent knows which action leads to success or failure.

[13] Each element is either ordinal (discrete or continuous) or nominal. In the following discussion, we focus on ordinal values; nominal values can be handled similarly. A binary element is a special case of discrete ordinal elements.

1, and activations below 0.1 are considered 0. So rules are in fact discretized and thus crisp/binary. [14]

Algorithms for learning explicit/conscious knowledge (rules) at the top level were devised. One algorithm that learns by using information in the bottom level is described as follows.

1. Update the statistics regarding the conditions of each rule.
2. Check the current criterion for rule construction, expansion, shrinking, and deletion.
2.1. If the result is successful according to the current criterion, and there is no rule matching that state and that action, then perform *construction* of a new rule: state $\longrightarrow$ action. Add the constructed rule to the rule network.
2.2. If the result is unsuccessful according to the current criterion, revise all the matching rules using *shrinking* and *deletion*.
2.2.1. Remove the matching rules from the rule network.
2.2.2. Add the revised versions of the rules into the rule network.
2.3. If the result is successful according to the current criterion, then generalize the matching rules.
2.3.1. Create new rules using *expansion*.
2.3.2. Add the expanded rules to the rule network to replace the original rules.

Let us discuss the details of the operations used in the above algorithm (including rule construction, shrinking, deletion, and expansion) and the criteria measuring whether a step is successful or not (used in deciding whether or not to apply some operators). At each step, we examine on the following information: (x, y, r, a), where x is the state before action a is performed, y is the new state entered after an action a is performed, and r is the reinforcement received after action a. We decide on whether or not to construct a rule based a simple criterion: We apply the *construction* operator if $r + \gamma e(y) - Q(x, a) > threshold$, which is fully determined by the current step (x, y, r, a).

The criterion for applying the *expansion* and *shrinking* operators, on the other hand, is based on a statistical test. The statistics are calculated in the following way: at each step, we update the following statistics for each rule condition and their variations, i.e., a rule condition plus/minus one value, with regard to the action a performed: PM (Positive Match) and NM (Negative Match), where positivity/negativity is determined by: whether or not we have $\max_b Q(y, b) - Q(x, a) + r > threshold$. Each statistic is updated with the following formula: $stat := stat + 1$ (where $stat$ stands for any statistic involved); at the end of each episode, it is discounted by: $stat := stat * 0.90$. [15] Based on these statistics, we employ the following measure in deciding whether or not to apply these operators:

$$IG(A, B) = log_2 \frac{PM(A) + 1}{PM(A) + NM(A) + 2} - log_2 \frac{PM(B) + 1}{PM(B) + NM(B) + 2}$$

where A and B are two different conditions that lead to the same action. (This is a widely used measure, for example, in many Inductive Logic Programming models, and well justified on the empirical basis. The measure compares essentially the percentage of positive matches under different conditions A and

[14] In addition, if there is more than one rule that leads to the same conclusion, an intermediate node is created for each such rule: all of the conditions of each rule are linked to the same intermediate node, and then all the intermediate nodes are linked to the node representing the conclusion. For more complex rule forms including predicate rules and variable binding, see Sun (1992). Such rules can be learned using more complex ILP techniques.

[15] The results are time-weighted statistics, which are useful in nonstationary situations.

B (with the Laplace estimator). If A can improve the percentage to a certain degree over B, then A is considered better than B. If a rule is better compared with the match-all rule (i.e, the rule with the condition that matches all inputs), then the rule is considered successful in the above algorithm (for deciding on expansion or shrinking operations). Here are the detailed descriptions of the rule revision operators:

- *Expansion*: if $IG(C, all) > threshold1$ and $IG(C', C) \geq 0$, where C is the current condition of the rule, C' is a modified condition (i.e., $C' = C$ plus one value), and *all* refers to no condition at all (for the same action specified by the rule), set $C'' = argmax_{C'} IG(C', C)$ as the new condition of the rule. Reset rule statistics. Any rule covered by the expanded rule will be placed in its children list. [16]
- *Shrinking*: if $IG(C, all) > threshold2$ and $IG(C', C) > 0$, where C is the current condition of the rule and C' is a modified condition (i.e., $C' = C$ minus one value range), set $C'' = argmax_{C'} IG(C', C)$ as the new condition of the rule. Reset rule statistics. Restore those rules in the children list that are not covered by the shrunk rule. If shrinking the condition makes it impossible for a rule to match, delete the rule.
- *Deletion*: The same as *Shrinking*.

Note that although the accumulation of statistics is gradual, the acquisition and revision of rules is one-shot and all-or-nothing. The process can be characterized as hypothesis testing as described by Bruner et al (1956) and Nosofsky et al (1993).

The above algorithm enables *bottom-up* learning. On the other hand, Stand-alone top-level learning can also be performed through hypothesis testing. Moreover, *top-down* learning can be performed by using the top-level knowledge to train the bottom-level network (in a supervised fashion). See Sun et al (1996, 1997) for details of these additional algorithms.

In the overall algorithm, Step 4 is for making the final decision on which action to take by incorporating outcomes from both levels. It allows different operational modes: e.g., relying only on the top level, relying only on the bottom level, or combining the outcomes from both levels weighing them differently. Specifically, we combine the corresponding values for an action from the two levels by a weighted sum; that is, if the top level indicates that action a has an activation value v (which should be 0 or 1 as rules are binary) and the bottom level indicates that a has an activation value q (the Q-value), then the final outcome is $w_1 * v + w_2 * q$. Stochastic decision making with Boltzmann distribution (based on the weighted sums) is then performed to select an action out of all the possible actions. The weights can be adjusted as described before. See Willingham et al (1989) and Jacoby et al (1993) for cognitive justifications of this method.

Figure 10 shows the two levels of the model. More explanations of the model can be found in Sun and Peterson (1998) and Sun et al (1996, 1999). We will not get into the details of the non-action-centered part of CLARION in this paper.

[16] The children list of a rule is created to keep aside and make inactive those rules that are more specific (thus fully covered) by the current rule. It is useful because if later on the rule is deleted or shrunk, some or all of these rules on its children list may be reactivated if they are no longer covered.

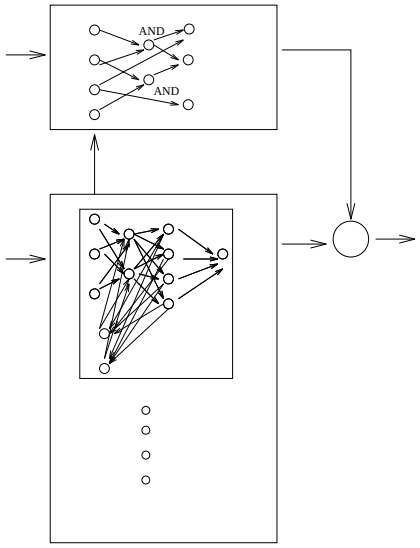

Figure 10: The implementation of CLARION. The top level contains localist encoding of propositional rules. The bottom level contains a connectionist network for capturing procedural skills. The interaction of the two levels and the information flows are indicated with arrows.

References

P. Ackerman, (1988). Determinants of individual differences during skill acquisition: cognitive abilities and information processing. *Journal of Experimental Psychology: General.* 1117 (3), 288-318.

J. R. Anderson, (1983). *The Architecture of Cognition*, Harvard University Press, Cambridge, MA

J. R. Anderson, (1993). *Rules of the Mind.* Lawrence Erlbaum Associates. Hillsdale, NJ.

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

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

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

L. Breiman, (1996). Bagging predictors. *Machine Learning*, 24, 123-140.

J. Bruner, J. Goodnow, and J. Austin, (1956). *A Study of Thinking.* Wiley, New York.

D. Chalmers, (1993). *Towards a Theory of Consciousness.* Ph.D Thesis, Indiana University.

P. Churchland, (1988). *Matters and Consciousness.* MIT Press. Cambridge, MA.

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

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

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

A. Collins and J. Loftus, (1975). Spreading Activation theory of semantic processing, *Psychological Review*, vol.82, pp.407-428.

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

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

A. Damasio et al, (1990). Neural regionalization of knowledge access. In: *Cold Spring Harbor Symp. on Quantitative Biology*, Vol.LV. CSHL Press.

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

D. Dennett, (1991), *Consciousness Explained.* Little Brown.

J. Dewey, (1958). *Experience and Nature.* Dover, New York.

Z. Dienes, (1992). Connectionist and memory-array models of artificial grammar learning. *Cognitive Science.* 16. 41-79.

Z. Dienes and D. Berry, (1997). *Psychonomic Bulletins and Reviews.*

H. Dreyfus and S. Dreyfus, (1987). *Mind Over Machine: The Power of Human In tuition*, The Free Press, New York, NY.

H. Dreyfus, (1992). *Being-in-the-world.* MIT Press, Cambridge, MA.

G. Edelman, (1989). *The Remembered Present: A Biological Theory of Consciousness.* Basic Book, New York.

J. Elman, (1990). Finding structures in time. *Cognitive Science.* 14, 179-211.

O. Etzioni, (1991). Embedding decision-analytic control in a learning architecture. *Artificial Intelligence*, 49, 129-159.

J. Feldman and D. Ballard, (1982). Connectionist models and their properties, *Cognitive Science*, pp.205-254, July 1982

O. Flanagan, (1992). *Consciousness Reconsidered.* MIT Press, Cambridge, MA.

W. Freeman, (1995). *Societies of Brains.* Lawrence Erlbaum Associates, Hillsdale, NJ.

J. Fodor, (1975). *The Language of Thought.* Crowell.

J. Fodor, (1983). *The Modularity of Mind.* MIT Press, Cambridge, MA.

J. Fodor and Z. Pylyshyn, (1988). Connectionism and Cognitive Architecture: A Critical Analysis, in: Pinker and Mehler (eds.) *Connections and Symbols*, MIT Press, Cambridge, MA.

J. Gelfand, D. Handelman and S. Lane, (1989). Integrating Knowledge-based Systems and Neural Networks for Robotic Skill Acquisition, *Proc.IJCAI*, pp.193-198. Morgan Kaufmann, San Mateo, CA.

J. Hasher and J. Zacks, (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General.* 108. 356-358.

N. Hayes and D. Broadbent, (1988). Two modes of learning for interactive tasks. *Cognition.* 28, 249-276.

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

M. Heidegger, (1927). *Being and Time.* English translation published by Harper and Row, New York. 1962.

J. Hendler, (1987). Marker Passing and Microfeature, *Proc.10th IJCAI*, pp.151-154, Morgan Kaufmann, San Mateo, CA.

G. Hinton, (1990). Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence*, 46, 47-76.

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

R. Jackendoff, (1987). *Consciousness and the Computational Mind.* MIT Press.

L. Jacoby, J. Toth, and A. Yonelinas, (1993). Separating conscious and unconscious influences of memory: measuring recollection. *Journal of Experimental Psychology: General*, 122 139-154.

W. James, (1890). *The Principles of Psychology.* Dover, New York.

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

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

F. Keil, (1989). *Concepts, Kinds, and Cognitive Development.* MIT Press. Cambridge, MA.

C. Kelley and L. Jacoby, (1993). The construction of subjective experience: memory attribution. In: *Consciousness.* eds. M. Davies and G. Humphreys. Blackwell, Oxford, UK.

J. Klahr, et al, (1989). *Production System Models of Learning and Development*, MIT Press, Cambridge, MA.

J. LeDoux, (1992). Brain mechanisms of emotion and emotional learning. In: *Current Opinion in Neurobiology.* Vol.2, No.2, 191-197.

P. Lewicki, M. Czyzewska, and H. Hoffman, (1987). Unconscious acquisition of complex procedural knowledge. *Journal of Experimental Psychology: Learning, Memory and Cognition.* 13 (4), 523-530.

P. Lewicki, T. Hill, and M. Czyzewska, (1992). Nonconscious acquisition of information. *American Psychologist.* 47, 796-801.

B. Libet, (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences.* 8, 529-566.

D. Lloyd, (1995). Consciousness: a connectionist manifesto. *Minds and Machines.* 5: 161-185.

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

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

A. Marcel, (1988). Phenomenal experience and functionalism. in: A. Marcel and E. Bisiach, *Consciousness in Contemporary Science.* Oxford University Press. Oxford, UK.

D. Mathis and M. Mozer, (1996). Conscious and unconscious perception: a computational theory. *Proc. of 18th Annual Conference of cognitive Science Society*, 324-328.

J. McClelland, B. McNaughton and R. O'Reilly, (1994). Why there are complementary learning systems in the hippocampus s and neocortex: insights from the successes and failures of connectionist models of learning and memory. Technical Report PDP.CNS.94.1, Carnegie Mellon University.

D. Meyer and D. Kieras, (1997). A computational theory of executive cognitive processes and human multiple-task performance: part 1. *Psychological Review.*

P. Merikle, (1992). Perception without awareness: critical issues. *American Psychologists.* 47, 792-795.

P. Merikle and E. Reingold, (1991). Comparing direct (explicit) and indirect (implicit) measures to study unconscious memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 17, 224-233.

J. Metcalfe and A. Shimamura, (eds.) (1994). *Metacognition : Knowing About Knowing.* MIT Press, Cambridge, MA.

D. Milner and N. Goodale, (1995). *The Visual Brain in Action.* Oxford University Press, New York.

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

R. Nisbett and T. Wilson, (1977). Telling more than we can know: verbal reports on mental processes. *Psychological Review.* 84 (3), 1977.

D. Navon and D. Gopher, (1979). On the economy of the human processing system. *Psychological Review.* 86, 214-255.

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

R. Nosofsky, T. Palmeri, and S. McKinley, (1994). Rule-plus-exception model of classification learning. *Psychological Review.* 101 (1), 53-79.

R. Penrose, (1994). *Shadows of the Mind.* Oxford University Press. Oxford, UK.

D. Perlis, (1985). Language with self reference I *Artificial Intelligence.* 25, 301-322.

M. Posner, G. DiGirolamo, and D. Fernandez-Duque, (1997). Brain mechanisms of cognitive skills. *Consciousness and Cognition.* 6, 267-290.

M. Posner and S. Petersen, (1990). The attention system of the human brain. *Annual Review of Neuroscience.* 13, 25-42.

M. Posner and C. Snyder, (1975). Facilitation and inhibition. In: *Attention and Performance.* eds. P. Rabbitt and S. Dornick. Academic Press.

M. R. Quillian, (1968). Semantic memory. In: M. Minsky (ed.), *Semantic Information Processing.* MIT Press, Cambridge, MA. pp.227-270.

H. Rachlin, (1994). Self control: beyond commitment. *Brain and Behavioral Sciences.*

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

A. Revonsuo, (1993). Cognitive models of consciousness. In M. Kamppinen (ed.), *Consciousness, Cognitive Schemata and Relativism.* 27-130. Kluwer, Dordrecht, Netheland.

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

P. Rosenbloom, J. Laird, and A. Newell, (1993). *The SOAR papers: Research on Integrated Intelligence.* MIT Press, Cambridge, MA.

D. Rosenthal, (1991). (ed.) *The Nature of Mind.* Oxford University Press, Oxford, UK.

D. Rumelhart, J. McClelland and the PDP Research Group, (1986). *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, MIT Press, Cambridge, MA.

S. Russell and E. Wefald, (1989). The principle of meta-reasoning. *Proc. of 1st International Conference on Knowledge Representation and Reasoning*, 406-411. Morgan Kaufmann.

D. Schacter, (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 501-518.

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

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

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

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

T. Shallice, (1988). *From Neuropsychology to Mental Structure.* Cambridge University Press.

R. Shiffrin and W. Schneider, (1977). Controlled and automatic human information processing II. *Psychological Review.* 84. 127-190.

S. Sloman, (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119 (1), 3-22.

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

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

L. Squire, B. Knowlton, and G. Musen, (1993). The structure and organization of memory. *Annual Review of Psychology.* 44, 453-495.

W. Stanley, R. Mathews, R.Buss, and S. Kotler-Cope, (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.

R. Sun, (1992). On Variable Binding in Connectionist Networks, *Connection Science*, Vol.4, No.2, pp.93-124.

R. Sun, (1994). *Integrating Rules and Connectionism for Robust Commonsense Reasoning.* John Wiley and Sons, New York, NY.

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

R. Sun, (1997). Learning, action, and consciousness: a hybrid approach towards modeling consciousness. *Neural Networks*, special issue on consciousness. 10 (7), pp.1317-1331.

R. Sun, E. Merrill, and T. Peterson, (1998). A bottom-up model of skill learning. *Proc.of 20th Cognitive Science Society Conference*, pp.1037-1042, Lawrence Erlbaum Associates, Mahwah, NJ. 1998.

R. Sun, E. Merrill, and T. Peterson, (1999). A model for bottom-up skill learning. submitted for publication.

R. Sun and T. Peterson, (1997). A hybrid agent architecture for sequential decision making. in: R. Sun and F. Alexandre (eds.), *Connectionist Symbolic Integration.* pp.113-138. Lawrence Erlbaum Associates, Hillsdale, NJ.

R. Sun and T. Peterson, (1998). Autonomous learning of sequential tasks: experiments and analyses. *IEEE Transactions on Neural Networks*, Vol.9, No.6, pp.1217-1234.

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

R. Sun and C. Terry, (1999). Modeling implicit learning using a bottom-up model. submitted for publication.

R. Sutton, (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. *Proc.of Seventh International Conference on Machine Learning.* Morgan Kaufmann. San Mateo, CA.

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

J. Taylor, (1997). The relational mind. in: A. Browne, (ed.) *Neural Network Perspectives on Cognition and Adaptive Robotics.* IOP, Bristol, UK.

W. Timberlake and G. Lucas, (1993). Behavior systems and learning: from misbehavior to general principles. in: ???

E. Tulving, (1972). Episodic and semantic memory. In: E. Tulving and W. Donaldson (eds.), *Organization of Memory.* 381-403. Academic Press, New York.

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

N. Ueda and R. Nakano, (1996). Generalization error of ensemble estimators. *IEEE International Conference on Neural Networks*, pp.90-95. IEEE Press.

D. Waltz, (1991). How to build a robot. *Proc.of Conf. on Simulation of Adaptive Behaviors.* S. Wilson. (ed.) MIT Press. Cambridge, MA.

C. Watkins, (1989). *Learning with Delayed Rewards.* Ph.D Thesis, Cambridge University, Cambridge, UK.

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