Pergamon

Neural Networks, Vol. 10, No. 7, pp. 1317–1331, 1997

Printed in Great Britain
0893–6080/97 $17.00+.00

PII: S0893-6080(97)00050-6

1997 SPECIAL ISSUE

Learning, Action and Consciousness: A Hybrid Approach Toward Modelling Consciousness

RON SUN

Department of Computer Science, The University of Alabama

(*Received* 2 *July* 1996; *accepted* 4 *December* 1996)

Abstract—*This paper is an attempt at understanding the issue of consciousness through investigating its functional role, especially in learning, and through devising hybrid neural network models that (in a qualitative manner) approximate characteristics of human consciousness. In doing so, the paper examines explicit and implicit learning in a variety of psychological experiments and delineates the conscious/unconscious distinction in terms of the two types of learning and their respective products. The distinctions are captured in a two-level action-based model* CLARION. *Some fundamental theoretical issues are also clarified with the help of the model. Comparisons with existing models of consciousness are made to accentuate the present approach. © 1997 Elsevier Science Ltd.*

Keywords—Neural networks, Hybrid systems, Consciousness, Implicit learning, Reinforcement learning, Procedural knowledge, Rule extraction, Dual representation.

1. INTRODUCTION

Amidst the widespread enthusiasm of recent years concerning the scientific study of consciousness, there are a large number of models being proposed (including computational models, which in turn include neural network models), and various claims have been made about them. These models capture to various extents experimental findings and pretheoretical intuitions about consciousness (see, e.g. Taylor, 1994; Schacter, 1990; Jackendoff, 1987; Shallice, 1988; Baars, 1988; Dennett & Kinsbourne, 1992; Penrose, 1994). Unfortunately, however, some of these models (such as Dennett & Kinsbourne, 1992; Shallice, 1988; Jackendoff, 1987) are aimed at a very high and gross level of explanation (e.g. overall architectures) and thus unable to provide more detailed predictions and explanations. On the other hand, existing computational, especially neural network, models tend to rush directly into complex neural physiological thickets (Taylor, 1994; Edelman, 1989) and thus may lose sight of forests. In addition, most existing models do not deal adequately with one crucial aspect of human consciousness: learning. In contrast to these approaches, we intend to stay at an intermediate and functional level; investigating the detailed functional roles of consciousness and determining how various aspects of the conscious and the unconscious should figure into the architecture of the mind (in terms of learning as well as performance). In other words, we posit a middle level between phenomenology and physiology/neurobiology, which might be more apt at capturing fundamental characteristics of consciousness. We will also link computational models of consciousness to parts of (phenomenological) philosophy that are concerned with consciousness.

As we will focus mainly on the learning aspect in consciousness, let us briefly describe the learning settings that we examined (from Sun et al., 1995). These settings are more complex than simple categorisation/classification and though action-based, involve more than just simple stimulus–response pairing. Psychological experiments involving dynamic decision making or artificial grammars will be discussed. In dynamic decision making (Berry & Broadbent, 1988), subjects were required to control the levels of an output variable by manipulating levels of an input variable. In one instance, subjects were to manage a simulated sugar production factory and the goal was to reach and maintain a particular level of productivity by manipulating the size of the workforce. In another instance, subjects were to interact with a computer simulated ‘‘person’’ and to maintain the behaviour of the person at ‘‘very friendly’’ by manipulating his/her own behaviour. In artificial grammar

Acknowledgements: This work was supported in part by Office of Naval Research grant N00014-95-1-0440. Thanks to Ed Merril and Diana Gordon for their comments.

Requests for reprints should be sent to Ron Sun, Department of Computer Science, The University of Alabama, Tuscaloosa, AL 35487, USA; Tel.: (205) 348-1667; e-mail: rsun@cs.ua.edu.

learning (Reber, 1989), subjects were presented with a string of letters that were generated in accordance with a simple grammar. Although subjects were unaware of the underlying grammars, they were asked to judge the grammaticality of novel strings. In addition, two navigation tasks were used for both psychological experiments and computational modelling in our lab (see Sun et al., 1995, 1996a, b). One is maze running, in which from a starting location, a subject has to find a goal location, using only local sensory information. The other is navigation through minefields, in which a subject is required to go from a starting location to a goal location within a given short period of time by navigating through a densely packed field of mines (which will explode if the subject gets too close). I will later show the relevance of these learning tasks to the study of consciousness.

In the rest of this paper, I will first present a cognitive architecture CLARION[1] for accounting for the distinction of the conscious and the unconscious (in Section 2). I will then show how it accounts for a number of phenomena related to the conscious/unconscious distinction in learning (which may also be referred to as the explicit and implicit distinction) (in Section 3). A discussion of some fundamental theoretical issues will take place after that (Section 4). A comparison to existing models of consciousness such as Baars (1988), Schacter (1990), and Damasio (1994) will follow, which will show the commonalities shared by some of these models and CLARION and the unique features of the present model (in Section 5). Some concluding remarks (Section 6) will complete the paper.

2. A HYBRID NEURAL NETWORK MODEL

A computational model that can tackle the learning tasks mentioned above is needed. It needs to satisfy some basic requirements as follows. It must be able to learn from scratch on its own (as human subjects often do in the learning tasks outlined earlier; Berry & Broadbent, 1988; Reber, 1989; and also Sun et al., 1996a). The model also has to perform concurrent, on-line learning. That is, it has to learn continuously from on-going experience in the world; for, as indicated by Medin et al. (1987), Nosofsky et al. (1994) and others, human learning is often gradual, on-going and concurrent, which is true of all the aforementioned tasks. As suggested by Anderson (1983) and many others, there are clearly two types of knowledge involved in human learning — procedural and declarative: while one is generic and easily accessible, the other is embodied and specific. Moreover, different types of learning processes are involved in acquiring different types of knowledge (Anderson, 1983; Keil, 1989; Smolensky, 1988; Stanley et al., 1989). Humans are able to learn procedural knowledge through trial and error (without a priori knowledge) in the aforementioned tasks. On top of low-level procedural skills, declarative knowledge can be acquired also through on-going experience in the world (see Stanley et al., 1989). Furthermore, it is important for declarative knowledge to be learned through the meditation of low level skills (i.e. bottom-up learning; see Sun et al., 1996a).

Procedural knowledge (skills) can be captured by subsymbolic distributed representation, such as that provided by a backpropagation network. Because of the implicit nature of procedural skills, details of such skills are in general inaccessible to consciousness (Anderson, 1983; Reber, 1989). A distributed representation naturally captures this property with representational units that are capable of accomplishing tasks but are in general uninterpretable and subsymbolic (Sun, 1994, 1995). (A symbolic representation may be used, but then this would require an artificial assumption that these representations are not accessible, while other similar representations are accessible — such a distinction is arbitrary.)

Procedural knowledge can be learned in a couple of different ways. In the case where correct input/output mappings are provided, straight backpropagation can be used on a neural network. Otherwise, reinforcement learning can be used (Sutton, 1990; Watkins, 1989). This is preferred because there is often no uniquely correct action in the aforementioned tasks, although feedback is usually available. Using reinforcement learning in neural networks, we can measure the goodness of an action through a payoff-reinforcement signal. An adjustment can be made to 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.

This level can be modular; that is, a number of small networks can co-exist each of which is adapted to specific modalities, tasks, or groups of input stimuli. This coincides with the well known modularity claim (Fodor, 1983; Karmiloff-Smith, 1986; Cosmides & Tooby, 1994), in that much processing in the human mind is done by limited, encapsulated (to some extent), specialized processors that are highly efficient. It is also similar to the idea of Shallice (1988) that a multitude of ''action systems'' compete with each other. There also has been some work in neural network and machine learning communities in developing modular systems, which are equally relevant.

On the other hand, declarative knowledge can be captured by a symbolic or a ''localist'' representation (Clark & Karmiloff-Smith, 1993), in which each unit has a clear conceptual meaning or interpretation. This allows declarative knowledge to be highly accessible and inferences to be performed explicitly (Smolensky, 1988; Sun, 1994, 1995).

[1] It was originally developed for modelling human skill learning; see Sun et al. (1995).

Declarative knowledge can be learned in a variety of ways. In this work, because of the dynamic on-going nature of the learning tasks, we need to be able to dynamically acquire a representation and to modify the representation subsequently if necessary, in an efficient or even one-shot fashion.

The difference in representing procedural and declarative knowledge revealed by the above discussion leads naturally to a two-level architecture, in which one level is procedural and the other declarative. This structuring can be argued on both psychological and philosophical grounds. Anderson (1983) put forward the dichotomy of separate yet interconnected declarative and procedural knowledge bases to account for a variety of learning data. Smolensky (1988) suggested that the separation of conceptual-level and subconceptual-level processing. The conceptual level possesses three characteristics: (1) public access; (2) reliability; and (3) formality. It can thus be modelled by symbolic processing. In contrast, skills, intuition, and the like are not expressible in linguistic forms and do not conform to the three criteria prescribed. Hence, skills and intuition constitute a different type of capacity, reflecting the ''subsymbolic'' processing at the subconceptual level (see also Shiffrin & Schneider, 1977). In a similar vein, Dreyfus and Dreyfus (1987) contrasted analytical and intuitive thinking, from a phenomenological analysis of human cognitive skill learning in which the fluent, holistic and situation sensitive way of solving problems (intuition) as observed in master level performers is in sharp contrast with the slow, deliberate thinking that often occurs in the novices (analytical thinking). Models have been proposed to account for such two-tiered structures, which often posit the existence of at least two separate components, each of which responds to one side of a dichotomy (e.g. Posner & Snyder, 1975; Schacter, 1990; Murphy & Medin, 1985; Keil, 1989; Sun, 1992a, 1994; Sun & Bookman, 1994). The dual representation hypothesis put forth in Sun (1994) stated that:

> It is assumed in this work that cognitive processes are carried out in two distinct levels with qualitatively different processing mechanisms. Each level encodes a fairly complete set of knowledge for its processing, and the coverage of the two sets of knowledge encoded by the two levels overlaps substantially.

Based on the above considerations, we developed CLARION: *Connectionist Learning with Adaptive Rule Induction ON-line*. It consists of two main components: the top level encodes explicit declarative knowledge, and the bottom level encodes implicit procedural knowledge. In addition, there is an episodic memory, which stores recent experiences in the form of ''input, output, result'' (i.e. stimulus, response, and consequence) that are recently-filtered (episodic memory will not be used in this paper and therefore will not be further discussed here) (see Figure 1).

An overall pseudo-code algorithm that describes the operation of CLARION is as follows:

1. Observe the current state x (in a proper representation).
2. Compute in the bottom level the Q-values of x associated with each of the possible actions a_i's: $Q(x, a_2), ..., Q(x, a_n)$. Select one action or a few based on Q-values.
3. Find out all the possible actions $(b_1, b_2, ..., b_m)$ at the top level, based on the input x (sent up from the bottom level) and the rules in place.
4. Compare the values of the selected a_i's with those of the b_j's (sent down from the top level), and choose an appropriate action b.

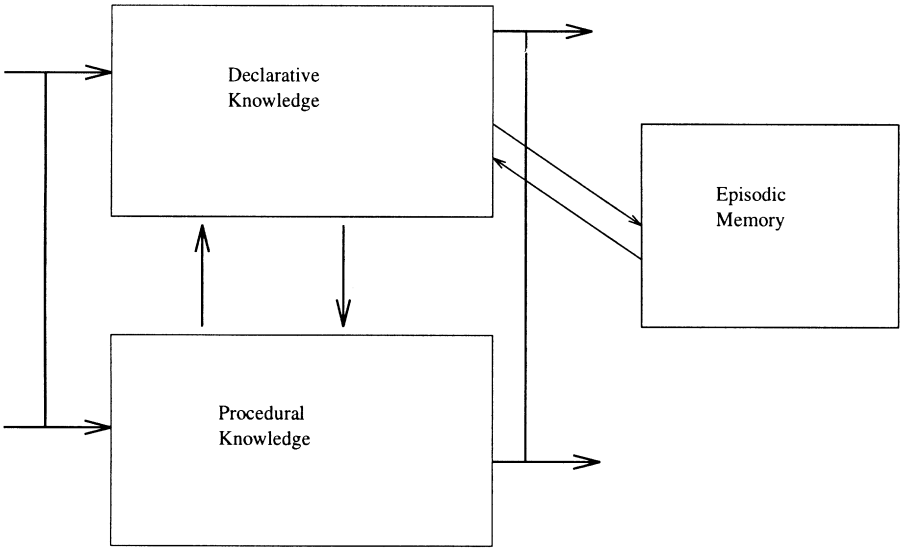

FIGURE 1. The CLARION architecture.

5. Perform the action b, and observe the next state y and (possibly) the reinforcement r.
6. Update Q-values at the bottom level in accordance with the *Q-Learning-Backpropagation* algorithm.
7. Update the rule network at the top level using the *Rule-Extraction-Refinement* algorithm.
8. Go back to step 1.

In the bottom level, a Q-value is an evaluation of the ''quality'' of an action in a given state: $Q(x, a)$ indicates how desirable action a is in state x (which consists of some sensory input). We can choose an action based on Q-values, e.g. by choosing the one that has the maximum Q-value in the current state or by choosing an action probabilistically based on Q-values. To acquire the Q-values, one option is to use the *Q-learning* algorithm (Watkins, 1989), a reinforcement learning algorithm.[2] In the algorithm, $Q(x, a)$ estimates the maximum discounted cumulative reinforcement that the agent will receive from the current state x on:

$$\max\left(\sum_{i=0}^{\infty}\gamma^i r_i\right) \tag{1}$$

where γ is a discount factor that favours reinforcement received sooner relative to that received later, and r_i is the reinforcement received at step i (which may be 0). The updating of $Q(x, a)$ is based on minimising

$$r + \gamma e(y) - Q(x, a) \tag{2}$$

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, after action a is performed, the (discounted) cumulative reinforcement that the agent will receive; 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 behaviour to emerge. Through successive updates of the Q-function, the agent can learn to take into account future steps in longer and longer sequences.[3]

We chose to use a four-layered network for implementation (see Figure 2), 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 implicit in 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 determined probabilistically the action to be performed based on the Boltzmann Distribution (i.e. Luce's choice axiom; Watkins, 1989):

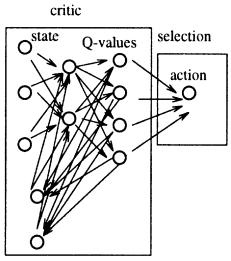

FIGURE 2. The *Q*-Learning method.

$$p(a|x) = \frac{e^{1/\alpha Q(x, a)}}{\sum_i e^{1/\alpha Q(x, a_i)}} \tag{3}$$

Here, α controls the degree of randomness (temperature) of the decision making process.[4] The combination of Q-learning and backpropagation facilitates the development of procedural skills in the bottom level, which can potentially be done solely on the basis of acting and exploring in the real world. This learning process performs both structural credit assignment and temporal credit assignment.

In the top level, declarative knowledge is captured in a simple propositional rule form. To facilitate correspondence with the bottom level and to encourage uniformity and integration (Clark & Karmiloff-Smith, 1993), we chose to use a localist network model for representing these rules. Basically, we connect the nodes representing conditions of a rule to the node representing the conclusion. However, we need to determine how we wire up a rule involving conjunctive conditions. There are a number of previous attempts (e.g. Sun, 1992b; Towel & Shavlik, 1993) that we can draw upon. For each rule, a set of links can be established, each of which connects to a concept in the condition of a rule to the conclusion of the rule. So the number of incoming links to the conclusion of a rule is equal to the number of conditions of the rule. If the concept in the condition is

[2] Supervised learning methods can also be applied, when correct mappings of an input and output are available.

[3] In terms of both simplicity and performance, Q-learning is best among similar reinforcement learning methods (Lin, 1992; Sun et al., 1995).

[4] The training of the backpropagation network is based on minimising the following:

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

where i is the index for an output node representing the action a_i. Backpropagation is then applied as usual to adjust the weights. Or, when correct mappings are available for each step, backpropagation can be directly applied.

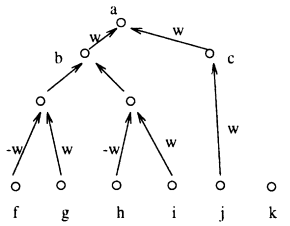

FIGURE 3. A network for representing rules: (1) $b\ c \rightarrow a$; (2) $\neg f\ g \rightarrow b$; (3) $\neg h\ i \rightarrow b$; (4) $j \rightarrow c$.

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 (as an obvious choice; other functions are also possible):

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

The threshold τ of a node is set to be n times $w - \theta$, 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 1, and activations below 0.1 are considered 0; so rules are crisp/binary.) 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 a rule are linked to the same intermediate node, and then all the intermediate nodes are linked to the node representing the conclusion (see Figure 3). (For more complex rule forms including predicate rules and variable binding, see Sun, 1992b).

To fully capture bottom-up learning processes, we devised an algorithm for learning declarative knowledge (rules) using information in the bottom level. The basic idea is as follows: if an action decided by the bottom level is successful (here, being successful could mean a number of different things, including the difference between the Q-value of the state before an action is performed and that after the action is performed, which comes from the bottom level; the details are specified in Sun et al., 1995), then the agent extracts a rule that corresponds to the action selected by the bottom level and adds the rule to the network. Then, in subsequent interactions with the world, the agent verifies the extracted rule by considering the outcome of applying the rule: if the outcome is not successful, then the rule should be made more specific and exclusive of the current case; if the outcome is successful, the agent may try to generalise the rule to make it more universal (Mitchell, 1982). (The detail of the algorithm can be found in Sun et al., 1995.)

At the top level, after rules have been learned, backward and forward chaining reasoning, means-ends analysis, counterfactual reasoning and explicit hypothesis testing can be used. These rules, expressed in the ''state → action result'' form (which constitutes a *schema*; cf. Waltz, 1991). allows powerful operations to be performed. Backward chaining means-ends analysis is accomplished at the top level through backward tracing of rule links from the ''result'', which is the new state entered after the ''action'' is performed, to the state, which is the state before the action is performed. This process is successively applied, with the derived ''state'' as the new state (the result) to be back traced, until reaching an initial state. All of the actions involved in the derivation are collected, which form a plan for accomplishing the desired final result. Counterfactual reasoning can also be applied (because we have information concerning conditions, actions, and results readily available): one can thus hypothetically alter either the conditions or the action of a rule, and see the change in (immediate or final) results. Such counterfactual reasoning can be used to justify (or explain) a chosen action, or a sequence of such actions (a plan). Explicit hypothesis testing, similar to techniques discussed by e.g. Bruner et al. (1956), Nosofsky et al. (1994), and Michalski (1983), can also be applied at this level of CLARION.

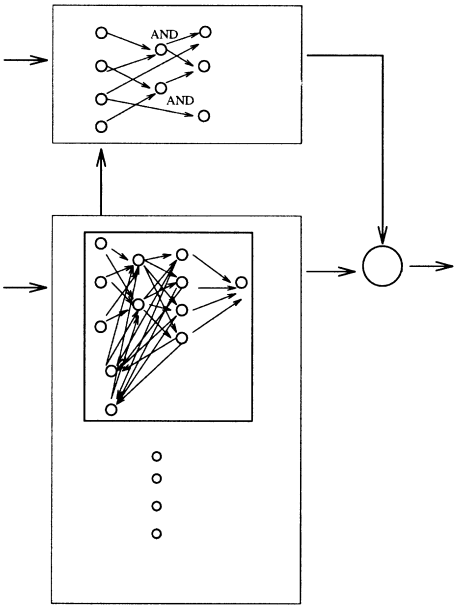

FIGURE 4. The implementation of CLARION.

The algorithm (at step 4) makes the final decision on which action to take by incorporating influences from both levels (as has been shown by Willingham et al., 1989, in humans, declarative knowledge can influence procedural performance). It allows different operational modes: (1) relying only on the top level; (2) relying only on the bottom level; or (3) combining the outcomes from both levels weighing them differently. The weights can change over time and in different situations. (These different operational modes roughly correspond to the

folk psychological notions of the intuitive mode, the deliberative mode, and the various mixtures of the two with different percentages of each; more later; see Sun et al., 1995.) Figure 4 shows the details of the two levels of the model.

The necessity of a two level architecture that incorporates two types of processes can be summed up as follows:

- In terms of representation, without the bottom level, the model will not be able to represent procedural skills properly. Such skills may involve graded, uncertain and inconsistent knowledge and autonomous stochastic exploration (with numeric calculation and probabilistic firing).
- In terms of learning, without learning in the bottom level, the model will not be able to learn from experience dynamically to acquire procedural skill. The bottom level captures *gradual* learning of skills, which is different from one-shot rule learning at the top level.
- Without the top level, the model will not be able to (1) represent generic, easily accessible, and crisp knowledge and (2) explicitly access and communicate that knowledge. When precision, consistency, and certainty are needed, declarative knowledge is preferred.
- Without rule learning, the model will not be able to acquire quickly and dynamically explicit knowledge for the top level from experience, and therefore have to resort to externally given declarative knowledge or to procedural knowledge exclusively.

There is ample biological evidence that indicates the existence of multiple pathways (in visual, linguistic, and other processing modes) some of which lead to conscious awareness, while others do not (e.g. one type is cortical while the other is subcortical), as described in Damasio (1994) and LeDoux (1992). For example, LeDoux (1992) described a cortical pathway from stimulus to thalamus to cortex, which produces conscious thoughts, and a subcortical pathway from stimulus to thalamus then to amygdala, which can lead directly to brain stem and effect actions without any explicit process. A two-level model such as CLARION approximates the separation of the two kinds of pathways to a certain degree, and suggests, in a concrete and tangible way, how subcortical processes can play a fundamental role in supporting and initiating consciousness in a bottom-up direction (more later).

3. PSYCHOLOGICAL DATA AND THE MODEL.

The crucial link between this model of procedural/declarative knowledge and the conscious/unconscious distinction in humans is in the psychological work on implicit learning (by e.g. Reber, 1989; Lewicki et al., 1992; Berry & Broadbent, 1988; Stanley et al., 1989; Willingham et al., 1989). Such work shows the dissociation between conscious and unconscious learning. Human knowledge, and its acquisition process, could be partially or completely unconscious. The connection from such illustrative data to our model lies in the ability of the model to account for some of the most important characteristics of human implicit/explicit learning, as will be sketched below. (More detailed comparisons and data can be found in Sun et al., 1996a, b)

3.1. The Difference of Conscious/Unconscious Learning

In the psychological data, there is a clear demonstration of the difference between conscious and unconscious (or, explicit and implicit) learning. Berry & Broadbent (1988) demonstrated this through an experiment using two similar dynamic decision tasks differing in the degree of saliency of the input/output relation. Human subjects were required to maintain the behaviour of a computer person at a ''very friendly'' level through their inputs. In the salient version, the computer responded in accordance with the subjects immediately preceding input. In the non-salient version, the computer responded in accordance with the input prior to that. Results suggested that subjects in the two conditions learned the tasks in very different ways: subjects in the non-salient condition learned the task implicitly while subjects in the salient condition learned the task explicitly, as demonstrated by tests of their explicit knowledge. Reber (1989) described a similar situation in artificial grammar learning. When complex hierarchical relations were needed to judge grammaticality, subjects tended to use implicit, unconscious learning; for example, when a sequence consisted of pairings of adjacent symbols that were ambiguous pair-wise but unambiguous when the entire sequence was viewed through hierarchical relations, such as in the case of 101110, implicit learning was preferred by the subjects. When only pair-wise relations were needed, such as in the case of 101010, subjects were more likely to use explicit, conscious learning by inducing an explicit rule. In other tasks, Cohen et al. (1990) also expressly demonstrated a dissociation between learning simple (pairwise) relations and learning complex hierarchical relations. A pattern emerging from the human data is that, if the to-be-learned relationships are simple, usually explicit/conscious learning prevails, while, when more complex relationships are involved, implicit/unconscious learning becomes more prominent. The implicit learning mechanism appears to be more structurally sophisticated and able to handle more difficult situations (Lewicki et al., 1992). It is important to note the inability of human subjects to articulate their implicitly learned knowledge, no matter how hard they tried (this is especially true in Lewicki et al., 1992). The subjects were often not even aware that they were learning. Nevertheless their performance improved over time, which demonstrated that their knowledge was unconscious.

This accords well with the CLARION model. In the model, one can freely move from one type of process to another, by engaging or disengaging the top level and its associated learning mechanisms (explicit processes that are consciously accessible, as discussed before), or the bottom level and its associated mechanisms (implicit processes that are not consciously accessible). Furthermore, in general, the bottom level is used to tackle more complex relations while the top level takes on simpler and crisper relations (cf. Reber, 1989; Seger, 1994). This is because the top level does not lend itself easily to the learning of complex structures due to its crisp, individuated, and symbolic representation and rigorous learning process. The bottom level, with its distributed network representation that incorporates gradedness and temporal information, handles complex relations better.

A specific example of this complexity difference is as follows. Implicit learning of sequences (e.g. artificial grammar sequences) is biased towards sequences with a high level of statistical structure with much correlation (Stadler, 1992). As has been demonstrated by Elman (1990) and by Cleeremans and McClelland (1991), recurrent backpropagation networks, as used in the bottom level of CLARION (in conjunction with *Q*-learning), can handle sequences with complex statistical structures, given proper training procedures. Dienes (1992) reported similar results, in which a simple network model outperformed other models in capturing sequence learning data. The rule learning mechanism, as used in the top level of CLARION, clearly has trouble handling such sequences. Therefore, in the circumstances in which a high level of statistical structure is involved in sequences, the bottom level prevails.

Note that there has been other work that demonstrated the distinction and dissociation of the two types of knowledge and proposed models based on that (e.g. Schacter, 1990; Shallice, 1988). However, some of the empirical work on which these models are based is concerned with abnormal subjects, most typically patients with brain damages. For example, Schacter (1990) discussed the following types of patients: amnesia (a selective inability to remember recent experience and to learn new information, typically due to lesions in the medial temporal lobe), blindsight (the inability to make certain responses in the absence of conscious perceptual awareness due to damages in the visual cortex), aphasia (impairment in processing syntactic or semantic information due to damages to a particular brain region), hemineglect (an impaired ability to attend to the side contralateral to the damaged hemisphere), and so on, all of which were characterised by dissociation of different types of information/knowledge somewhat similar to situations discussed above. Schacter (1990) proposed a model for accounting for the dissociation (see Section 5 for details). Although the model was, I believe, on the right track, the support for it was not as strong as it could have been, because in brain damaged patients, it was possible that certain reconfiguration and reallocation might have taken place (Shallice, 1988) and thus rendered the findings less applicable to normal human subjects. In this work, I only examine experimental findings from normal human subjects and thus results obtained may be generalised to a wider range of settings.

3.2. Delayed Explication of Unconscious Processes

In the implicit learning literature, implicit performance typically improves earlier than explicit knowledge that can be verbalised by the subject (Stanley et al., 1989). For example, in dynamic decision tasks, although performance quickly rises to a high level, subjects' verbal knowledge improves far slower; the subjects cannot provide usable verbal knowledge until near the end of their training (Stanley et al., 1989). Bowers et al. (1990) also showed delayed explication of implicit processes. When subjects were given patterns to complete, they showed implicit recognition of what a proper completion might be even though they did not have explicit recognition of a correct completion. The implicit recognition improved over time and eventually, an explicit recognition was achieved. In all of these cases, as suggested by Stanley et al. (1989) and Seger (1994), we may hypothesise that, due to the fact that explicit knowledge lags behind but improves along with implicit knowledge, explicit knowledge is in a way extracted from implicit knowledge. Cleeremans and McClelland (1991) also pointed out this possibility in discussing their data and models.

Several developmental theorists have considered a similar process in child development. Karmiloff-Smith (1986) suggested that developmental changes involve representational redescription. In young children, first low level implicit representations of stimuli were formed and used, then, when more knowledge was accumulated and stable behaviours developed, through a redescription process, more abstract representations were formed that transformed low-level representations and made them more explicit and usable. Based on data on perceptual analysis and categorization in infancy, Mandler (1992) proposed that relatively abstract ''image-schemas'' were extracted from perceptual stimuli, which coded several basic types of movements. On top of such image schemas, concepts were formed using information therein. She suggested that it was likely that an infant gradually formed ''theories'' of how his/her sensorimotor procedures work and thereby gradually made such processes explicit and accessible. Finally, Keil (1989) suggested that conceptual representations were composed of an associative component (with frequency and correlational information; Hasher & Zacks, 1979) and a ''theory'' component (with explicit knowledge; Murphy & Medin, 1985). Developmentally, there was a clear shift from associative to theory based representations in children. In data concerning learning

concepts of both natural and nominal kinds, simple similarity-based or prototype representations dominated at first, but gradually more explicit and focused theories developed and became more prominent. Keil (1989) pointed out that it was unlikely that theories developed independently, but rather they developed somehow from associative information that was already available. These findings further testify to the ubiquity of an implicit-to-explicit transition (Clark & Karmiloff-Smith, 1993).

CLARION readily captures this kind of bottom-up process. The bottom level develops implicit, embodied skills on its own (Section 2, eqn (2)), while the top level extracts explicit rules using algorithm *Rule-Extraction-Refinement* (Section 2). Thus, the delayed bottom-up learning naturally falls out of the model.[5]

3.3. Differences in Conscious/Unconscious Processes: Flexibility, Generalizability, and Robustness.

It has been shown that implicit learning produces less flexible knowledge than explicit knowledge (Seger, 1994; Berry & Broadbent, 1988; Stanley et al., 1989; Karmiloff-Smith, 1986). Seger (1994) argued that implicit learning results in knowledge that was more tied to the specific stimulus modality of the learning environment and less manipulable. Based on psycholinguistic data, Karmiloff-Smith (1986) observed that with the growth of explicit representations, more and more flexibility was shown by subject children. CLARION can account for the higher degree of flexibility of explicit, conscious knowledge relative to implicit, unconscious knowledge. Due to the explicit (i.e. localist) representation used at the top level of CLARION (which stores explicit knowledge), a variety of explicit manipulations can be performed that are not available to the bottom level. For example, backward and forward chaining reasoning, counterfactual reasoning, explicit hypothesis testing learning, and so on can be used individually or in combination. These capacities lead to heightened flexibility in the top level. The bottom level employs only backpropagation networks and thus cannot have the same flexibility.

As observed in many experiments, following explicit learning, subjects are able to handle novel stimuli in a similar way (or in other words, to generalise). In artificial grammar learning, Reber (1967, 1976) found good transfer to strings using different letters but based on the same grammar. Berry and Broadbent (1988) showed that subjects trained on a dynamic decision task could transfer to another task with a similar cover story and identical underlying relations. Generalisation has been demonstrated in neural network models by e.g. Elman (1990) and many others. Elman (1990) reported good generalisation of sequences by recurrent backpropagation networks in grammar learning. Pollack (1991) found generalisation of such networks to arbitrarily long sequences. As in human learning, generalization in neural networks is based in part on similarity of old and new sequences but also in part on certain structures exhibited by the sequences. Thus, the bottom level of CLARION, which incorporates a backpropagation network, has the capability to capture the generalization exhibited in human implicit learning. (Explicit processes, as in the top level of CLARION, can also generalise, albeit in a different way as discussed in Sun et al., 1995.)

It has also been observed that implicit processes are more robust than explicit processes (Reber, 1989) in the face of internal disorder and malfunctioning. For example, Hasher and Zacks (1979) found that encoding of frequency information (an implicit process) was correctly performed by clinically depressed patients, even though they could not perform explicit tasks consciously. Warrington and Weiskrantz (1982) found that amnesics were more successful in performing implicit rather than explicit memory tasks. This effect is consistent with the dual representation framework of CLARION: while the top level employs localist representation and is thus more vulnerable to malfunctioning, the bottom level utilises a distributed representation that is more resistant to damages and faults, as demonstrated amply in neural network models.

3.4. Unconscious Initiation of Action

Existing evidence indicates that unconscious processes often (if not always) initiate actions in skilled performance in advance of conscious awareness. Libet (1985) reported that electrophysiological ''readiness potentials'' (RPs) always precede conscious initiation of an act that is fully endogenous and voluntary. After a conscious intention to act appears, whether the action actually takes place or not can still be decided consciously by a subject within a time period of somewhere between 100 and 200 ms. As suggested by Libet (1985), the role of the conscious mind is not to initiate a specific course of action, but to control and influence (implicitly selected and initiated) actions.

This view is consistent with that of Willingham et al. (1989) that the role of explicit processes is to influence the implicit process but not to directly take control of skill learning or performance. Willingham et al. (1989) posited that the effects from the two processes are ''superimposed'' on each other, so that each type complement each other. Kelley and Jacoby (1993) also insisted that an important function of the explicit mind is to oppose, or counterbalance, the influence of the implicit mind.

The aforementioned view is also consistent with

[5] There is also evidence that explicit knowledge may develop independently. Willingham et al. (1989) reported such data. These data rule out the possibility that one type of knowledge is *always* preceded by the other type, at least under their experimental conditions. To account for this phenomenon, in CLARION, explicit hypothesis testing can be employed in the top level for learning rules, independently of the bottom level, as mentioned before.

voluminous data on the ever-present role of unconscious processes in all kinds of tasks: lexical priming, semantic processing, visual attention, unconscious perception and so on (as discussed in Velmans, 1991; Marcel, 1983). Velmans (1991) summarized evidence for the existence of implicit (preconscious) analysis of input stimuli, implicit processing of semantic content of word pairs in ''shadowing'' experiments, and implicit processing of bilingual messages in similar experimental settings. Most of these findings support the possibility that unconscious processes start before conscious processes take hold.

CLARION can readily accommodate this phenomenon, in that the bottom level, which captures unconscious processes, can work independently and initiate processing without the involvement of the top level. However, after the initiation of action and, consequently, the activation of the relevant nodes, the corresponding representations at the top level can then be activated by the bottom-up information flow (see Section 2). The activated explicit representations and their associated processes at the top level will in turn influence the implicit processing at the bottom level, in way of modifying and rectifying its outcomes and decisions (through the combination mechanism; see Section 2). Thus the implicit processes, which directly control actions in skilled performance, incorporate the results of explicit processes from the top level.

3.5. Synergy Between the Conscious/Unconscious Processes

Why are there two separate (although interacting) systems, one conscious and the other unconscious? Based on earlier discussions, we may hypothesize that each system serves a unique function and the two are complementary to each other; that is, there may be a synergy between the conscious and the unconscious. Such a synergy may show up by speeding up learning, improving learned performance, and facilitating transfer of learned skills.

In terms of speeding up learning, Stanley et al. (1989) reported that in a dynamic decision task (the sugar factory task), subjects' learning improved if they were asked to generate verbal instructions for other subjects along the way during learning. Willingham et al. (1989) found that those subjects who acquired full explicit knowledge appeared to learn faster.

In terms of learned performance, Willingham et al. (1989) found that subjects who verbalized while performing were able to attain a higher level of performance, because the requirement that they verbalised their knowledge prompted the formation and utilization of explicit knowledge. In high-level skill acquisition, Gick and Holyoak (1980) found that good problem solvers could better state rules that described their actions in problem solving. This phenomenon may be related to the self-explanation effect (Chi et al., 1989):

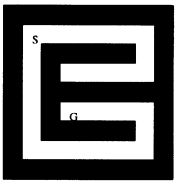

FIGURE 5. The Maze. The starting position is marked by ''S'' in which the agent faces upward to the upper wall. The goal is marked by ''G''.

subjects who explained the examples in textbooks more completely did better in solving new problems. In all these cases, it may well be the explication and the use of explicit knowledge that helped the performance.

In terms of facilitating transfer of skills, Willingham et al. (1989) obtained some suggestive evidence that explicit declarative knowledge facilitated transfer of skilled performance. It was reported that: (1) subjects who acquired explicit knowledge in a training tasks tended to have faster response times in a transfer task; (2) these subjects were also more likely to acquire explicit knowledge in the transfer tasks. In high-level domains, Ahlum-Heath and DiVesta (1986) also found that the subjects who were required to verbalize while solving the Tower of Hanoi problem performed better on a transfer task after training.

Sun et al. (1995) reported some simulation experiments that demonstrated CLARION was able to exhibit analogous synergy effects in learning, performance, and transfer through the interaction of the two levels. The simulation experiments were conducted in two domains: maze running and navigation through minefields. The details of the experiments and complete data can be found in Sun et al. (1995). Briefly, in the maze task, a subject/agent was to find a unknown target in the maze and had only rudimentary sensory inputs regarding its immediate left, front and right side, indicating whether there was a wall, an opening, or the goal; the agent could move forward, turn to the left, or turn to the right, until it found the target (see Figure 5). In terms of speeding up learning, the differences in learning speeds between *Q*-learning (which in CLARION captures unconscious learning at the bottom level) and CLARION (which includes both unconscious and conscious learning) were very significant. In terms of trained performance (measured by the average number of steps needed to reach the target in one episode), CLARION outperformed pure *Q*-learning by large margins again. We also compared the trained performance of the bottom level of CLARION alone (after the training of the entire system together, including *Q*-learning) with the performance of pure *Q*-learning, and discovered that the explication of skills not only improved the performance of the whole system, but it also improved the *Q*-learning part when included as part of CLARION. We also assessed the performance of trained models in a new and larger maze. CLARION transferred much better than *Q*-learning alone

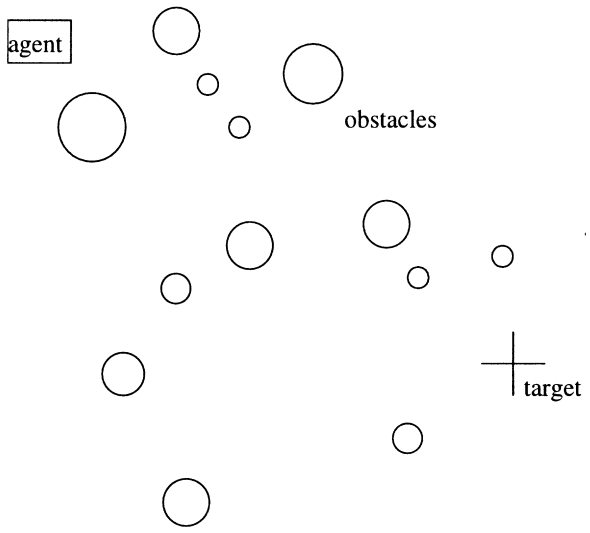

FIGURE 6. Navigating through mines.

(in terms of number of steps to reach the goal in one episode). Furthermore, by comparing the corresponding performance of the top level, the bottom level and the whole CLARION model, it was clear that often learned rules alone (the top level) performed better in transfer than the bottom level, as well as than the whole CLARION model, which showed that explicit knowledge facilitated transfer.

In the simulated navigation task shown in Figure 6, the subject/agent had to navigate an underwater vessel through a minefield to reach a target location. The agent received only local information from a number of instruments, as shown in Figure 7. Using only this information, the agent decided (1) how to turn and (2) how fast to move, and within an allotted time period, could either (a) reach a target (which is a success), (b) hit a mine (a failure), or (c) run out of fuel (a failure). In terms of learning speeds, the superiority of CLARION over *Q*-learning was statistically significant. To access transfer, after training models on 10-mine minefields, we assessed performance of these models in new minefields that contained 30 mines. CLARION outperformed *Q*-learning. The difference between the best transfer of Q-learning and the best transfer of CLARION was statistically significant. In sum, CLARION is able to replicate similar findings in human conscious/unconscious learning.

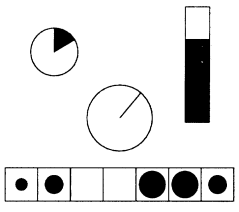

FIGURE 7. The navigation input. The display at the upper left corner is the fuel gauge; the vertical one it the upper right corner is the range gauge; the round one in the middle is the bearing gauge; the 7 sonar gauges are at the bottom.

4. THEORETICAL ISSUES

Some theoretical issues concerning consciousness will be discussed below in relation to the CLARION model.

4.1. Casual Efficacy of Consciousness

Is consciousness epiphenomenal as some have claimed? To see the casual role of consciousness, it is useful to examine the available cognitive. The following has been observed from patients who suffered the loss of some of the capacities of their consciousness due to brain damages (Marcel, 1988):

- They lost the ability to act on parts of the environment that were not accessible to their explicit/conscious mind (as in the case of blindsight patients who could not grasp objects in the blind field).
- They lost the ability to form an integrated self-concept (as in the case of amnesiac patients).
- They lost the ability to learn new complex tasks that required explicit, verbal instructions.
- They lost the ability to form explicit plans of actions before acting on them.

These effects indicate certain casual efficacy of consciousness, and are consistent with CLARION. In the model, these effects follow from the loss of the mechanisms in the top level of CLARION, such as backward chaining reasoning (planning), verbal instruction taking, and the use of explicit knowledge (regarding self and objects in the world).

Through contrasting ''aware'' vs ''unaware'' conditions on their experiments with human subjects, Kelley & Jacoby (1993) showed that conscious awareness per se had a distinct effect in subsequent behaviour. The two different conditions produced two different causal attributions in the subjects: one to true causes (in the ''aware'' condition) and the other to spurious causes (in the ''unaware'' condition; see Nisbett & Wilson, 1977 for causal attribution as error-prone, post hoc interpretation); consequently, different causal attributions led to different actions on the part of the subjects making the attributions. This kind of causal role in consciousness is consistent with that of the top level of CLARION, which shares the responsibility of controlling actions using explicit knowledge. A second mechanism on top of unconscious processes can offer counterbalance, and thus can have clear survival values to the agent possessing it.

4.2. Human Intuition

While the top level of CLARION captures conscious processes, the bottom level may capture *intuition* to some extent (as a form of skill): This level has the characteristics of being implicit, inaccessible, and holistic, which are also characteristics of human intuition (James, 1890; Dreyfus & Dreyfus, 1987). According to Smolensky

(1988), as mentioned before, intuition and skill are not expressible in linguistic forms and constitute a different kind of capacity, reflecting ''subsymbolic'' processing. Dreyfus & Dreyfus (1987) suggested that intuition is manifested in the fluent, holistic and situation sensitive way of dealing with the world, unique to humans and not captured by conventional symbolic computation. These identified characteristics can be found in the bottom level of CLARION to some extent.

It was hypothesised by Reber (1989) that human intuition may be the direct result of implicit, unconscious learning: Through the gradual process of implicit learning, ''tacit'' (implicit) representations emerge that capture environmental regularities and are used in direct coping with the world (without the involvement of any introspective process). Intuition is the end product of this process of unconscious and bottom-up learning (Reber, 1989). Bowers et al. (1990) also suggested that intuition is the outcome of an unconscious, implicit process (which later becomes explicit due to the emergence of a coherent pattern of activation) in the context of discovery. CLARION indeed uses implicit learning to develop tacit (implicit) representations in the bottom level and thus acquires intuition in the sense identified above.

4.3. Two Types of Consciousness

We can also examine the two levels of CLARION using the perspective of phenomenological philosophy. Heidegger (1927) emphasised a basic mode of existence, that is, the immediate comportment with the world. Normally, when going about its daily business, an agent is not *thematically* conscious of routine activities. Everyday routine activities are mostly made up of non-deliberate ''primordial'' coping. For example, in normal perception, we are usually not having thematic experience of the world (Dreyfus, 1992). An agent's ''openness onto the world'' is fundamental and makes possible the secondary experience of deliberate looking or trying to see (Dreyfus, 1992). Comportment is prior to any (explicit) belief, (explicit) knowledge, or (explicit) representation; it is a direct connection between an agent and its existential context. It is comportment with the world that is in fact a more fundamental kind of consciousness, according to Heidegger, and in this view, consciousness is non-representational (i.e. without explicit representation).

This is in contrast with the notion of ''thematic'' consciousness, which involves a focused, meditated awareness of the object of consciousness (akin to the common notion of consciousness). Thematic consciousness is representational because it treats awareness itself as an object (see also Clark & Karmiloff-Smith, 1993). As has been argued extensively by Heidegger (1927), thematic consciousness can indeed arise, but it presupposes a nonthematic, nondeliberate, direct, and on-going way of dealing with the world (i.e. comportment); for direct comportment with the world is a *necessary* means for coping with a complex world that exhibits complex regularities. Explicit representations are derived from direct comportment with the world. Derived representation and thematic consciousness come into play, e.g. during breakdown in which established routines get disrupted and thus alternative ways are necessary (see also Sun, 1994, Chapter 8). In the light of the above, the distinction between the two levels in CLARION can be corresponded roughly to the distinction between comportment and thematic consciousness. Specifically, the bottom level captures the implicit and routine activities and embodies ''comportment''. This is because the bottom level embodies skills resulting from and used for directly coping with the world and involves distributed representation, which is hence unable to present explicit traces of its processes (Smolensky, 1988). However, some events (e.g. certain implicit processes or explicit verbal inputs; Velmans, 1991) may lead to activation of corresponding explicit representations and processes at the top level and therefore lead to (thematic) conscious awareness. This is because at the top level explicit/localist representation is used, which makes it possible to articulate the content that is present and trace the processes as they are occurring (Sun, 1994, Chapter 2). Furthermore, due to the fact that the top level of CLARION is derived from, mediated by, and grounded in the bottom level, which has direct interactions with the external world and develops bottom-up from such interactions, (thematic) consciousness is clearly grounded, in this model as in humans, in the interaction between the agent and the external world.

4.4. Qualia

But what about qualia? Qualia refer to the phenomenal quality of conscious experience. Block (1994) distinguishes access consciousness and phenomenal consciousness, whereby access consciousness refers to the utilization of the content of consciousness while phenomenal consciousness refers to the subjective feel of conscious experience. Although it has been a major difficulty to understand phenomenal consciousness/qualia (''the hard problem''; Chalmers, 1992), some speculations may be made here in relation to CLARION: qualia (phenomenal consciousness) may be accounted for by the totality of a multi-modal (see the next section regarding modalities), multi-level organization and its total collective states, which are of extremely high complexity involving external perception (of many modalities), internal perception, action decision making, explicit concepts, etc. The complexity of this organisation may explain the difficulty (or impossibility) of describing phenomenal qualities (qualia) of consciousness, which is the most striking feature of phenomenal consciousness (as argued by Nagel, 1974). In this approach, a particular kind of phenomenal quality may be accounted for by a

particular region of total-state space (involving the totality of all the aforementioned aspects) or the manifold as termed by Van Gulick (1993), which gives rise to the sense of what something is like (Nagel, 1974). Clearly, such regions depend on particular *functional* organizations of modules and levels (Chalmers, 1992) that support such a space of total-states. Qualia are thus (partially) the result of functional organisations (architectures) of cognitive apparatuses. In CLARION, qualia are (partially) the result of the two-level organization, on top of all the detailed, intricate structures involved in various fine-grained modalities (the detail of which is not covered here). Equally important, such regions arise from the interaction of the agent and the world, as the consciousness is grounded in learning, action, and the world in which the agent exists. Thus, phenomenal consciousness is in general derived from a complex integrative organisation of cognitive apparatuses that develop as a result of the interaction between the agent and the world. This is true both in an ontological and in an ontogenetic sense. Note that, though access consciousness has a clear causal role in behaviour (as discussed earlier), phenomenal consciousness is less clear in this regard (Block, 1994).

4.5. Functionalism

An important question regarding consciousness is whether there is a physical basis for consciousness and what it is. It is pretty much agreed upon among contemporary philosophers that there is indeed a physical basis but they disagree on what constitutes that physical basis. In the CLARION model, we basically stipulate that the physical basis of consciousness is made up of the detailed architecture, the fine-grained functional organisation (e.g. the two-level modular framework with detailed sensory modality structures at a very fine level), of one's cognitive apparatus, in interaction with one's world. The distance between the physical and the phenomenological/psychological is so great that intermediate levels (functional levels) are necessary for studying cognition, and this is especially true in the case of studying consciousness. The present approach is a form of fine grained functionalism (Chalmers, 1992), which states that consciousness is invariant across systems with the same functional organization at a sufficiently fine-grained level (i.e. the principle of organizational invariance), as argued for elegantly by Chalmers (1992). It is a weak form of functionalism, in that it is not just causal connections between functional states that are important, but also the level at which we identify functional states. This approach is also interactionalism in the sense that the interaction of internal and external systems (the agent and its world), on the basis of internal (developing) fine-grained functional organizations, is crucial in giving rise to conscious experience (Heidegger, 1927).

5. COMPARISONS

The following comparisons will further explicate and accentuate the CLARION model in accounting for issues of consciousness. CLARION captures important features of existing models. The comparisons reveal that CLARION has the potential to account for the integration of sensory modalities, global consistency, and the unity of consciousness.

First we can compare CLARION with the model of Baars (1988) (see Figure 8), in which 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 modules in the bottom level of CLARION, and the global workplace may be roughly captured by the top level, which ''synthesizes'' the bottom level modules and is essential in conscious processing. One difference is that CLARION does not emphasise as much internal consistency (Marcel, 1983): it is believed to be limited as a phenomenon in consciousness and may have only limited roles in the emergence of consciousness. Global broadcasting in Baars' model (Baars, 1988; Revonsuo, 1993) can be viewed as the integration of the two levels of representations (with the bottom-level representations dispersed within multiple modules) in CLARION, which does produce somewhat consistent outcomes (which lead to the unity of consciousness; Baars, 1988; Marcel, 1983; more on this later).

We can also compare CLARION with the model of Schacter (1990) (see Figure 9), which is based on neurophysiological findings of the dissociation of different types of knowledge in brain damaged patients as mentioned earlier. It is similar to CLARION, in that it includes a number of ''knowledge modules'' that perform specialized and unconscious processing (analogous to bottom-level modules in CLARION) and send their

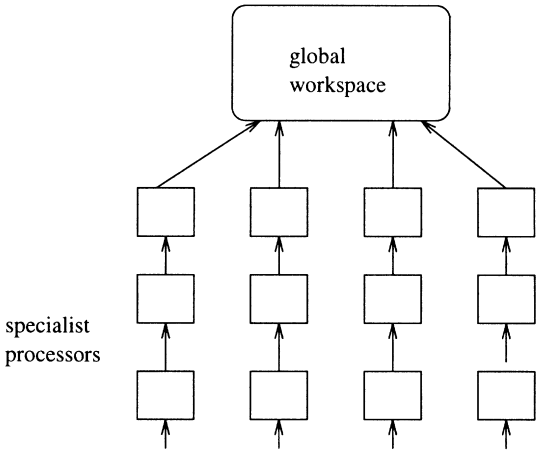

FIGURE 8. Baars' model of consciousness.

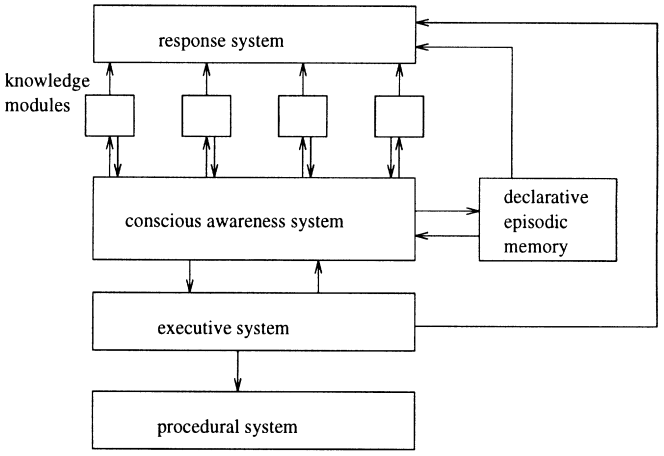

FIGURE 9. Schacter's model of consciousness.

outcomes to a ''conscious awareness system'' (analogous to the top level in CLARION), which gives rise to conscious awareness. Schacter's explanation of some disorders (e.g. the loss of short-term memory or explicit learning abilities, as mentioned earlier) is that certain brain damages result in the disconnection of some of the modules from the conscious awareness system, which leads to their inaccessibility to consciousness. An alternative explanation offered by CLARION is that disorders may not be due to disconnected modules, but the loss of some explicit learning and performance mechanisms at the top level (resulting from brain damages etc.).

Finally, we can examine Damasio's neuroanatomically motivated model (Damasio, 1994; Revonsuo, 1993). The model (see Figure 10) hypothesised 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 hypothesised the global ''multi-modal convergence zone'', which integrated information across modalities also through reverberation via recurrent connections. In CLARION, different sensory convergence zones may be roughly captured by bottom-level modules, each of which takes care of sensory inputs of one modality (at a properly fine level), and the role of the global multi-modal convergence zone (similar to the ''global workspace'' in a way) may be played by the top level of CLARION, which has the ultimate responsibility for integrating information (and serves as ''conscious awareness system''). The widely recognised role of reverberation (Damasio, 1994; Taylor, 1994) may be captured in CLARION through using recurrent connections within modules at the bottom level and through multiple top-down and bottom-up information flows across the two levels, which lead to the unity of consciousness that is the synthesis and integration of all the information present (Marcel, 1983; Baars, 1988).

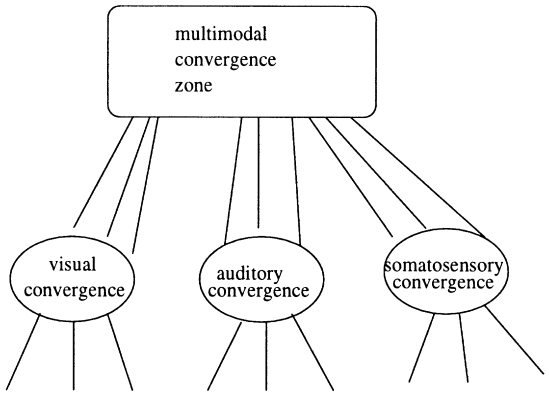

FIGURE 10. Damasio's model of consciousness.

6. CONCLUDING REMARKS

This paper presented a hybrid neural network model for learning that incorporated the distinction of declarative and procedural knowledge, and succeeded to some extent in accounting for the distinction of the conscious and the unconscious (or the explicit and the implicit). More specifically, the CLARION model applied neural network and machine learning techniques to explain complex human learning and consciousness in normal human subjects. It accounted for phenomena in psychological literature on learning and development in terms of the two levels in the model and their associated mechanisms. The model readily accommodated important features of existing models of consciousness. We also had something to say about theoretical issues such as qualia and intuition on the basis of the model, which helped clarify complex issues in a tangible way. The key issue for future research is scaling up the model by incorporating a variety of sensory information and dealing with complex environments, in order to allow rudimentary forms of consciousness to emerge, because, I believe, complexity is a necessary condition for consciousness.

REFERENCES

Ahlum-Heath, M., & DiVesta, F. (1986). The effect of conscious controlled verbalization of a cognitive strategy on transfer in problem solving. *Memory and Cognition*, *14*, 281–285.

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

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

Berry, D., & Broadbent, D. (1984). On the relationship between task performance and associated verbalizable knowledge. *Quarterly Journal of Experimental Psychology*, *36A*, 209–231.

Block, N. (1994). On a confusion about a function of consciousness. *Brain and Behavioral Sciences*.

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

Bruner, J., Goodnow, J., & Austin, J. (1956). *A study of thinking*. New York: Wiley.

Chalmers, D. (1993). *Towards a theory of consciousness*. Ph.D Thesis, Indiana University.

Chi, M., Bassok, M., Lewis, M., Reimann, P., & Glaser, P. (1989). Self-explanation: how students study and use examples in learning to solve problems. *Cognitive Science*, *13*, 145–182.

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

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

Cohen, A., Ivry, R., & Keele, S. (1990). Attention and structure in sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 17–30.

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

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

Dennett, D., & Kinsbourne, M. (1992). Time and the observer. *Behavioral and Brain Science*, *15*, 183–200.

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

Dreyfus, H., & Dreyfus, S. (1987). *Mind over machine: the power of human in tuition*. New York, NY: The Free Press.

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

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

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

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

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

Gick, M., & Holyoak, K. (1980). Analogical problem solving. *Cognitive Psychology*, *12*, 306–355.

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

Heidegger, M. (1927). *Being and time* (English translation, 1962). New York: Harper and Row.

James, W. (1890). *The principles of psychology*. New York: Dover.

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

Keil, F. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

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

LeDoux, J. (1992). Brain mechanisms of emotion and emotional learning. In *Current opinion in neurobiology*, **2**(2) (pp. 191–197).

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

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

Lin, L. (1992). Self-improving reactive agents based on reinforcement learning, planning, and teaching. *Machine Learning*, *8*, 293–321.

Mandler, J. (1992). How to build a baby. *Psychology Review*, *99*(4), 587–604.

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

Marcel, A. (1988). Phenomenal experience and functionalism. In A. Marcel & E. Bisiach (Eds.), *Consciousness in contemporary science*. Oxford, UK: Oxford University Press.

Medin, D., Wattenmaker, W., & Michalski, R. (1987). Constraints and preferences in inductive learning: an experimental study of human and machine performance. *Cognitive Science*, *11*, 299–339.

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

Mitchell, T.(1982). Generalization as search. *Artificial Intelligence*, *18*, 203–226.

Murphy, G., & Medin, D. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289–316.

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

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, *4*, 435–450.

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

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

Posner, M., & Snyder, C. (1975). Facilitation and inhibition. In P. Rabbit & S. Dornick (Eds.), *Attention and performance*. Academic Press.

Pollack, J. (1991). The induction of dynamic recognizers. *Machine Learning*, *7*(2/3), 227–252.

Reber, A. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*, 855–863.

Reber, A. (1976). Implicit learning of synthetic languages: the role of instructional set. *Journal of Experimental Psychology. Human Learning and Memory*, *2*, 88–94.

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

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

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

Seger, C. (1994). Implicit learning, Psychological Bulletin, *155*(2), 163–196.

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge University Press.

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

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

Stadler, M. (1992). Statistical structure and implicit serial learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 318–327.

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

Sun, R. (1992a). A connectionist model for commonsense reasoning incorporating rules and similarities. *Knowledge Acquisition*, *4*, 293–321.

Sun, R. (1992b). On variable binding in connectionist networks. *Connection Science*, *4*(2), 93–124.

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

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

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

Sun, R., Peterson, T., & Merrill, E. (1995). *Hybrid architecture and situated learning* (Technical Report TR-CS-96-0019). University of Alabama. submitted to a journal.

Sun, R., Peterson, T., & Merrill, E. (1996a). *Bottom-up skill learning* (Technical Report TR-CS-96-0021). University of Alabama. submitted to a journal.

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

Sutton, R. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. *Proc. of 7th International Conference on Machine learning*. San Mateo, CA: Morgan Kaufmann.

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

Towel, G., & Shavlik, J. (1993). Extracting Refined Rules from Knowledge-Based Neural Networks. *Machine Learning. 7*.

Van Gulick, R. (1993). Understanding the phenominal mind. In M. Davies & G. Humphries (Eds.), *Consciousness*. Oxford, UK: Blackwell.

Velmans, M. (1991). Is human information processing conscious? *Behavioral and Brain Sciences*, *14*, 651–726.

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

Warrington, E., & Weiskrantz, L. (1982). Amnesia: a disconection syndrome?. *Neuropsychologica*, *20*, 233–248.

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

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