Coward, L. A. and Sun, R. (2004). Some Criteria for an Effective Scientific Theory of Consciousness and Examples of Preliminary Attempts at Such a Theory. Consciousness and Cognition. In press.

Some Criteria for an Effective Scientific Theory of Consciousness and Examples of Preliminary Attempts at Such a Theory

L. Andrew Coward
School of Information Technology
Murdoch University
Perth, Western Australia
Australia
landrewcoward@shaw.ca

Ron Sun
Department of Cognitive Sciences
Rensselaer Polytechnic Institute
110 8th Street, Troy, NY 12180
USA
rsun@rpi.edu

Abstract

In the physical sciences a rigorous theory is a hierarchy of descriptions in which causal relationships between many general types of entity at a phenomenological level can be derived from causal relationships between smaller numbers of simpler entities at more detailed levels. The hierarchy of descriptions resembles the modular hierarchy created in electronic systems in order to be able to modify a complex functionality without excessive side effects. Such a hierarchy would make it possible to establish a rigorous scientific theory of consciousness. The causal relationships implicit in definitions of access consciousness and phenomenal consciousness are made explicit, and the corresponding causal relationships at the more detailed levels of perception, memory, and skill learning described. Extension of these causal relationships to physiological and neural levels is discussed. The general capability of a range of current consciousness models to support a modular hierarchy which could generate these causal relationships is reviewed, and the specific capabilities of two models with good general capabilities are compared in some detail.

Key words: access consciousness; phenomenal consciousness; implicit memory; explicit memory; skill learning; neurophysiology; hybrid architecture; cognitive architectures

1. Introduction

Investigations of consciousness have ranged from attempts to develop computational theories to the more limited goal of identifying the neural correlates of consciousness. One major element has been an argument around whether a scientific account of consciousness is possible, or whether there could be an explanatory gap in understanding how physiological processes generate the "what it is like" of consciousness.

The paradigm for a successful scientific theory is the physical sciences. This paper reviews some of the characteristics of physical theories as a model for what is needed for consciousness. Such theories establish hierarchies of description on many levels of detail in which causal relationships on one level can be mapped into causal relationships on any other level. It is then argued that the requirement to learn without undesirable side effects

on prior learning will have forced the brain into the form of a modular functional hierarchy which has properties making it an effective vehicle for a hierarchical scientific theory.

The causal relationships implicit in generally used definitions of access and phenomenal consciousness are then made explicit. The equivalent but more detailed causal relationships observed in explicit and implicit mental processes are reviewed. A range of architectural models of consciousness are evaluated for their potential capability to support the causal relationships at the highest level, and two models are found to have such potential. The ability of these two models to support the causal relationships found in explicit and implicit mental processes are then considered in more detail, and the consistency of the models with psychology and physiology briefly evaluated.

The conclusion is reached that models of this type have some potential to become scientific theories of consciousness with explanatory capability analogous with theories in the physical sciences.

2. Characteristics of Scientific Theories

What could a scientific theory of consciousness tell us, and what would be the limits of such a theory? The physical sciences are often taken as the paradigm for rigorous scientific theory. In this domain theories have achieved a high degree of mathematical sophistication over the last 400 years. To define the nature of an effective theory of consciousness it is useful to analyze the nature of theories in the physical sciences. The aim of our discussion is to give a relatively simple view of what science aims to achieve. More extensive discussion from a similar viewpoint can be found in, for example, Lakatos (1970) or Salmon (1984).

At the highest level, domains of human experience are the roots from which the physical sciences have developed. Experience of movement in day-to-day life, in hunting or in warfare, has led to theories of motion from Newtonian mechanics through Hamiltonian dynamics to special relativity and quantum electrodynamics. The ability to derive dates for planting crops from an understanding of the behaviour of stars and planets has led ultimately to theories of gravity from Newtonian to general relativity. The experiences of materials processing such as metalworking led to successively deeper sciences of matter from chemistry through atomic theory to quantum mechanics and on to string theory. Deeper theories provide accounts that unify multiple apparently independent domains at higher levels. This unification is achieved by using small sets of relatively simple concepts which can be combined in different ways to generate the concepts and causal relationships of the higher level theories. However, deeper theories do not in general completely replace higher level theories in the practice of science, although they may result in significant revisions. For example, the deeper level theory of quantum mechanics has not replaced the use of theories at the level of chemistry, and even research at the frontiers of quantum mechanics makes extensive use of classical physics [Greene 1999, page 380].

A successful theory on any level creates entities at that level of descriptive detail and causal relationships between those entities that correspond exactly with a range of data. Entities at a higher level package sets of deeper level entities in such a way that the higher level causal relationships can be specified without reference to the internal structure of those higher level entities. However, causal relationships between detailed entities must generate all the causal relationships which exist between the higher level entities. Some higher level relationships may be generated probabilistically from more detailed

relationships, for example, the behaviour of gases from the statistics of large numbers of gas molecules [1].

This definition of a successful theory has some similarities with the approach proposed by Machamer et al (2000), but with some important differences. Their approach is that the world should be conceived as being composed of entities and activities, and mechanisms are composed of entities and activities. Our definition of a causal relationship is essentially equivalent to their mechanism, and their definition of an activity is in our view equivalent to a way in which two entities can interact. The critical difference is that in our view, a driving force for the creation of multiple levels of description is that physical scientists, like all human beings, can only handle a limited amount of information at one time. They must therefore use a high level theory for thinking about broader phenomena, and then focus through successive levels of detail on more detailed theories. To keep the information content of descriptions on every level within human capabilities, entities must be arranged in hierarchies and defined in such a way that the interaction between two entities at a high level is a very small proportion of all the interactions within the population of more detailed entities which makes up the two higher level entities. Whether the entities have any "absolute" reality or are simply defined by humans as an aid to understanding is immaterial to our discussion.

For example, pre-atomic or macro chemistry established entities like sulphuric acid and caustic soda, and on a somewhat more detailed level entities like acids, alkalis, and salts. A consistent causal relationship (or law) on that level is acid plus an alkali generates a salt plus water. The number of different possible chemicals in macro chemistry is extremely large. Atomic theory uses less than one hundred different types of atom, and by means of intermediate entities like hydroxyl and hydrogen ions can generate the acid-plus-alkali law. However, a full end-to-end description of a macro chemistry process would be long and convoluted at the atomic level, because it would include a description of the behaviours of all the approximately 10^{24} atoms making up a typical chemical reaction. The information content of such a description would make it too complex to be grasped within human cognitive capacities. In practice the physical sciences establish multiple levels of description to bridge the gap between macro phenomena and the most fundamental theories like quantum mechanics.

Bridging is achieved by creating descriptions at a detailed level for small but typical processes within a high level phenomenon, and determining consistency with the high level phenomenon as far as possible. Consistency means that the causal relationships between the high level entities in the process element are equivalent to the causal relationships between the groups of detailed entities making up the high level entities. Any inconsistency would invalidate the theory. If a large and representative sample of such consistency tests all give positive results, the detailed theory is considered valid. In this situation a correspondence exists between high and detailed level descriptions. In a successful theory, causal relationships at the deeper level predict unexpected but experimentally confirmable causal relationships at higher levels.

---

[1] It might also be noted that in the physical sciences the relationships between causal relationships at higher and lower levels can be subtle, not just a simple correspondence. Direct causal relationships at a higher level may become mere statistical tendencies at a lower level, and many causal relationships at the lower level may disappear at the higher level.

Note that the actual high level processes which occur in experience are only a tiny subset of all the high level processes which could conceivably occur given the detailed level entities. Although any conceivable high level process could be described in terms of the detailed entities, the subset which actually occurs is determined by the actual configuration of entities which happens to exist (in physical science terms, the boundary conditions). The identity of the subset cannot be determined a priori from the properties of the individual entities alone. In this sense the high level processes are "emergent".

A scientific theory of consciousness, by analogy with the physical sciences, would create descriptions of equivalent causal relationships on the level of consciousness and via a number of intermediate levels to a computationally and physiologically detailed description. There would be fewer types of entity and causal relationship at more detailed levels, but higher level entities would be made up of sets of more detailed entities, and the causal relationships at high level would be generated by the causal relationships between the equivalent entities at more detailed levels. However, the causal relationships between higher level entities could be described without reference to their internal structure.

The capability, at least in principle, to map phenomenological properties to neuron properties is an essential aspect of an effective theory. The ability of an intermediate theory to accurately model high level phenomena is a necessary but not sufficient condition for effectiveness. Thus in astronomy, the Ptolomaic theory of planetary motion based on epicycles around a hierarchy of mathematical points could account for observations at least as accurately as the Copernican model when that model was proposed. Addition of more epicycles could enable it to match observations to yet higher degrees of accuracy. However, a model based on orbits around mathematical points had much less capability for deeper level modeling than the Copernican model in which an orbit was generally related to the presence of an identifiable astronomical object at a central point in the orbit.

3. Mechanistic views of consciousness

Such a theory would be a mechanistic explanation, i.e. an explanation which "treats [the cognitive] systems as producing a certain behaviour in a manner analogous to that of machines developed through human technology" (Bechtel et al 1993). What can be expected of such a theory of consciousness is an appreciation of how causal relationships at a deeper level give rise to the phenomena of consciousness at the phenomenological level, a better understanding of causal relationships at the psychological level, and a understanding of how silicon-based (electronic) systems might generate conscious phenomena.

A mechanistic explanation is generally taken to be computational in the broad sense of that term. Here the term "computation" is 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; although we do not believe in its relevance). In terms of the sufficiency of such computational explanations, Jackendoff (1987) proposed the following hypothesis[2]: "Every phenomenological

[2] Critics such as Edelman 1989, Freeman 1995, Damasio 1994, Penrose 1994, Searle 1980) failed to show that computation, in general, cannot account for the nature of consciousness, although they had some legitimate complaints about specific computational approaches and models.

distinction is caused by/supported by/projected from a corresponding computational distinction".

A theory of consciousness which is analogous with a theory in physical sciences must first establish a set of entities or conditions C1, C2, C3, C4 etc. at a high level with consistent causal relationships such as the presence of C1 and C2 results in C3[3]. Then, the theory must also establish a smaller set of entities or conditions at a lower level so that different combinations of entities or conditions c1, c2, etc. correspond with the high level C1, C2, C3, C4 etc. in such a way that if C1 plus C2 causes C3 at the high level of description, then at the detailed level the combination of c1, c2 etc. corresponding with C1 plus C2 at high level causes the combination of c1, c2 etc. corresponding with C3 at high level. Descriptions at the high level contain less densely packed information than descriptions at a detailed level. This means that the number of entities which are needed on the detailed level in general should be smaller than the number of entities at the high level, and the number of different types of causal relationships at the detailed level generally should be smaller as well.

The lower density of descriptions at the higher level means that it is possible that many states at the detailed level could correspond with the same state at the higher level. We may generalize the token identity theory concerning psychophysical correspondence (see e.g. Braddon-Mitchell and Jackson 1996): instances of a high level (psychological) state are instances of detailed level (physical) states (i.e. correspondence of categories across levels). However, consistent with the supervenience principle (no mental difference without a physical difference, see Kim 1993), no two different states at high level can correspond with the same state at a detailed level.

3.1 Implications of the Requirement for Simple Functional Architectures

In support of the multiple level approach toward consciousness, consider the use of such an approach in other areas. It is not always realized that, for the design and management of complex systems (e.g. computers), hierarchies of descriptions exist on many levels of detail. In such systems, a "functional architecture" exists which makes it possible to describe the operation of the system on many levels of detail. For example, carrying out a spell check within a word processing program could be described at the highest level as "check the spelling of all the words in this document". At a more detailed level, the description might include elements such as "load the spelling tool from hard disk to RAM". At yet more detailed levels the description would include software instructions such as " (test = a) whileTrue: [do ….]" and at an even deeper level would include assembly code instructions such as "add 1,xy". In such a functional architecture, the functionality of the system at the highest level is expressed as functional requirements which are separated into relatively independent modules. Each module is separated into sub-modules, sub-modules into yet smaller components and so on until the smallest functional elements of the system are reached. In a complex real time system such as a telecommunications switch, (Nortel Networks 2001) there may be seven or eight levels of description from highest level functional requirements to executable machine code level,

[3] Such causal relationships might include: perception of a red object and a verbal input of "What color is that" causes the perceiver to say "That is red"; or perception of a red object and a verbal input of "What are your feelings right now" causes the perceiver to say "I feel sad"

and a similar number of levels of description for the hardware, from system description to transistor, which interact at different levels with the software hierarchy. The modularization is strongly constrained: all modules on one level should perform roughly the same proportion of system operations, and although information exchange between modules is essential to coordinate functionality, such information exchange should be minimized as far as possible. A module is defined as a group of system operations which interact (i.e. exchange information) more strongly within the group that outside the group. The widely used concept of information hiding introduced by Parnas (1972), in which a module processes information hidden from other modules, is another way of expressing this definition. Minimization of information exchange overall is equivalent to requiring that the difference between internal and external interaction is as large as possible across all modules in the system. The reasons for modular architectures are that such architectures make it easier to modify functionality, diagnose and repair problems, and perform relatively independent design by different designers (see e.g. Kamel, 1987; Bass et al. 1998).

To understand these reasons, imagine a design process in which (in a caricature of biological evolution) a large number of technicians were given a large number of devices and told to connect at random. Periodically the result would be tested, and eventually a system found which performed as required. There are some severe problems with such a system. There are no blueprints which can guide the process of building another copy: the only option is to duplicate the original, device by device, connection by connection. If an error were made in such a duplication process a functional problem would result, but there would be no easy way to use the knowledge of the problem to identify and correct the error. Similarly, if a device failed during operation it would be very hard to identify which device was defective. Finally, device level changes would generate complex and unpredictable functional changes, and if there was a need to modify the functionality in a controlled fashion there would be no way to identify what device level changes would produce the desired functional modifications (Coward 1999b). There should therefore be relatively simple logical paths which can relate high level system functionality to the operations of transistors and elements of code. The hierarchy of modules as described makes it possible for such paths to exist[4]. Such a hierarchy of modules is effectively a hierarchy of descriptions of exactly the same functionality at different levels of detail. Biological brains are not the result of a deliberate design process, but are constrained by the needs to be constructed from DNA "blueprints", recover from construction errors and damage, and learn new behaviors without introducing random, undesirable changes into existing behaviors. Coward (2001) has therefore argued that selection pressures deriving from the advantages of ease of construction, recovery from damage, and unconfused learning might result in brains being constrained to adopt simple functional architectures as defined earlier. Brains may therefore have an intrinsic "description" hierarchy[5] which

[4] The requirement for a hierarchy of this type only occurs if the system must perform a very complex combination of interacting features with limited information handling resources in such a way that some features can be changed without side effects on other features. In most artificial neural network and robotic applications the set of features perfomed is small relative to the resources available, and the need for the hierarchy is much less.

[5] For example, Goguen (1977) proposed a formal system description approach in which objects and relationships between objects are specified, and higher level objects are constructed from more detailed objects and relationships. Goguen applied the approach to describing music.

could link the phenomena of consciousness to those of physiology. Within such an architecture, high level descriptions of the phenomena of consciousness could be mapped through successively more detailed descriptions.

In cognitive science many different types of module have been proposed, including peripheral modules, domain specific modules, and anatomical modules. Proposed peripheral modules include early vision, face recognition, and language. Such modules take information from a particular modality and only perform a specific range of functions. Proposed domain specific modules include driving a car or flying an airplane (Hirschfeld and Gelman 1994; Karmiloff-Smith 1992). In such modules highly specific knowledge and skill is well developed in a particular domain but does not translate easily into other domains. Anatomical modules are isolatable anatomical regions of the brain which work in relative independence (Damasio 1994). In an early discussion of cognitive modules, Fodor (1983) proposed that such modules should have proprietary input domains and dedicated neural hardware, and generate information in accordance with algorithms which are not shared with or accessible to other modules (i.e. information hiding as defined by Parnas 1972).

All these module definitions can be interpreted as different instantiations of the requirement for less external interaction than internal interaction. Different types of modules may be an effective way to achieve minimized information exchange in different parts of the system[6]. Marr (1982) argued that complete understanding of an information processing device required understanding on three levels: the levels of computational theory, representation and algorithm, and hardware implementation. However, Marr's position was that the three levels are only loosely coupled, since "The choice of algorithm is influenced, for example, by what it has to do and by the hardware in which it must run. But there is a wide choice available at each level, and the explication of each level involves issues that are rather independent of the other two." Functional architecture based description hierarchies are rather different. The functionality described on each level is precisely the same, the difference is only in the information density (level of details) and the length of the description, and in the smaller number of entities needed by descriptions at the more detailed level. Causal relationships at one level will always have to correspond with causal relationships at other levels.

4. The Definition of Consciousness

Even the concept of consciousness is controversial, because of the wide range of different phenomena to which the term "conscious" is applied, and because of the difficulty of objective measurement. A recent attempt at a definition was that of Block (1995) who distinguished between access consciousness, monitoring consciousness, self consciousness, and phenomenal consciousness. Access conscious was defined as the ability to report and act upon experiences. Block suggested that this ability is equivalent

[6] The most effective modules may not correspond in any simple fashion with system features. In functionally complex electronic systems modules are defined so that the operation of one feature requires the participation of as few modules as possible, but in general one module or one group of modules will not correspond with one feature on any level, because such one-to-one correspondence would not be compatible with minimizing intermodule interaction (see e.g. Kamel 1987; Bass, Clements and Kazman 1998). For analogous reasons in a cognitive system modules may well not correspond with skill domains or other obvious cognitive features. Hence the inability to find modules corresponding with such features is not an argument against a modular functional hierarchy (Coward 2001).

to the existence of some representation of the experience in the brain, the content of which is available for verbal report and for high level processes such as conscious judgments, reasoning, and the planning and guidance of action. Monitoring consciousness refers to thoughts about one's sensations as distinct from the sensations. Self consciousness refers to thoughts about self. Phenomenal consciousness refers to the qualitative nature of experience, for example why the experience of the colour red feels as it does and not like something else.

In order to establish a description hierarchy based theory, the first requirement is a definition of consciousness which includes causal relationships between cognitive entities. To make the causal relationships implicit in the Block definition explicit, consider the following scenario and behaviours which might result.

In the scenario, a person with a sore ankle is out walking with a companion, and encounters a tree partially blocking the path. One behaviour is simple avoidance: stepping around the tree in a way which minimizes the stress on the sore ankle. A second behaviour is speaking a warning "mind the tree" to the companion. A third behaviour is to make the comment "Do you see that tree? It's a Hemlock ". A fourth behaviour is to notice the pain from the ankle, and take a little longer to decide how to place the feet in getting around the tree. A fifth behaviour is to plan how the feet are placed to minimize stress, remember how a stumble caused the soreness, and to become irritated at the carelessness which caused the stumble.

The first behaviour can generally be unconscious. Sensory input from the tree and the sore ankle etc. generates some activation state internal to the brain which leads to avoidance behavior but does not lead to higher cognitive functions or verbal report. The third behaviour falls within the definition of access consciousness. The internal brain activation state in response to the tree (or "representation") generates both cognitive processing and a complex verbal report. The second behaviour is of interest because it appears to indicate that simple verbal warnings can be initiated by an unconscious activation without much cognitive processing.

The entities and causal relationships suggested by these scenarios could be visual input from a tree, pain input from the ankle, unconscious activation, conscious activation (using this term in preference to "representation" because the latter term carries implications which may or may not be present), avoidance behaviour, verbal warning behaviour, higher cognitive behaviour (including associative thinking, and verbal report of associative thinking). Sensory input can cause an unconscious and/or a conscious activation. An unconscious activation can cause avoidance behaviour and simple verbal warnings. A conscious activation can cause avoidance behaviour and simple verbal warnings, and can also cause higher cognitive processing and complex verbal reports of that processing.

In the fourth behaviour, the pain from the ankle generates a conscious activation which includes associative thinking and affects behaviour generation and which also has the capability of generating a complex verbal report. This fourth behavior corresponds with monitoring consciousness. In the fifth behaviour, visual input from the feet generates a conscious activation, again with associative thinking, effects on emotion (creating irritation), and the capability of generating a complex verbal report. This fifth behaviour corresponds with self consciousness. The fourth and fifth behaviours do not generate any qualitatively different causal relationships.

Phenomenal consciousness is a somewhat different case because it is not immediately obvious what causal relationships are relevant, although clearly there is something which it is believed justifies further explanation. The argument is made that there is "something that it is like" to experience a pain or a colour. Furthermore, although there is nothing that it is like to be a book or a rock, there is something that it is like to be you or me (Nagel 1974).

A possible approach, which Dennett (2001) has labeled heterophenomenology, is to regard first person reports as phenomena which must be explained from a third person viewpoint, and focus on the behaviour which results from different conscious activations. The conscious experience of a pain or a colour is felt to have a complex structure which is dynamic and may be different for experiences of similar conditions on different occasions. However, it is extremely hard to describe that complexity. Hence the causal relationships associated with phenomenal consciousness are firstly that conscious activations can generate verbal reports of the presence of complexity and of differences between activations in response to similar conditions, but cannot generate verbal reports of the details of the complexity. Secondly, conscious activations in response to similar conditions can generate different behaviours such as different emotional states.

The causal relationship that conscious activations can generate complex verbal reports and unconscious activations cannot is essentially equivalent to the widely used accessibility distinction between conscious and unconscious processes (see e.g. Clark 1992, Hadley 1995). Directly accessible means, for example, that such processes can be expressed verbally without complex intermediate interpretive or transformational steps. Unconscious processes are not accessible directly (see, e.g., Heidegger 1927, Dreyfus and Dreyfus 1987, Berry and Broadbent 1988, Reber 1989, Sun 1999).

5. The Function of Consciousness

Another question which must be addressed is the functional or behavioural value of consciousness. One suggestion is the veto view of Libet (1985), or a more general counterbalance view (Kelley and Jacoby 1993). However, in these views, the reason unconsciously initiated actions need counterbalance, whether in the form of occasional veto or some other form, is not addressed (Sun 1999). Reber (1989), Stanley, Mathews, Buss, R. and Kotler-Cope (1989), and others have proposed that conscious and unconscious processes are suitable for different situations. **T**he earlier language/planning view of Crick and Koch (1990) suggests that consciousness enables the use of language and explicit planning. This view of theirs does not indicate the advantages over language and planning generated by unconscious processes. Jaynes (1976) suggested that conscious processing generates better behaviour in very complex social situations requiring consideration of past experiences of many different and unique individuals. Interestingly, Jaynes suggested that conscious processing makes use of mechanisms originally established to support speech. The relationship between language and conscious processing has been extensively debated. Carruthers (2003) has reviewed a number of alternatives, from views that language is conceptually necessary for conscious thinking to views that language is a support structure for the construction of human thought processes.

Sun (1994, 1999) suggested that conscious and unconscious processes are based upon different representations of the external world, and that the two processes are therefore

alternative approaches to generating behaviour. In this view there is synergy between the two approaches (Sun 1994, 1995, 1999). As indicated by psychological data (e.g., on implicit learning and implicit memory), 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). Hence there can be synergy between the two types of process.

There is psychological evidence in favour of the synergy view, (Willingham et al. 1989; Stanley et al, 1989). Synergy has also been demonstrated through computer simulation, in the domains of commonsense reasoning (Sun 1994, 1995) and in skill learning of tasks similar to the cognitive experimental tasks (Sun 1999; Sun et al. 2001).

Conscious processes also enables meta-level processes (Nelson 1993). Such meta-level processes and manipulation can include selection of reasoning methods, controlling the direction of reasoning, and evaluating its progress. Meta-level processes can be developed on top of meta-level processes, providing many levels of self-control of mental processes. Meta-level control of unconscious processes is more difficult, and the need for such processes provides another need for synergy.

At a deeper level of description (Sun 1999) has suggested that conscious processes are characterized by explicit (i.e., localist or symbolic) representations with explicit meta-level control. Unconscious processes are characterized by distributed representations without meta-level control. Synergy results from the interaction of the two different types of representations and therefore two types of processes for generating behaviour. This approach has been implemented in the CLARION system (Sun et al. 2000).

This synergy view incorporates some of the other views of the functional role of consciousness discussed earlier. Synergy would mean an effective veto on an unconsciously generated behaviour in some circumstances. Similarly, as in the situational difference view, in some cases it may be advantageous to use only the conscious or only the unconscious process. For example, if a task has been extensively practiced there is no longer a need for synergy, and unconscious processes alone are adequate. Conscious processes are then available for other tasks, as in the phenomenon of automatization. In the synergy view conscious language/planning is used on top of unconscious processes because of improved performance through the interaction of both types of processes.

Another alternative view of the function of conscious processes is the expansion of the range of information available to determine appropriate behavior in a given situation (Jaynes 1976; Baars 1988, Coward 1990). In this view, unconscious and conscious states are more of a continuum: as the range of information activated in response to a perceived object expands, the range of available alternative behaviours expands. To give a simple example, consider what happens when an object such as a tree is perceived. Direct sensory input generates some kind of mental activation which could be interpreted as alternative behavioral recommendations including to walk around the tree. However, if the range of activity is expanded to include information recorded in the past when similar activations were present, it could include memories of past experiences in which trees have been a factor, including information recorded about the activities during those experiences. This function might be of limited value in generating behavior in relatively simple situations such as in avoiding walking into the tree, but would be particularly

important for generating behavior in extremely complex situations or to generate complex verbal reports. Activation of information about a number of past situations with some resemblance to the current situation, including the behaviors adopted and their consequences, might make it possible to synthesize a more sophisticated behavior in the current situation. In this view speech perception operates by expanded activation of information (Coward 1999b) and therefore operates by a similar mechanism to consciousness. Baars (1997) suggests that consciousness is a major biological adaption which has acquired multiple functions in the general domain of developing appropriate behaviour in response to novel, challenging, and information rich events. Such functions include learning under very novel conditions, providing context information to supplement direct perceptual information, and various prioritization and decision making functions. The capability to expand the information available to support generation of behaviour beyond immediate perceptual information would support many of the more specific functions suggested by Baars. The expanded activation of information corresponds with the fringe of non-sensory experiences which in Mangan's (2001) argument surrounds a conscious experience.

At a deeper level of description a mechanism is required to activate the additional information most likely to be relevant for behaviour. One mechanism is to activate information which has often been active in the past at the same time as currently active information (Coward 1990). This past activity mechanism requires that all the units of recorded information which form the basis for declarative knowledge are conditions which have actually occurred in the past, but which are simply combinations of raw sensory inputs or of other conditions recorded without a priori guidance. Such conditions will therefore only correlate partially with cognitively significant features. These conditions are therefore functionally ambiguous in the sense that a cognitive feature would be indicated by the presence of any large subset of the set of conditions recorded when objects with the feature have been experienced in the past, and one condition could be part of the set for many features. In unconscious processing, only conditions actually present in current sensory inputs are activated. In conscious processing a population of additional conditions is activated. These additional conditions are those often present in the past when the already active conditions were also active. The expanded population will contain small subsets of the information which would be active in response to direct perception of a range of sensory objects which have often been present in the past at the same time as the current sensory object. This approach has been implemented in the recommendation architecture (Coward 2001).

The functional role of phenomenal consciousness has been regarded as a far more difficult question. The concept of 'qualia', referring to the 'phenomenal content' of conscious experience (Nagel 1974; Chambers 1993; Block 1995) presents difficulties for the view that the defining feature of a mental state is the set of causal relationships which it has with other states and with behaviour. Searle (1980) made the argument that a functional organization capable of generating behavior is not a necessary and sufficient condition for consciousness. It might then be argued that if cognitive functioning can occur without qualia, then qualia may not have a functional role. However, such a second logical step is not valid. There could be a range of different states at a detailed level which generate the same externally observed behavior, some of which correspond with phenomenal consciousness and others do not. For example, a wide range of different

functions can be implemented in an electronic system using general purpose microprocessors. With the same transistor technology, most such functions could be implemented to run faster with special purpose hardware. The fact that a function can be implemented without a microprocessor does not prove that microprocessors have no functional role.

A more urgent issue is the physical nature of phenomenal consciousness. As suggested in section 4, one approach is a greater structural and dynamic complexity to conscious processes. In the representational difference view, qualia might be accounted for by the totality of a multi-modal multi-level organization and its collective states. These total-states are of very high complexity because they involve a nexus of external perception in many modes: internal states, emotion, implicit and explicit memory, and so forth (Sun 1999). This nexus was termed the "manifold" by Van Gulick (1993), and the "superposed complex" by Lloyd (1995). In this approach, as argued by Sun (1999), complexity of organization may explain the difficulty of describing phenomenal experiences (i.e. qualia).

An analogous conclusion follows from the expanded ambiguous information view. The unconscious perception of, for example, the colour red, would correspond with activation of only those information combinations which had been almost invariably present when a red colour had been perceived in the past. As the activation is expanded, it includes more fragments of information derived from an increasing range of objects which happened to be red, or objects which happened to be present when red objects were also present in the past (Coward 1999b). Initially, this population could cause a verbal report of the initial object (i.e. red) but not in general a verbal report of other objects. However, as the active population expands, it may at some point contain enough information derived from one object type to activate yet more information present when such objects were present in the past, creating a conscious activation. Populations would be sensitive to the exact combination of recent experiences (i.e. the starting point for the definition of similarity), the degree to which the population was expanded (i.e. the degree of similarity to the starting point which was required for activation), and the profile of past experience (i.e. which objects had been present at similar times). The populations corresponding with conscious activations would therefore have a complexity which was not in general accessible to verbal report, and exhibit considerable variation between individuals and in response to apparently similar perceived conditions.

In both the synergy and the ambiguous information activation accounts, the complexity of organization is thus used explain the 'irreducibility' of phenomenal experience, and the difficulty of describing qualia. In both cases qualia are the result of the functional architecture of the system, and the activations which generate the subjective experience of qualia also serve useful functions (Nelson 1993. Sun 1999; Coward 1999b).

6. The Psychology of Consciousness

There has been extensive experimental work in the areas of perception, memory and skill acquisition (see for example Kirsner et al. 1998). This work has often been interpreted in terms of theoretical distinctions between implicit and explicit mental processes and between declarative and procedural knowledge, There has been criticism of these distinctions, but we believe that the level of description immediately below the general definition of consciousness and its functions should describe observed mental

processes in a causal fashion consistent with the causal relationships at the higher level. Although we use the implicit and explicit terminology, this use of experimental results is independent of theoretical interpretations of these results [7].

In general, explicit and implicit phenomena have been distinguished by the presence or absence of access consciousness (Schacter and Graf 1986). The distinction between implicit and explicit had one of its origins in the observation of memory deficits (for example corresponding with the kinds of memory tasks amnesiacs can and cannot perform; Dunn 1998). In amnesiacs such as patients with Korsakoff's syndrome (Butters 1984), the most obvious symptoms are the inability to create new declarative memories. However, there is an ability to learn skills even though there is no memory of the skill learning experience. For example, performance in solving the Tower of Hanoi problem steadily improves over a number of sessions, even though at the beginning of each session the patient has no memory of seeing the problem before (Cohen et al. 1985).

6.1 Implicit and explicit perception

In the dichotic listening experiments of Triesman (1960), subjects were presented with two sections of very different text, one to each ear. They were asked to repeat aloud (to shadow) the text heard by one specific ear only. Part way through the presentation, the two texts were switched between ears. The subjects tended to respond by (erroneously) switching to repeating the text from the other ear (i.e. the meaningful continuation). At the end of the presentation, some content could be recalled from the shadowed text but not from the other text.

In terms of the high level definitions of consciousness, the input of the shadowed text generated an activation corresponding with access consciousness. This activation generated verbal report (i.e. shadowing) and the recording of information which was accessible to later conscious activations. The input of the unshadowed text to the unselected ear generated an unconscious activation which could change the processing but could not generate verbal report or record information accessible to a subsequent conscious activation.

At this deeper level there are therefore a number of causal relationships. Firstly, sensory inputs can generate both a conscious and an unconscious activation at the same time. There may only be one conscious activation, but there could be multiple unconscious activations with respect to different sensory inputs from those generating the conscious activation. Secondly, only a conscious activation can result in recording of information which is accessible to a future conscious activation. Thirdly, an unconscious activation performs enough semantic or equivalent processing of its sensory inputs to generate a cognitively appropriate change to the source of inputs to the conscious activation.

These hypothesized causal relationships are consistent with the causal relationships at the higher, general consciousness level and add detail to those causal relationships.

---

[7] For example, implicit learning is a controversial topic. However, the existence of implicit processes is not in question, what is in question is their extent and importance (Stadler and Frensch 1998). If we allow for the possibility that both implicit and explicit processes coexist and interact with each other, we can move beyond the controversies that focused mainly on the details of implicit learning. Therefore in this paper we will not get into these controversies.

6.2 Implicit and explicit memory

Consider now the differences between explicit and implicit memory (Kirsner 1998). In laboratory testing of both normal and amnesiac subjects, two frequently used types of explicit memory testing are free recall and recognition. In free recall, subjects are presented with lists of words or other objects and in a subsequent test phase asked to recall the items on the list. Recognition testing is similar except that in the test phase subjects are supplied with items and asked if they occurred in the original list. A typical test of implicit memory is repetition priming. In one form of repetition priming, subjects are presented with letter strings which may or may not be English words. The subject is asked to classify each letter string as a genuine word or meaningless string, and the time taken to reach a decision measured. In the first phase of the experiment a series of strings are presented. In the second phase another series containing both strings from the first series and new strings are presented. It is found that the reaction time to classify repeated strings is shorter than for new strings.

There are a number of dissociations between explicit and implicit memory which indicate separate underlying mechanisms (Kirsner 1998). In explicit memory, unless subjects are asked to make old/new judgments immediately, accuracy is low and declines systematically as a function of time (Scarborough et al. 1977; Tulving et al. 1982). Explicit memory is associated with access consciousness, while implicit memory is not (Schacter and Graf 1986; Shimamura 1986). Performance in explicit memory testing is strongly affected by organization of the material to be remembered, but such organization has minimal effect on implicit memory (Schacter et al. 1989). Implicit memory is more strongly reduced by changes to the lexical form of words than explicit memory (Forbach et al. 1974).

Consider now the causal relationships implicit in these results. Visual perception of a letter string can generate a conscious activation. Information recorded in such an activation includes links between the word and the context in which the word appeared (i.e. on a list which included specific other words as part of a test). However, this information can only be accessed by a conscious activation and the ability to access the information declines rapidly with time. A later conscious activation generated by a combination of a word and a reminder of the context can test whether information linking word and context was recorded in the past, or the conscious activation generated by a reminder of the context can generate a conscious activation which can in turn generate speaking words which were present in that context. A visual perception of a letter string can generate an unconscious activation, and the activation can generate a simple verbal behaviour (e.g. saying yes or no).

6.3 Implicit and explicit skill learning

A concept used to describe cognitive skill acquisition is the distinction between declarative and procedural knowledge. Declarative knowledge is said to be knowledge of objects and events which is recorded during higher cognitive processes or consciousness and can be accessed by such processes. Procedural knowledge is said to be knowledge of how to perform a cognitive skill and is inaccessible to conscious processes (Anderson 1983).

Procedural knowledge can be highly efficient once it has been developed and can work independently without consciousness in many cases (e.g., Anderson 1983, 1993; Dreyfus

and Dreyfus, 1987). Cognitive skills can be acquired without acquisition of new declarative knowledge, as demonstrated by the earlier discussion of Korsakoff's syndrome patients learning to solve the Tower of Hanoi problem.

Declarative knowledge may be acquired later than procedural knowledge. In a dynamic control task, Stanley et al. (1989) found that the development of declarative knowledge paralleled but lagged behind the development of procedural knowledge. Even in high-level cognitive skill acquisition such as learning to design psychological experiments, VanLehn 1995) and Schraagen (1993) report that such learning often involves generalizing specific knowledge to form generic schemas in addition to specializing general knowledge to specific situations. There can be inconsistencies between procedural and declarative knowledge of a skill, as revealed by studies of specialist expertise. "In some areas of expertise, there is a dissociation between what experts say they do and what they do ..... [An expert verbal] description typically bears only a superficial relationship to the expertise" (Speelman 1998). Verbal descriptions by experts describing their expertise often correspond with beginner methods rather than actual methods, and requiring a verbal description can even result in the expert reverting to the less effective beginner method (Bainbridge 1977; Berry 1987).

However, declarative knowledge is also advantageous in many situations. Declarative knowledge can speed up the learning process when constructed on-line during skill learning (Mathews et al. 1989, Sun et al. 1996), declarative knowledge can facilitate the transfer of a skill (Willingham et al. 1989, Sun and Peterson 1999) by speeding up learning in new settings, and declarative knowledge can help in the communication of knowledge and skills to others.

There are therefore a number of causal relationships that can be hypothesized in explicit and implicit learning. Conscious activations can record and access declarative knowledge, but unconscious activations cannot. Both conscious and unconscious activations can generate behaviour and lead to procedural knowledge. However, conscious activations may generate behaviours which are inconsistent with those generated by unconscious activations. A conscious activation will often be less effective in generating skilled behaviour than an unconscious activation.

7. Architectural Models of Consciousness

At the next level of descriptive detail, separation of the cognitive system into subsystems is required. The separation must be such that the causal relationships between processes in the different subsystems correspond with the causal relationships at higher levels. The argument was made earlier that there are reasons for the existence of a functional modular hierarchy in the brain which provides the appropriate separation. In such a hierarchy, information exchange between modules is required to coordinate their different functions, but in order to make functional change without side effects feasible, the separation into modules must be such that this information exchange is minimized overall as far as possible (Coward 2001). This information exchange minimization requirement may conflict with the otherwise desirable objective that subsystems correspond exactly with different major cognitive functionalities.

Many architectural models have aimed at a two subsystem separation which corresponds with the distinction between implicit and explicit mental processes. The differences between the two subsystems in such architectural models can be in knowledge

organization, in knowledge content, in knowledge representation, or in knowledge processing.

For instance, Anderson (1993) proposed in his ACT-R model that there are two types of knowledge: declarative knowledge is represented by semantic networks, and is consciously accessible; procedural knowledge is represented by production rules, and is inaccessible. The difference thus lies in the different way of organizing knowledge: whether knowledge is organized in an action-centered way (procedural knowledge) or in an action independent way (declarative knowledge). Both types of knowledge are implemented symbolically (using either symbolic semantic networks or symbolic production rules). The model has difficulty with the qualitative phenomenological differences between the conscious and the unconscious (e.g. in terms of conscious accessibility). Although the knowledge organization is apparently different between semantic networks and production rules (with different degrees of action-centeredness), the difference appears insufficient to account for the qualitative phenomenological difference, since both are symbolically represented (along with numerical measures) and fundamentally the same. As Lebiere, Wallach and Taatgen [1998] have acknowledged in their work on implicit and explicit learning in ACT-R, "One of the defining properties of implicit learning, the fact that it is not a conscious process, is harder to operationalize….. The closest you can get … is the notion that implicit learning is not guided by learning intentions, but is rather a by-product of normal processing".

Hunt and Lansman's (1986) model is almost the exact opposite of Anderson's model, although the emphasis is on knowledge access (as opposed to knowledge organization). In their model, 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. Despite the different emphases, the issue with this model is the same, the differences between conscious and unconscious processes in the model do not on their own level and in their own terms generate the necessary causal relationships of conscious and unconscious processes. (A number of other views had the same issue, such as Bower 1996, Logan 1988, and so on).

There have also been various proposals in neurobiology that there are different processing pathways in the brain, some of which lead to conscious awareness while others do not. For example, Milner and Goodale (1995), Damasio et al (1990), and LeDoux (1992) proposed various versions of this view. Likewise, Schacter (1990) and Revonsuo (1993) suggested, based on neuropsychological data, that multiple modules coexist in the brain, each of which performs specialized processing (without incurring conscious awareness), with the exception of one module that is solely responsible for conscious awareness. Each of the specialized modules sends its output to the conscious module and thus makes the output consciously accessible. The issue of these biologically motivated two-system views is that, although there is ample biological evidence that indicates the existence of multiple pathways (in visual, language, and other processing modes), some of which are correlated with conscious awareness while some others are not, it is less clear why some result in consciousness while others do not, that is, what is different, mechanistically or computationally, between these different pathways.

Yet another two system view is based on the representational difference. As proposed in Sun (1994, 1995, 1999), different representational forms (in different subsystems)

may be used to explain the qualitative phenomenological difference between the conscious and the unconscious. According to connectionist theorizing (Sun 1995),in localist (or symbolic) representation, one distinct entity (e.g., a node in a connectionist model) represents a concept. Therefore, the representation is easily accessible. In distributed representation, a non-exclusive set of entities (e.g., a set of nodes in a connectionist network) are used for representing one concept and the representations of different concepts overlap each other; that is, a concept is represented as a pattern of activations over a set of entities (a set of nodes). Therefore, the representation is not easily accessible (relatively speaking). The mechanistic difference in accessibility between the two types of representation accounts for the phenomenological difference in accessibility between conscious and unconscious. To put this in a more simplistic way, verbal reports (speech) can be generated more easily from a localist representation. This view has been implemented in the CLARION model, which has two components using localist and distributed representations respectively (see Sun et al. 2000).

In the recommendation architecture (RA) model (Coward 1990, 1999a, 2000), an exact correspondence between the two subsystems and implicit and explicit mental processes is not present. Knowledge content, representation and processing differ between the two subsystems. In one subsystem, called clustering, knowledge is recorded about individual perceptual events. This knowledge is functionally (i.e. behaviourally) ambiguous. The behavioural ambiguity limits undesirable behavioural side effects of recording new information (Coward 2001). In the other subsystem, called competition, consequence feedback is used to create and record knowledge of the behavioural implications of the perceptual knowledge recorded in clustering. In this model, both explicit and implicit processes require knowledge recorded in both subsystems. The difference is that a much wider range of information is activated in conscious processes.

There have also been proposals that different processing modes in the same system give rise to conscious and unconscious processing (Dennett 1991). For example, Baars (1988) suggested that some sort of coherence in the activities of the brain gives rise to consciousness. Baars (1988, 1997) developed the classical theatre metaphor for consciousness into a specific theory. In this global workspace theory there are many processes active within a working memory "stage", and an attention "spotlight" highlights one of these processes. The highlighted process is the current content of consciousness. Working memory processes compete to enter the "spotlight", and information generated by the successful process is provided to a wide range of unconscious processes for learning and behaviour generation. Franklin and his collaborators (Franklin and Graesser 1999) have implemented a system with a global workspace architecture, and demonstrated that the system generates a wide range of behaviour analogous with consciousness. Mathis and Mozer (1996) proposed that being in a stable attractor of a neural network leads to consciousness. Crick and Koch (1990) suggested that synchronous firing of neurons in the gamma band of the EEG leads to conscious awareness. Damasio (1994) associated a reverberation of information flows in various cortical and sub-cortical areas with consciousness.

An attempt to establish a relationship between coherence and conscious is part of the theory proposed by O'Brien and Opie (1998). In this theory the brain is modeled as a large collection of neural networks. Any stable pattern of activation in one network corresponds with a symbolic representation, and any symbolic representation is

phenomenally conscious. Another attempt is that of Tononi and Edelman (1998), who argue that a neural activation corresponding with consciousness should interact more strongly internally than with the rest of the system, and should be able to give rise to a large repertoire of different activity patterns. Such properties would account for the observations that conscious experiences are integrated and the number of different conscious experiences which can be accessed over a short time is large.

The difficulty with these views is that despite many useful insights, there is no explanation why coherence (whether it is in the form of attractors, reverberation, or synchronous firing) leads to consciousness. In other words, it is not clear how the difference between coherent and incoherent states generates the causal relationships observed at the highest level of description of conscious phenomenology.

8 Criteria for an Effective Architectural Model

Given the objective of a mechanistic or scientific theory of consciousness defined as a consistent hierarchy of causal descriptions from cognitive to physiological levels, what are the essential requirements for an architectural model ?

Firstly, causal relationships at the definition and psychological levels must have corresponding causal relationships in the model. For example, the relationships that conscious processes can cause complex verbal reports but unconscious processes cannot must follow from the structure of the model.

Secondly, the qualitative differences between conscious and unconscious processes must be reflected in qualitative differences between corresponding processes in the model. However, it is not enough to identify differences in the model, it must be clear how the difference in the model processes causes different behavioural responses, such as verbal reports expressing the difficulty of describing conscious perceptions.

Thirdly, the major subsystems of the model should somehow correspond with major physical structures in the brain, and causal relationships between the subsystems must correspond with causal relationships between those physical structures. For example, the type of connectivity between subsystems and sequence of subsystem activation during a cognitive process must resemble that observed in major brain structures. This correspondence must also exist at more detailed levels, with model subsystems corresponding with modules observed in the brain, such as layers, areas and columns in the cortex and nuclei in subcortical structures. At a yet deeper level the causal relationships between devices in the model must correspond with causal relationships between neurons, although clearly detailed modeling at this level has not yet been achieved.

To illustrate the need for consistent causal relationships between psychological descriptions and models, consider the differences between three implemented models: the ACT model (Anderson 1993), CLARION (Sun et al. 2000), and the RA (Coward 2001). In the ACT model there are two types of knowledge, but both are symbolically represented (along with numerical measures). There is nothing in the model which demonstrates at the model level that conscious processes cause verbal reports but unconscious processes do not. In the case of CLARION, knowledge coded in local representations can easily generate speech but knowledge coded in distributed representations cannot easily do so. The difference in causal relationships at the psychological level is intrinsic to the model. In the RA, physical behavioural responses

can be generated by much smaller activated information populations than those required to generate complex verbal reports, although physical behaviours can also be generated by the larger activated populations. Again, the causal relationships at the psychological level are intrinsic to the model.

To illustrate the issue of the qualitative difference between conscious and unconscious perceptions, in ACT both types of activation are symbolic (along with numerical measures). In CLARION the type of representation is different, and in RA the type of activated information is different [8]. Thus in ACT there is no clear causal relationship between the different nature of conscious and unconscious activations and the type of verbal reports generated by conscious activations (e.g. "these activations have a richness which is hard to express in words") but in CLARION and RA there are such causal relationships.

Thus both CLARION and RA have potential to establish the modular description hierarchies required for a scientific theory of consciousness, while many alternative models have more serious deficiencies in this area. The consistency of these two models with the causal relationships found in implicit and explicit mental processes will next be examined in more detail.

9 Descriptions of CLARION and RA

As a first step, the architectures of the two models will be described in more detail.

9.1 CLARION model

The CLARION model has been implemented electronically and demonstrated a range of cognitive behaviours (Sun et al. 2001).

9.1.1 The high level architecture

The CLARION model contains three major components as illustrated in figure 1. The top level and the bottom level both produce behavioral recommendations in response to an input condition, and a separate component selects which behavior to carry out.

Both top and bottom levels receive the same sensory inputs, generate behavioral recommendations, and receive feedback on the consequences of accepted recommendations. The bottom level encodes procedural rules using distributed representations. Each module within the level corresponds with a behavior, and is generally uninterpretable in symbolic terms. Modules are evolved by consequence feedback (i.e. by reinforcement learning; e.g. Watkins 1989). The top level of the architecture encodes propositional rules using symbolic representations. Each unit or module within the level has a clear conceptual meaning/interpretation (i.e. a semantic label). Conceptual units are established by selecting a generalized input condition which resulted in generation of a successful behavior by the bottom level, and coding it as a propositional rule leading to the same behavior. Such propositional rules can then also generate behaviors and can be evolved independently of modules in the bottom level. The two levels thus in time generate independent behavioral recommendations, and the selection component selects a behavior from the range of alternatives generated by the two levels.

[8] More details later.

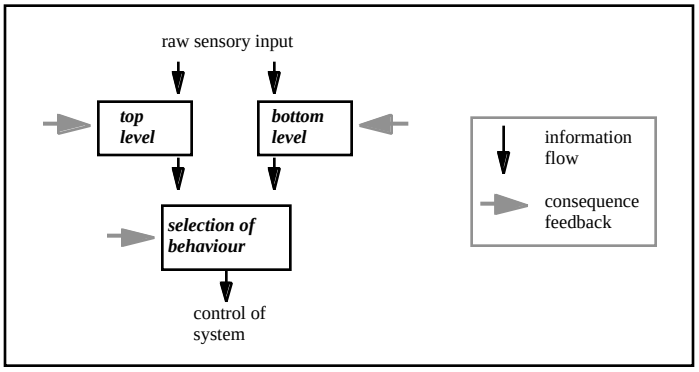

Figure 1 High level view of the CLARION architecture

9.1.2 The bottom level

Modules within the bottom level recommend alternative behaviors. In response to an input condition, the output from each module is a numeric value which can be interpreted as the relative value of the behavior corresponding with current input conditions. Modules are implemented as three layer backpropagation networks. The I/O mappings are acquired by reinforcement learning algorithms. Use of a reinforcement learning algorithm such as Q-learning (Watkins 1989) allows sequential behaviors to emerge. The system can learn to take into account future steps in longer and longer sequences. The bottom level is more sensitive to subtle or complicated forms of information. Processes operating at this level may be aptly described as associative, complex and fuzzy. The similarity of two items, defined as the similarity between their representational patterns, plays a major role, because connections between patterns direct the processing.

9.1.3 The top level

Modules within the top level consist of propositional rules. Each rule is implemented as a localist connectionist network. Rule learning occurs in response to input conditions. If some action decided by the bottom level is successful then the level extracts a rule that corresponds to the decision and adds the rule to the rule 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 (shrinking); if the outcome is successful, the agent may try to generalize the rule to make it more universal (expansion). If two rules give similar results, they are merged. If a rule needs to be shrunk but cannot be without making it unresponsive to all input conditions, it is deleted. So rules are in effect discretized and thus crisp/binary.

One input condition may correspond with multiple rules. In these circumstances, one rule is chosen from the matching set (by voting or randomly). It is possible to follow exact rules. Processes in the top level are more crisp and discrete, and thus more precise, reliable and selective. The top level allows explicit control and manipulation (due to explicit, localist representation) for deciding, altering and controlling reasoning methods. Each node has a clear conceptual meaning/interpretation (i.e. a semantic label).

9.1.4 Selection of behavior

The different characteristics of the two levels makes the combination of the outcomes from the two levels advantageous. This subsystem makes the final decision of which action to take by incorporating outputs from the two levels. The corresponding values for an action from the two levels are combined in a weighted sum. The top level indicates that its recommended action has an activation value, which is 0 or 1 because rules are binary, and the bottom level indicates that its recommended action has a numeric activation value. Stochastic decision making (with Boltzmann distribution) based on the weighted sums is then performed to select an action. Relative weights of the two levels can be automatically set based on the probability matching of the relative performance of the two levels (which is commonly observed in animal behavior). That is, if the success rate of the decisions made by the top level is sb and the success rate of the bottom level is st, then the weights are st / (sb + st) for the top level and sb / (sb + st) for the bottom level. At a particular moment (and in a particular task), whether the top level, the bottom level, or both are used may be altered by a number of factors (as known from the psychological literature) e.g. instructions, or situational demands such as complexity or multiplicity of tasks, and so on (Sun et al. 2001).

9.1.5 General properties of the model

As a preliminary to a more detailed comparison of the model with the psychological phenomena described in section 6, the processes in the model which correspond with four major psychological processes are described in general terms. These processes are: acquisition of a skill; recognition that a perceived object or condition is familiar (i.e. has been perceived before); reminiscence of a past object or condition; and generation of verbal reports describing a mental state.

In the CLARION model, skills can be acquired by two different types of learning modes: top-down learning and bottom-up learning. In top-down learning, top-level knowledge is learned first and then assimilated into the bottom level. In bottom-up learning explicit knowledge is extracted from implicit knowledge in the bottom level which is learned through trial and error first. Depending on the circumstances there might be mixed learning (both top-down and bottom-up), separate learning (learning within each level separately), or no learning at all. Recognition that an object or condition is familiar depends upon the existence of a large number of rules in the top level which correspond with the condition, and/or knowledge in the neural networks in the bottom level. Recognition that a new condition closely resembles a particular past condition can be performed, first of all, at the bottom level, which is similarity-based and associative, but also through the interaction of the two levels which leads to the explicit recognition of what a new condition is similar to by bottom-up partial activation of top level nodes. Reminiscence of a past object or condition could be implemented by some top level nodes recommending activation of other nodes even in the absence of the input condition for those nodes, and speech could be implemented by top level nodes recommending generation of speech elements.

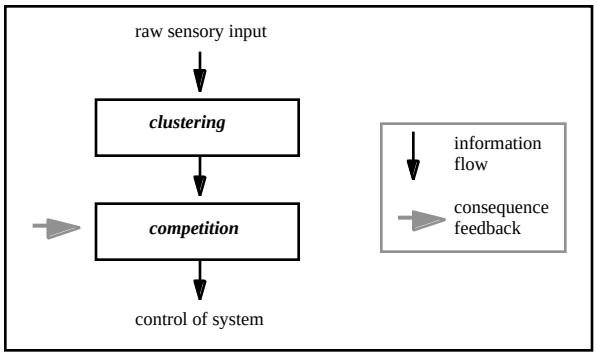

Figure 2 High level view of the recommendation architecture

9.2 Description of RA Model

The RA model has been implemented electronically and demonstrated a number of cognitive behaviours (Coward 2000; 2001).

9.2.1 The high level architecture

The RA has a major functional separation into two subsystems as illustrated in figure 2. Clustering receives a sequence of input states which are sets of relatively raw sensory inputs. The subsystem defines and records specific combinations of the sensory inputs present within one input state. A signal is generated indicating the first recording and any subsequent repetition of a combination. Combinations are defined within a number of different ranges of complexity, where the complexity of a combination is the number of sensory inputs which contribute to it. Combinations are defined until every input state contains some recorded combinations within every range of complexity. Detection of combinations within some ranges of complexity determines when and where additional combinations will be recorded. Detection of combinations within other ranges of complexity form the outputs to competition. Competition interprets different subsets of its inputs from clustering as different behavioural recommendations, and selects one behaviour. Consequence feedback to competition is used to enhance interpretations and selections, but consequence feedback is not used to directly change the combinations recorded by clustering.

9.2.2 The clustering subsystem

Clustering is organized in a sequence of layers. The first layer selects, records and detects repetition of specific combinations of raw inputs, and subsequent layers select, record and detect specific combinations of the combinations detected by the previous layer. Each layer therefore selects, records and detects combinations within a different range of combination complexity. A combination is selected once and recorded, and subsequent possible changes to the combination are limited to ensure that the different functional meanings which have been assigned to the presence of the original combination by different recipients are not excessively distorted.

In order to prevent excessive recording, sophisticated processes are required to manage when and where combination recording or change can occur. These processes are controlled by signals indicating the level of detection of combinations within modules. A primary function of a modular structure of columns and areas which is overlaid on the

layers is to provide this change management process. Some modules determine whether an adequate output from a layer is present and if so inhibit further recording in the layer. Some modules determine which other modules are the most appropriate location for recording and excite such recording. No behavioural knowledge is used in defining the combinations, so combinations will not correspond exactly with behavioural categories.

If the outputs to competition are significant but still inadequate to generate behaviour, there is an alternative to recording additional combinations. This alternative is to activate inactive combinations in early layers which were recorded when combinations currently active in later layers were also recorded, or were often active in the past when the currently active combinations were also active. This process introduces information derived from different input states which are similar to or were present at the same time as input states similar to the current input state. Invoking this process will in general be a behaviour which must be selected on the basis of current clustering outputs by the competition subsystem.

9.2.3 The competitive subsystem

The competitive subsystem is organized into modules corresponding with different behaviours. High level modules correspond with major types of behaviour (aggressive, avoidance, food seeking etc.), the most detailed modules correspond with sequences of motor commands. Outputs from clustering are provided first to all high level modules. There is a competition between these modules which selects one behaviour type. The effect of this selection is to release clustering outputs corresponding with the selected behaviour type to all the modules corresponding with more specific behaviours of the selected type. A competition between these more detailed modules selects one more specific behaviour. Further stages of competition may select even more specific behaviours within the general type selected.

The competition process adds the weights assigned in each module to each input reaching the module from clustering and from other competition modules, and selects the module with the largest input weight. Subsequent consequence feedback changes the weights of inputs into the modules corresponding with the selected behaviour type. This weight change modulates the probability of similar sets of clustering outputs generating similar behaviour in the future. Note that the absence of direct consequence feedback to clustering plus the separation of competition into behaviour modules means that consequence feedback in response to one behaviour does not affect past consequence learning by other behaviours.

9.2.4 General properties of the RA model

In the RA, declarative knowledge corresponds in a general sense with the recording of combinations within clustering, and procedural knowledge with the relative weights of inputs to competition modules. However, expressing declarative knowledge requires use of competition weights, and accessing the weights requires activation of combinations in clustering. Skill acquisition is assignment of behavioural weights to existing and/or new combinations detected by clustering. Recognition that a perceived object or condition is familiar is based upon the degree of new combination recording required to generate an output from clustering. In reminiscence, the competition subsystem interprets its current clustering outputs as a recommendation to activate additional relatively simple

combinations recorded in the past when currently active high complexity combinations were also recorded. These relatively simple combinations may in turn activate more complex combinations. The effect is to activate information recorded from objects similar to or present in the past at the same time as the currently perceived object. Enough information may be activated to generate a behaviour such as speaking the name of one such (remembered) object. Generation of speech requires activation of an adequate level of information derived from all the objects or conditions forming the subjects of speech.

10. Accounting for Causal Relationships at the Highest Level

In sections 4 and 5 a number of causal relationships were identified at the levels of the definition of consciousness and the functional role of consciousness. This sections evaluates how causal relationships within the models correspond with these high level relationships. In summary these causal relationships are firstly that sensory input can cause a conscious activation, an unconscious activation, or both. Secondly, an unconscious activation can cause physical behaviour but not complex verbal reports. Thirdly, a conscious activation can cause physical behaviour or complex verbal reports. Fourthly, a conscious activation can generate other conscious activations of associated objects including self. Finally, there is a qualitative difference between unconscious and conscious activations in response to the same perceived object, and between conscious activations in response to the same object at different times. The complexity of a conscious activation and the differences between different conscious activations in response to the same object can cause complex verbal reports, but the details of the complexity are difficult to describe, or in other words cannot cause complex verbal reports.

CLARION model:

In the CLARION model unconscious activations are activations in the bottom level, conscious activations are in the top level. Processes at either level can generate behaviour, but representations in the top level are more explicit and therefore more easily expressed in speech. The symbolic representations in the top level can readily generate other symbolic representations by associative inferential logic. The qualitative difference between local and distributed representations gives rise to the qualitative phenomenal differences between conscious and unconscious activations.

RA model

In the RA model, unconscious activations are activations within clustering which include only information combinations most similar to the currently perceived object. Conscious activations include this information plus a large population of somewhat less similar information which will typically include information combinations recorded in the past when objects with some degree of similarity to the currently perceived object were present or when other objects were present at the same time as the currently perceived object. However, the information derived from any other object will in general be a small subset of the information which would be activated in response to a direct perception of such an object. Both conscious and unconscious activations can generate physical behaviour via competition, but complex verbal reports require extensive activation of associated information. The conscious activation in response to a perceived object can

contain enough information combinations derived from other objects to cause further conscious activations which in turn generate verbal reports of those objects. The qualitative difference between unconscious and conscious activations derives from the much larger volume of associated information combinations. In a given conscious activation there will be large numbers of subsets of the activation which are also subsets of an activation which could generate naming behaviour for a different object, but which are too small to generate such behaviour alone. The presence of many such subsets which are unable to generate verbal behaviour corresponds with the perception of complexity which cannot be expressed verbally in conscious experiences.

11. Accounting for Causal Relationships at More Detailed Psychological Levels

It is not enough to account for a general difference between conscious and unconscious processes. A viable theory must demonstrate the capability to account for detailed psychological causal relationships. To illustrate this process, we will compare the ability of the two models to account for phenomena across the range of psychological processes discussed earlier.

11.1 Implicit and explicit perception

The causal relationships identified in section 6.1 are firstly that sensory inputs can generate both a conscious and an unconscious activation at the same time. An unconscious activation may be generated by a different set of sensory inputs from the conscious activation, and multiple unconscious activations may be possible generated by different sensory input sets. Only a conscious activation can record information accessible by subsequent conscious activations. An unconscious activation performs enough semantic or equivalent processing to generate a cognitively appropriate change to the source of inputs to the conscious activation.

CLARION model:

In CLARION, the bottom level can process information from different sources simultaneously, due to its modular structures. In order for information to be consciously accessible, however, it has to be represented at the top level. In the top level of CLARION, conceptual nodes perform meaning analysis and generate speech recommendations. Other nodes generate recommendations biasing the relative probability of acceptance of alternatives from the bottom levels, which processes all inputs but generates results which may not be accessible without the top level. Yet other nodes respond to external instructions. The shift of text shadowing from one ear to the other following meaning analysis rather than explicit instructions, as in the experiments of Triesman (1960), can be modeled as the top-level nodes responsible for meaning analysis overruling nodes for following external instructions, which results in only information of consistent continuity, after processing by the bottom level, reaching the top level nodes.

RA model:

In the RA model, multiple unconscious activations correspond with multiple independent populations of activated information combinations within clustering. In each population all activated combinations are present in the current sensory input from one domain. The extent of the conscious activation means that in general only one such

population can be supported at a time. The extent of the conscious activation also means that any information recorded in clustering could include information derived from a wide range of different earlier perceptual conditions. Hence some of the recorded information could be activated by a later conscious activation generated by a repetition of one of those conditions. However, an unconscious activation can only include information derived strictly from the perceived condition. All unconscious activations generate outputs from clustering which can be interpreted by competition as recommendations to generate conscious activations. The source of sensory inputs to a conscious activation can therefore be switched immediately if the appropriate recommendation is accepted.

11.2 Implicit and Explicit memory

The causal relationships identified in section 6.2 are firstly that perception of a letter string can generate an unconscious activation, and the unconscious activation can generate a simple verbal behaviour. Information linking the unconscious activation to the behaviour makes subsequent behaviour generation more rapid. Secondly, perception of a word can also generate a conscious activation. Information is recorded which includes links between the word and the context in which the word was perceived. The ability to access this information declines rapidly with time. Thirdly, the conscious activation generated by a combination of a word and a reminder of a context can test whether information linking word and context was recorded in the past. Alternatively, a reminder of the context can activate information recorded when the context was present which may contain enough information to generate verbal reports of words perceived in that context.

CLARION model:

In CLARION, explicit memory is in general driven by processes at the top level, while implicit memory is driven by bottom level processes. The faster time decay for explicit memory results from a higher decay parameter in the top level. Because of the separation between the levels, only a conscious activation can access information recorded during a previous conscious activation. The symbolic organization of information in the top level makes if effective for associative activations.

RA model

The increase in speed due to unconscious priming is a result of higher input weights in competition for the specific combinations within clustering which are present in the letter string. The combinations do not change, and the weights could only change if a similar behaviour under similar conditions occurred .

An information combination recorded during a conscious activation in response to a word may include information from the context in which the word is perceived, from the word, and from other words on the list immediately preceding the word. A subsequent activation of information derived from the context may contain enough information to generate a secondary activation containing much more information about a word on the list. The probability of subsequent associative activation of information within clustering is determined by the frequency of recent simultaneous activation, and the ability to activate information recorded during the test will therefore decline with time.

11.3 Implicit and explicit skill learning

The primary causal relationship identified in section 6.3 is that it is possible to record two types of knowledge, often labeled declarative and procedural, which are not always consistent with each other. Conscious activations can access declarative knowledge, but unconscious activations cannot. Both conscious and unconscious activations can generate physical behaviour and record procedural knowledge, but procedural knowledge is not directly accessible to conscious activations. Conscious and unconscious activations may generate inconsistent behaviours, in which case the unconscious activation is generally more consistent with current procedural knowledge.

CLARION model:

The CLARION model was originally developed in the context of skill learning, and allows learning of skills of varying complexity. It provides a model for initial implicit learning shifting to explicit learning with a bottom-up learning mechanism that extracts explicit rules from neural networks (Sun et al. 2001), and vice versa. Inconsistencies between explicit and implicit knowledge could arise through learning differences between the two levels (that is, reinforcement learning vs. rule learning). It provided a process by which declarative knowledge can accelerate the acquisition of implicit knowledge and vice versa through simultaneous use of two sets of learning and performance mechanisms (Sun et al. 2001).

RA model

Unconscious activations in clustering include only information combinations actually present in the currently perceived sensory condition. Skill learning has associated competition weights in favour of skilled behaviour with these combinations. In order to generate a complex verbal report, the activation must be expanded to include a much larger population of combinations which are not present in the current sensory condition. The combinations in this larger population will not have been present as often during skill learning, and the weights of these combinations into different possible behaviours will therefore not have been as well adjusted to generate the most appropriate skilled behaviour. A conscious activation may therefore result in a different, less skilled behaviour.

For a highly skilled behaviour, a conscious activation can therefore only be present if the skill is not operating at its highest level. Furthermore, conscious activations cannot access relative weight information directly because the weights are not represented as combinations which can be activated. Complex verbal reports can therefore only describe less skilled versions of behaviour.

Early learning can be accelerated by conscious activations, because expanded activations can help identify and record the types of combination most likely to provide discrimination between conditions in which different skilled behaviours are required.

12 Correspondence with Physiology

The issues here are whether a model maps into known physiology. If so, firstly there should be a functional correspondence between the major subsystems of the model and major neural structures such as cortex and subcortical structures. When there are activations in a model subsystem which cause activations in another subsystem there

should be a causal relationship between activations in the corresponding neural structures. Secondly, there should be functional correspondence between more detailed subsystems of the model and the substructures of major neural structures such as layers, areas and columns in the cortex and nuclei in the basal ganglia and thalamus. Thirdly there should be functional correspondence between devices used in the model and physiological neurons.

CLARION model:

The central role of modularity in motor control learning has been argued from the neurophysiological standpoint (e.g., Wolpert and Kawato 1998). In relation to the methods of modularization used in CLARION, it has been argued that such modules reside mainly in the supplementary motor area, while the selection of modules is likely done in the basal ganglia (see e.g. Bapi and Doya 1999). However, Houk et al (1995) suggested that action modules resided in the basal ganglia in the form of matrix regions. The learning (both the updating of modules and the updating of the module selection mechanism) is likely to be controlled by processes in the pre-supplementary motor area (Nakamura et al. 1997).

There has been indications that in the human brain, the temporal lobe is often correlated with consciously controlled sequence learning (Keele et al. 1998). In implicit memory research, however, it is found that explicit recall of sequences is often correlated with frontal lobe activities (Posner et al. 1997). Thus, it may well be the case that the temporal lobe and/or the frontal lobe are both responsible for the explicit processing corresponding to the top level of CLARION. The bottom level of the model, responsible for implicit processing, is likely distributed across a wide range of brain regions including the pre-supplementary motor area, the supplementary motor area, and the motor cortex (Nakahara et al. 1997).

RA model

The RA model requires a major separation between clustering and competition. Clustering is a modular hierarchy in which devices that permanently record conditions are organized into layers, columns and areas which group conditions by degree of similarity in such a way that detection of conditions can be used both to recommend behaviours and to determine when and where additional conditions will be recorded. Major clustering subsystems generate attention focus, general behavioural, and specific behavioural recommendations. Permanent recording of conditions creates the need for a resource management function, and damage to this function would result in a specific range of behavioural and memory deficits. Ensuring minimization of information exchange within clustering requires a management process which includes a periodic rerun of selections of past experiences. Competition is separated into three components which select the currently most appropriate attention focus, general behaviour type, and specific behaviour within the general type from the alternatives provided by clustering.

The cortex separation into sensory, associative and motor subsystems and organization into layers, areas and columns (Calvin 1995) resembles the clustering architecture. The thalamus, basal ganglia, and cerebellum subcortical structures have functions (Kingsley 2000) resembling those of the competition subsystems. Damage to the hippocampus results in deficits strongly resembling those produced by damage to the resource

management subsystem (Coward 1990). The requirement that cortex neurons select and permanently record specific combinations of inputs requires experimental test.

13. Discussion

It has been argued that a scientific theory of consciousness requires a hierarchy of consistent causal descriptions of phenomena on many levels from the level of phenomenal consciousness to psychological and physiological levels, in which the deeper levels have lower information density but higher descriptive complexity. A consistent description at one level alone is necessary but not sufficient to demonstrate theoretical viability. It has been further argued that biological brains are likely to have been constrained into simple functional architectures (successive levels of isomorphic descriptions) by the needs to be constructed, to recover from damage, and to learn without undesirable side effects, and that the hierarchy of descriptions implicit in such architectures can be the basis for such a scientific theory.

The outlines of such a theory at high levels has been constructed with causal relationships for access consciousness as the highest level description, and such relationships within perception, memory, and learning as the next level. The framework for causal relationships at the physiological and neural levels has been briefly discussed. Causal relationships which could be a mechanistic basis for phenomenal consciousness have been suggested, although much more detail remains to be worked out.

It has then been argued that many current cognitive models may not have the capability to support these causal relationships, and may therefore be inadequate as the basis for a scientific theory. Two current models have been demonstrated to have some capabilities of this type: CLARION and RA. The capabilities of these models to account for the observed causal relationships were then compared.

The comparison shows that both models can account for phenomena on multiple levels, but the model architectures and approaches are radically different.

In CLARION, the perceived complexity of phenomenal experience and the difficulty of verbal reports of that complexity are due to the complexity of the multimodal activation states corresponding with conscious experience. In RA the complexity and verbal description difficulty are generated by the population of fragmentary associated information which is activated in conscious experiences around the core of directly perceived information.

In CLARION, the observed differences between implicit and explicit perception, memory, and skill learning are generated by the location of symbolic and distributed information, behaviour and learning in separate subsystems. In RA there is no internal symbolic information. The different processes are generated by a combination of differences between conscious and unconscious activations and differences in the way information is recorded in two separate subsystems.

In CLARION there is a major separation between top level, bottom level and behaviour selection subsystems. The top level can provisionally be identified with temporal and frontal lobes, the bottom level with various motor areas, and behaviour selection with the basal ganglia. In RA, the major separation between clustering and competition can be identified with the separation between cortex and subcortical structures such as the thalamus, basal ganglia, and cerebellum. Clustering requires layers, columns and areas corresponding with those observed in the cortex.

At the device level, CLARION is a connectionist theory. Devices receive inputs which are outputs from other devices and generate outputs if the total input weight exceeds a threshold. Learning occurs by modification of input weights. In RA, similar device algorithms are used in competition, but in clustering the primary learning algorithm is that a device records specific active input combinations when appropriate change management signals are present, and generates an output in response to any repetition of a recorded combination.

CLARION thus has the advantage of greater conceptual simplicity and explanatory parsimony. RA has the advantage of a stronger potential for consistency with neural substructures such as cortex layers, areas and columns. Further work is required to create different predictions by the two models for the same cognitive or physiological phenomena, which would make it possible to compare such models based on detailed experimental results.

This paper has developed a number of suggestions of possibilities for theories of consciousness, based on theoretical and experimental work on computational models. Much more detail on these suggestions remains to be worked out, but we believe that these suggestions will provoke further thoughts and experimental work on various aspects of consciousness.

14. Conclusions

Firstly, a scientific theory of consciousness requires construction of a hierarchy of consistent causal descriptions from physiology through a series of intermediate levels to conscious phenomena. It is inadequate to only look for neural correlates of consciousness or to model cognitive data without reference to physiological plausibility or phenomenological analysis. Secondly, although entire conscious processes could in principle be described end-to-end in detail in terms of the activities of large populations of neurons, such descriptions would not be comprehensible to a human intelligence. Scientific understanding depends upon the selection of key elements of conscious phenomena and the creation of intermediate models for such elements. Thirdly, the two example theories, CLARION and RA, demonstrate that such a research program is feasible, and have properties which may become parts of an eventual full scientific theory.

Acknowledgements

Ron Sun acknowledges the support provided in part by ONR grant N00014-95-1-0440 and ARI grant DASW01-00-K-0012.

References

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

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

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

Baars, B. (1997). In the Theatre of Consciousness. *Journal of Consciousness Studies*, 4, 4, 292 - 309.

Bainbridge, L. (1977). Verbal reports as evidence of the process operator's knowledge. International *Journal of Man-Machine Studies*, 11, 411-436.

Bapi, R. and Doya, K. (1999). MFM: Multiple forward model architecture for sequence learning. In: L. Giles and R. Sun, (eds.) *Proceedings of the IJCAI'99 Workshop on Neural, Symbolic, and Reinforcement Methods for Sequence Learning*.

Bass, L., Clements, P. and Kazman, R. (1998). *Software Architecture in Practice*. Addison-Wesley.

Bechtel, W. and Richardson, R. C. (1993). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Princeton: Princeton University Press.

Berry, D.C. (1987). The problem of implicit knowledge. *Expert Systems*, 4, 144-151.

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

Block, N. (1995). On a confusion about a function of consciousness. *Brain and Behavioral Sciences* 18, 227 - 287.

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

Braddon-Mitchell, D. and Jackson, F. 1996: *Philosophy of Mind and Cognition*. Oxford, Blackwell.

Butters, N. (1984). Alcoholic Korsakoff's Syndrome: An Update. *Seminars in Neurology* 4, 2, 226-244.

Calvin, W.H., (1995). Cortical Columns, modules, and Hebbian cell assemblies. In M.A. Arbib, Ed., *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA: Bradford Books/MIT Press.

Carruthers, P. (2003). The Cognitive Functions of Language. *Behavioural and Brain Sciences*, 26, in press.

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

Cohen, N.J., Eichenbaum, H., Deacedo, B.S., and Corkin, S. (1985). Different memory systems underlying acquisition of procedural and declarative knowledge. In D.S. Olton, E. Gamzu, and S. Corkin (Eds.), *Memory dysfunction: an integration of animal and human research from preclinical and clinical perspectives* , 54-71. New York: New York Academy of Sciences.

Coward, L. A. (1990). *Pattern Thinking.* New York: Praeger.

Coward, L.A. (1999a). A physiologically based approach to consciousness, *New Ideas in Psychology*, 17, 3, 271-290.

Coward, L.A. (1999b). A physiologically based theory of consciousness, in Jordan, S. (ed.), *Modeling Consciousness Across the Disciplines*, 113-178, Maryland: UPA.

Coward, L.A. (2000). A Functional Architecture Approach to Neural Systems. *International Journal of Systems Research and Information Systems*, 9, 69 - 120.

Coward, L.A. (2001). The Recommendation Architecture: lessons from the design of large scale electronic systems for cognitive science. *Journal of Cognitive Systems Research*, 2, 2, 111-156

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

Damasio, A. et al (1990). Neural Regionalization of Knowledge Access. In *Cold Spring Harbour Symposium on Quantitative Biology LV The Brain*. CSHL Press.

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

Dennett, D. (2001). *Are We Explaining ConsciousnessYet ?* Cognition 79, 221 - 237.

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

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

Dunn, J. (1998). Implicit Memory and Amnesia, in Kirsner, K., Speelman, C., Maybery, M., O'Brien-Malone, A., Andersen, M., and MacLeod, C. (Eds). *Implicit and Explicit Mental Processes*, 99-117. New Jersey: Erlbaum.

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

Forbach, G.B., Stanners, R.F., and Hochhaus, L. (1974). Repetition and practice effects in a lexical decision task. *Memory and Cognition*, 2, 337-339.

Franklin, S. and Graesser, A. (1999). A Software Agent Model of Consciousness. *Consciousness and Cognition* 8, 285 - 305.

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

Goguen, J. A. (1977). Complexity of Hierarchically Organized Systems and the Structure of Musical Experiences. *International Journal of General Systems* 3, 233 - 251.

Greene, G. (1999). The Elegant Universe. Norton.

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

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

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

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

Houk, J. Adams, J. and Barto, A. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. Houk, J. Davis, and D. Beiser, (eds.) *Models of Information Processing in the Basal Ganglia*. MIT Press, Cambridge, MA.

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

Jaynes, J (1976). *The Origin of Consciousness in the Breakdown of the Bicameral Mind*. Boston: Harvard.

Kamel, R. (1987). Effect of Modularity on System Evolution. *IEEE Software*, January 1987, 48 - 54

Karmiloff-Smith, A. (1992). Beyond Modularity. MIT Press.

Keele, S., Ivry, R., Hazeltine, E., Mayr, U. and Heuer, H. (1998). *The cognitive and neural architecture of sequence representation*. Technical report No.98-03, Institute of Cognitive and Decision Sciences. University of Oregon.

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

Kim, J. (1993) *Supervenience and Mind*. Cambridge University Press

Kingsley, R.E. (2000). *Concise text of neuroscience*. Baltimore MA: Lippincott, Williams and Wilkins.

Kirsner, K. (1998). Implicit Memory, in Kirsner, K., Speelman, C., Maybery, M., O'Brien-Malone, A., Andersen, M., and MacLeod, C. (eds), *Implicit and Explicit Mental Processes*, 13-36, New Jersey: Erlbaum.

Kirsner, K., Speelman, C., Maybery, M., O'Brien-Malone, A., Andersen, M., and MacLeod, C. (1998). *Implicit and Explicit Mental Processes*. New Jersey: Erlbaum.

Lakatos, I. (1970). Falsification and Methodology of Research Programs. I. Lakatos and A. Musgrave (Eds.) *Criticism and the Growth of Knowledge*, Cambridge: Cambridge University Press.

Lebiere, C., Wallach, D., and Taatgen, N. A. (1998). Implicit and Explicit Learning in ACT-R. Proceedings of the Second European Conference on Cognitive Modeling.

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

Leslie, A. (1994). ToMM, ToBY and Agency: Core architecture and domain specificity. In *Mapping the Mind*, L. Hirschfeld and S. Gelman, eds. Cambridge University Press.

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

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

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

Machamer, P., Darden, L. and Craver, C. (2000). Thinking About Mechanisms. *Philosophy of Science* 67, 1 - 25.

Mangan, B. (2001). Sensation's Ghost: The Non-Sensory "Fringe" of Consciousness. *Psyche*, 7, 18.

Marr, D. (1982). *Vision.* W.H. Freeman: New York.

Mathews, R. Buss, R. Stanley, W. Blanchard-Fields, F. Cho, J. and Druhan, B. (1989). Role of implicit and explicit processes in learning from examples: a synergistic effect. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 15, 1083-1100.

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

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

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

Nakahara, H., Doya, K., Hikosaka, L. and Nagano, S. (1997). Reinforcement learning with multiple representations in the basal ganglia loops for sequential motor control. *International Joint Conference on Neural Networks*, 1553-1558.

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

Nortel Networks (2001). DMS-100/500 Feature Planning Guide. http://www.nortelnetworks.com/products/01/dms100w/doclib.html

O'Brien, G. and Opie, J. (1998). A Connectionist Theory of Phenomenal Experience. *Behavioural and Brain Sciences*, 22, 127-148.

Parnas, D. L. (1972). On the Criteria to be Used in Decomposing Systems into Modules. *Communications of the ACM*, 15, 12, 1053-1058.

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

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

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*. Kluwer, Dordrecht, Netherland.

Salmon, W. (1984). Scientific Explanation and the Causal Structure of the World. Princeton: Princeton University Press.

Scarborough, D.L., Cortese, C., and Scarborough, H.S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, 3, 1-17.

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

Schacter, D.L., Bowers, J., and Booker, J. (1989). Intention, awareness and implicit memory: The retrieval intentionality criterion. In S. Lewandowsky, K. Kirsner, and J. Dunn (Eds.). *Implicit Memory: Theoretical Issues*. Hillsdale NJ: Erlbaum.

Schacter, J.C. and Graf , P. (1986). Preserved learning in amnesic patients: Perspectives from research on direct priming. *Journal of Clinical and Experimental Neuropsychology*, 8, 727-743.

Schraagen, J. (1993). How experts solve a novel problem in experimental design. *Cognitive Science*, 17, 285-309.

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

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

Shimamura, A. (1986). Priming effects in amnesia: Evidence for a dissociable memory function. *Quarterly Journal of Experimental Psychology*, 38A, 619-644.

Speelman, C. (1998). Implicit Expertise: Do We Expect Too Much from Our Experts. In Kirsner, K., Speelman, C., Maybery, M., O'Brien-Malone, A., Andersen, M., and MacLeod, C. (Eds.), *Implicit and Explicit Mental Processes*. New Jersey: Erlbaum.

Stadler, M.A. and Frensch, P.A. (1998). *Handbook of Implicit Learning.* Sage Publications.

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

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

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

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

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

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

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

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

Tononi, G. and Edelman, G.M. (1998). Consciousness and Complexity. *Science* 282, 1846 - 1851.

Triesman, A.M. (1960). Contextual clues in selective listening, *Quarterly Journal of Experimental Psychology*, 12, 242-248.

Tulving, E., Schacter, D.L., and Stark, H.A. (1982). Priming effects in word-fragment completion are independent of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8, 4, 336-342.

Van Essen, D.C. and Andersen, C.H. (1995). Information processing strategies and pathways in the primate visual system. In *An Introduction to Neural and Electronic Networks*. Academic Press.

Van Gulick, R. (1993). Understanding the phenomenal mind. In: *Consciousness*. eds. M. Davies and G. Humphreys. Blackwell, Oxford, UK.

VanLehn, K. (1995). Cognitive skill acquisition. In: J. Spence, J. Darly, and D. Foss, (eds.) *Annual Review of Psychology*, 47. Annual Reviews Inc. Palo Alto, CA.

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

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

Wolpert, D. M. and Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, 11, 1317-1329.