Implications of Resource Limitations for a Conscious Machine

L. Andrew Coward and Tamas O. Gedeon
Department of Computer Science
Australian National University

Abstract
A machine with human like consciousness would be an extremely complex system. Prior work has demonstrated that the way in which information handling resources are organized (the resource architecture) in an extremely complex learning system is constrained within some specific bounds if the available resources are limited, and that there is evidence that the human brain has been constrained in this way. An architectural concept is developed for a conscious machine that is within the architectural bounds imposed by resource limitations. This architectural concept includes a resource driven architecture, a description of how conscious phenomena would be supported by information processes within that architecture, and a description of actual implementations of the key information processes. Other approaches to designing a conscious machine are reviewed. The conclusion is reached that although they could be capable of supporting human consciousness-like phenomena, they do not take into account the architectural bounds imposed by resource limitations. Systems implemented using these approaches to learn a full range of cognitive features including human like consciousness would therefore require more information handling resources, could have difficulty learning without severe interference with prior learning, and could require add-on subsystems to support some conscious phenomena that emerge naturally as consequences of a resource driven architecture.

Key words: consciousness; information model; system resource architecture; system design

Introduction

Theoretical arguments, supported by experience with the design of extremely complex electronic systems and by comparisons with the mammal brain, have demonstrated that any system which must perform a sufficiently complex combination of behavioural features with limited information handling resources will experience severe constraints on its architectural form [Coward 2001]. The implementation of a "conscious machine" with a full range of cognitive capabilities and the ability to bootstrap those capabilities by learning from experience is likely to require extensive information handling resources. The primary value of considering the design of a conscious machine is probably to achieve a deeper understanding of human consciousness [Aleksander 2005], and the value of such consideration is reduced if the impact of resource limitations on architecture is not taken into account.

There are many different ways in which a system to perform a given combination of features could be implemented. However, as the number and complexity of the required features increases, practical considerations place increasingly severe constraints on the architectural form of the system. For example, the resource driven system architecture for a complex electronic system has a separation between memory and processing subsystems, and communication busses that link them. Memory is separated into domains with different access times (e.g. cache, RAM, hard drive and CD memory). There may be a primary processor and various peripheral processors. Busses with various bandwidths connect different parts of the system. A hard wired instruction set and registers are defined which can efficiently perform all the information processes required by system applications.

Coward [2000, 2001] has argued that unless there are no limits to available information handling resources, any system that is designed under external intellectual control to perform a large number of different features tends to be constrained into this

architectural form. Any particular application or feature will call upon all or most of these different resources, and the same resources will be shared across many different applications and features. Functional architectures and user manuals are important to specify the features of the system from a user's point of view, by describing how the features perform. However, their feature descriptions do not discuss how system resources are organized to support the performance of the features. An attempt to implement the system features by using the functional architecture or user manual as the resource driven architecture (i.e. physical modules corresponding with features or functions) would result in impractical levels of resource requirements and other practical problems.

These conclusions [Coward 2001] are based on analysis of the architectural impacts of a number of practical considerations, including (1) the need to perform a large number of behavioural features with limits on the physical resources required for information recording, information processing and internal information communication; (2) the need to add and modify features without side effects on other features; (3) the need to protect the many different meanings of information generated by one part of the system and utilized for different purposes by each of a number of other parts of the system; (4) the need to maintain the association between results obtained by different parts of the system from a set of system inputs arriving at the same time (i.e. maintain synchronicity); (5) the need to limit the complexity of the system construction process; and (6) the need to recover from construction errors and subsequent physical failures or damage.

Any learning system will tend to require more resources than a system designed under external intellectual control to perform the same set of behaviours. Coward [2000, 2001] has argued that for a complex learning system, the same practical considerations will also tend to constrain architectural form. However, for a learning system the need to learn new and modified features rather than having such changes imposed under external intellectual control results in a qualitatively different set of architectural constraints called the recommendation architecture. There will tend to be a primary separation between two subsystems called clustering and competition. The clustering subsystem is a hierarchy of modules that defines and detects similarity circumstances in the sensory and other information available to the system. The competition subsystem interprets each detection of a similarity circumstance as a recommendation in favour of a range of different behaviours, each with a different weight, and determines and implements the behaviour most strongly recommended across all currently detected similarity circumstances.

Each module in clustering detects a different similarity circumstance. Such a similarity circumstance can also be viewed as a set of similar information conditions, and detection of the circumstance means that a high proportion of the set of conditions is present. Resource limitations imply that the management of different behaviours will need to share the same set of similarity circumstances, and the need to learn without interference with prior learning therefore implies that there will be severe restrictions on the ways in which the similarity circumstance of a module can change.

As described in Coward [2001], if the ratio of number of behaviours to available resources is high, new conditions can be added to the set for a module if they are fairly similar to other conditions already in the set, but with some limited exceptions conditions cannot be removed once added. As a result, modules cannot be evolved to correspond with cognitive features or categories. Such

modules are analogous with statistically defined independent components [Hyvärinen et al 1999], although unlike such components they are evolved continuously with experience.

These architectural constraints only apply if there are strong limits to information handling resources relative to the number of behavioural features that must be learned. However, Coward [2001; 2005a] has pointed out that natural selection will generate pressures on architectural form analogous with the practical considerations, and presents a wide range of evidence that the human brain has been constrained within these theoretical architectural limits.

In human beings, cognitive capabilities including consciousness are dependent on an immense amount of learning. For example, speech capabilities are often regarded as essential to human consciousness [e.g. Block 1995], and such capabilities require an extensive learning process. Information derived from sensory experience must be recorded and organized to support different kinds of representations of that experience. In addition, behaviours appropriate to representations corresponding with different sensory experiences etc. must be learned. Any external guidance of this learning has to be provided through sensory inputs influenced by a teacher, but such a teacher does not have access to the internal operations of the brain. A conscious machine given its capabilities largely by design and not requiring such a learning process might be possible, but might also have fewer lessons for understanding human consciousness. For example, as discussed later, some machine consciousness proposals limit their discussion of learning to creating associations between representations and behaviours, but do not address the issue of how an adequate set of representations can be bootstrapped from experience with very limited a priori knowledge and ongoing guidance. This approach will reduce the resources required, but the lessons for human consciousness which can be drawn from such an approach may be less reliable.

If a conscious machine is required to learn a high proportion of its capabilities from experience, and continue to learn and modify its own behaviours on an ongoing basis, the architectural limits on complex learning systems will be a reasonable guide to the definition of the architecture for such a fully featured conscious machine. However, it is important to note that given enough resources, a system with an architecture outside the constraints could learn the same range of behaviours. With enough resources, different types of representational learning or behaviour could be supported with separate resources without sharing. Such an organization of resources to correspond with externally visible functionality is analogous to an implementation of a user manual for a regular electronic system. Coward and Sun [2007] have argued that a number of consciousness models are of this “user manual” type. A small subset of the full range of conscious capabilities could be implemented with this approach, but problems would occur as such a system was scaled up to perform an increasingly wide range of capabilities. Eventually the available information handling resources would be inadequate, or severe problems with interference between earlier and later learning would develop. Hence a critical test of any proposal for a conscious machine is whether it includes a strategy by which a very wide range of capabilities can be bootstrapped from experience with limited, plausible a priori knowledge and ongoing guidance.

Furthermore, behavioural phenomena that emerge naturally within an architecture satisfying the constraints might need to be specially implemented (for example by additional subsystems) in another type of architecture. For example, in systems within the

recommendation architecture constraints, modules that are activated when their similarity circumstance is present in a specific sensory input state are likely to have behavioural recommendation strengths relevant to that input state. However, the severe restrictions on module changes mean that modules that have often been active in the past at the same time as such active modules, or modules that changed at the same time in the past as such modules may also have relevant recommendation strengths. Indirect activation of modules on the basis of such temporally correlated past activity may therefore have behavioural advantages which would not be present for modules in a different architecture which could change more freely with learning [Coward and Gedeon 2005]. One such advantage is support for autobiographic memory requiring recall of unique events on the basis of a single exposure [Coward 2005a, b]. As discussed below, such indirect activations are important for supporting conscious phenomena in the recommendation architecture approach.

Two initial steps in any major design project are *firstly* to specify the desired functions of the system and *secondly* to define an architectural concept, and taking these two steps within the architectural constraints imposed by resource limitations will be the focus of this paper.

The desired functions need to be described on a number of levels of detail. One level is a list of different system features and how they operate, described in a way that can easily be understood by a system user. This first level is essentially equivalent to a user manual. The second level describes the functions needed to support the system features, in terms of subsystems performing functions that are readily understood by an outside observer, with these functions combining in various ways to deliver individual features. This second level is the functional architecture, which in some ways is a more general user manual. When the desired functions of the system include "human consciousness", an important issue is the specification of what exactly is meant by that term. In this paper, several types of consciousness will be discussed including access consciousness, phenomenal consciousness, stream of consciousness and self consciousness. The term “conscious” is often not precisely defined, and to avoid the problems created by such weak specification the focus will be on specific examples of cognitive phenomena of these consciousness types, and how such specific examples could be supported within the architecture.

The architectural concept for a complex system has three elements. One element is a system architecture which separates the system into a number of physical modules, with different modules performing different types of information processes in a resource efficient fashion. The second element is a step by step description of how each of a representative range of features will be performed by a sequence of information processes supported by the modules. The third element is a demonstration that the key information processes can be implemented with the available technology. An adequate design concept implies that there is a high level of confidence that the system could be implemented, and in a commercial environment justifies investment of the substantial design resources needed to build a complete system. This paper aims to present a conscious machine architectural concept of this type.

The driving force for the system architecture is effective use of resources. A module is a set of physical resources that performs a group of similar information processes, where "similar" means that the processes can all be performed efficiently by the same physical

resources. The architecture therefore separates the available system resources into modules that can perform different types of information processes very efficiently. Because the primary emphasis is on resources, and system features or functions as defined by an outside observer will typically require information processes supported by many different modules, the relationship between modules and such features or functions will be very complex. Thus in a complex system there are some similarities between the user manual and the functional architecture, but radical differences between these descriptions and the system architecture that describes how the resources of the system are organized to deliver the features. In this paper it will be assumed that a resource conserving system architecture is required, as argued by Coward [2001].

The second element in the design concept is demonstration of the capability of the system architecture to support a representative selection of the required system features. This demonstration is qualitative but must be reasonably detailed in terms of the steps by which a feature operates. In the case of a conscious machine it must show that sequences of information processes supported by the modules of the architecture can combine to deliver a representative cross section of conscious phenomena.

The third element in the design concept is demonstration that the different required information processes can be implemented using available technology. This demonstration generally means an actual implementation of each critical information process.

An architectural concept must thus provide a framework in which resources are organized to deliver a set of information processes, a demonstration that each information process can be implemented (generally with an actual implementation), and a set of scenarios by which a representative sample of system features are performed using the information processes.

The approach in this paper will be *firstly* to define consciousness in terms of phenomena in physiological systems; *secondly* to provide an overview of the arguments in Coward [2001] that a range of practical considerations constrain the system architecture and information processes of any complex learning system whether biological or electronic into a recommendation architecture form; *thirdly* to summarize how physiology suggests that the human brain has been constrained within the recommendation architecture limits; *fourthly* to specify how conscious phenomena can be understood as information processes within this physiological architecture; and *fifthly* to review how the same architecture and information processes can be implemented in electronic form.

A range of alternative approaches to modelling consciousness will be discussed, leading to the conclusion that their architectural concepts do not adequately take into account the practical considerations that apply to any sufficiently complex system. Often an architecture is defined in which subsystems correspond with major system features or functions rather than types of information process specified in terms of the implementation technology. This type of architectural concept is effectively a user manual that describes how features work in a way that is easy for an outside observer to understand, or a functional architecture that describes how major groups of similar features work. If such architectures are treated as system architectures, they will be difficult to implement for a complete range of features without excessive resources and other practical problems.

Consciousness features are often defined generally in these alternative approaches, and not in terms of a reasonably detailed sequence of steps, and the issue of learning complex combinations of detailed features without severe interference between earlier and

later learning not addressed. It may well be possible to implement limited subsets of conscious features with such approaches, but in addition to practical problems, features which emerge naturally because of the form of the resource driven architecture in the human brain could need to be added as additional subsystems in a functional architecture type implementation.

Modelling Consciousness

As discussed in Coward and Sun [2004, 2007], the objective of a physiologically realistic model of consciousness is not to "understand" subjective individual experience in a philosophical sense. Rather, the objective is to model specific, observable phenomena labelled "conscious" in terms of physiological structures, in such a way that observed cognitive inputs (i.e. inputs described in cognitive terms such as perceptions) result in specified physiological states, and these physiological states cause other physiological states, and so on, to generate the observed high level cognitive outputs. This position follows the argument of William James [1904] that although consciousness is clearly a function in experience which thoughts perform, it is not an entity or quality of being which can be contrasted with the material world. A "conscious machine" will similarly need to model specific, observable conscious phenomena, but using electronic rather than physiological structures to implement the required information mechanisms.

There is, however, controversy over which phenomena should be labelled conscious. An important part of the definition of "consciousness" is a contrast between human behaviours that are conscious and those that are unconscious. As a simple example of this contrast, consider the phenomenon of dichotic listening [Treisman 1960]. This phenomenon occurs when different texts are presented verbally to the left and right ears of human subjects and the subjects are told to shadow (i.e. repeat) the text delivered to just one specific ear. If the texts are switched between ears during the shadowing, subjects switch ears to the meaningful continuation without being aware of having switched. Subjects have no memory of the unshadowed text. Thus although there is no memory of the text presented to the unshadowed ear, such text must be able to influence behaviour because it is able to trigger a switch. The argument is that the attended text enters consciousness, while the unattended text influences behaviour but does not enter consciousness. A physiological model needs to describe the physiological states that result from the presentation of the two texts and how those states generated other states, and demonstrate that those later states result in the ability to switch from one ear to the other when the texts switch between ears, but create future memories of only one text. It is of course possible that physiological models of this type may give insight into the issues of philosophical understanding.

In this paper, four aspects of consciousness phenomena will be discussed: access consciousness; phenomenal consciousness; stream of consciousness including imagining of non-existent circumstances; and self consciousness. Access consciousness and phenomenal consciousness were proposed by Block [1995] as major different types of consciousness. Access conscious is defined as the ability to report and act upon an experience and requires 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. The phenomenon of dichotic listening is an example of access consciousness compared with unconsciousness.

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. Part of the issue here is the difficulty in generating exact descriptions of such feelings, and the possibility that such feelings may be different for each individual. Stream of consciousness is the concept formally introduced by William James [1892], and will be defined as the ability to experience streams of mental images that have limited connection with current sensory inputs and may be of situations never actually experienced or even impossible in reality. Self consciousness refers to the capability to include self images in a meaningful way in streams of consciousness. These different types of consciousness are of course not fully independent.

There is no a priori guarantee that the many different phenomena which have been labelled "conscious" are supported by exactly the same physiological mechanisms. It could also be that it is the occurrence of a sequence of physiological processes in a particular order that results in "conscious" phenomena, but all of the same processes in a different order might not result in a "conscious" process. In this context, searches for "neural correlates of consciousness", or detectable physiological states that occur consistently at the same time as a wide range of conscious phenomena, may be somewhat simplistic.

As a starting point for creation of a physiologically realistic model, it is essential to have specific examples of these phenomena and other phenomena that are not "conscious" for comparison purposes. In the case of access consciousness, one such phenomenon is dichotic listening as described above. To provide phenomena for different types of consciousness, consider the following scenario (similar to that discussed in Coward and Sun [2004]). In the scenario, a person is out walking with a companion, and encounters a tree partially blocking the path. One behaviour is simple avoidance: stepping around the tree. A second behaviour is to make the comment "Mind the tree" to a companion. A third behaviour is to focus attention on the tree and "become aware" of the tree as a tree. A fourth behaviour is to experience a series of mental images initiated by the sensory experience of the tree, resulting in a comment "Up in the mountains I saw a whole area covered with trees like it. I wonder what it would be like to have a group of trees like that in my garden". The cognitive processes suggested by these scenarios could be visual input from a tree, unconscious activation, conscious activation, avoidance behaviour, and higher cognitive behaviour (including associative thinking, and verbal report of associative thinking).

The first behaviour can generally be unconscious. Sensory input from the tree and the path etc. generates some activation state internal to the brain that leads to avoidance behavior but does not lead to higher cognitive functions, verbal report, or in many cases even later memory. The second behaviour is generation of a simple comment and may not require complex cognitive processing. The internal brain activation state in response to the tree generates both motor behaviour and a simple verbal report.

The third behaviour appears to have some relationship with what is often called phenomenal consciousness. Becoming aware of the tree as a tree is subjectively a richer experience of the tree, the experience is difficult to express in speech and may be very individual specific. The cognitive phenomena resulting from this third behaviour are therefore verbalization of the presence of a richer experience that cannot be described verbally in detail, but may be very idiosyncratic.

The fourth behaviour is an example of a stream of consciousness including imagining of non-existent circumstances ("trees like that in my garden"), in this case leading to a specific verbal behaviour. As will be discussed later, it is possible that an initial step could be to become aware of the tree as a tree; this step could lead to simple comments, and/or be followed by a stream of consciousness resulting in more complex comments.

The objective of this paper is to indicate how an electronic system modelled on physiology could be given sensory inputs, and from the internal states generated by those inputs would in turn generate states corresponding with observed "conscious" cognitive outputs consistent with the internal states.

Architectural Constraints on Complex Learning Systems

Significant limits can be observed on the range of architectures exhibited by practical systems. It is striking, for example, that every electronic system above a certain level of behavioural complexity has the memory, processing, input/output and common communication bus structure. Both hardware and software in complex electronic systems tend to be organized into modules, with far more communication within modules than between them. It is also striking that structures including cortex, hippocampus, cerebellum, thalamus, hypothalamus and basal nuclei (or ganglia) appear ubiquitously in the mammal brain, across species with widely differing behavioural challenges. Analogous structures are visible in the avian brain despite 300 million years of independent evolution [see e.g. Shimizu and Bowers 1999].

As discussed in Coward [2000; 2001], the source of many of the architectural constraints on complex electronic systems is the group of practical considerations listed earlier. In the case of systems such as brains that must learn complex combinations of behaviours, there are analogous practical considerations [Coward 2001; 2005a]. Possible effects of some of these considerations on learning system architectures have been discussed by a number of authors. For example, the need to learn without interference with prior learning, a need analogous with feature modifiability in electronic systems, has been widely discussed as the "catastrophic interference" problem [McClelland et al 1995; Robins, 1995; French, 1999]. The need to maintain the association between results of system processing on inputs received from the same object at the same time has been widely discussed as the "binding problem" [Von der Malsburg, 1981; 1999; Shadlen et al 1995].

The specific architectural requirements that apply to complex learning systems have been labelled the recommendation architecture [Coward 2001], and in a very general sense this architecture is analogous with the memory, processing architecture for systems in which complex combinations of behaviours are defined under external intellectual control, although the two architectures are qualitatively different.

At the highest level, the architectural limits on complex learning systems require a separation between a modular hierarchy and a component hierarchy[1]. The modular hierarchy organizes information derived from experience by defining and recording conditions that occur within that experience, and detecting each condition when it is recorded and any subsequent repetition. A module defines a set of similar conditions and indicates the detection of any significant subset by producing an output. Components in the component hierarchy correspond with different behaviours, sequences of behaviours or types of behaviour. The component hierarchy receives module outputs, and interprets each such output as a recommendation in favour of a range of different behaviours, each recommendation having a weight. The component hierarchy performs a competitive process, which determines and implements the most strongly recommended behaviour at each point in time.

In order to economize on resources, the conditions detected by one module must be used to support many different behaviours. However, this creates a problem for modifiability, since a change to a module condition set to support learning of one behaviour will tend to introduce undesirable side effects on all the other behaviours supported by the same module. As a result, the degree and type of changes to module condition sets must be severely constrained. For a system that learns a complex combination of behaviours, a module can add a condition to its group if it is similar to and occurs at the same time as conditions already in its group, but with some tightly limited exceptions cannot subsequently change or delete the condition [Coward 2001]. Modules thus detect a "similarity space" (or receptive field) that gradually expands but does not contract.

Reward feedback following a behaviour changes the recommendation weights recently active in the component hierarchy in favour of that behaviour, but cannot change condition definitions in the modular hierarchy. Hence modules cannot be evolved to correspond with unambiguous cognitive circumstances (such as features or categories). Rather, modules must be evolved into a set of statistically semi-independent units that can discriminate between cognitive circumstances with behaviourally different implications. In other words, the group of module outputs in response to instances of one category is sufficiently different from those in response to instances of a different category that high integrity behaviours can be achieved. Electronic simulations have demonstrated that such discrimination can be achieved with module changes restricted to gradual expansion [Gedeon et al 1999; Coward 2000; Coward et al 2001; Ratnayake et al 2004].

In order to achieve a high integrity behaviour, a reasonable range of behavioural recommendations must be generated in response to every input state. This is equivalent to a requirement that at least a minimum level of module outputs must be generated in response to each input state. This requirement is the primary driving force determining the need for recording of additional conditions: if the

[1] In recommendation architecture terminology, the difference between a module and a component is that modules can exchange information with complex behavioural meanings, while exchanges between components can only have simple behavioural meanings. This difference is driven by the use of reward feedback within the component hierarchy [Coward 2001; 2005].

number of modules producing an output is below a minimum level, conditions present within the current input state are added to modules until the minimum level is reached. Such additions must be sufficiently similar to existing module conditions, and new modules could be defined if the current input state does not contain conditions sufficiently similar to existing modules. The new conditions added to a range of modules constitute the "memory trace" of the current experience.

A module in the modular hierarchy detects a group of conditions that are relatively similar, and with as little overlap as possible with the groups of conditions detected by other modules. A condition is defined by a group of sensory inputs; each with a specified state, and the condition occurs if a high proportion of the inputs are in the state specified for the condition. Two conditions are similar if there is significant overlap in the sensory inputs that define them, and they tend to occur in the same sensory input states.

There are some conceptual similarities between modules and the components in independent components analysis [Hyvärinen et al 1999] in the sense that modules decompose a sequence of input states into partially statistically independent "features" in an unsupervised manner. As in the case of independent components [e.g. Bartlett et al 2002], modules will not generally correspond exactly with clear cognitive features. No one module will be present if and only if a particular feature is present. Rather, there will be a group of modules, which tend to be present more often when a particular feature is present. The critical differences between independent components and modules are that modules can continuously evolved with experience by appropriate addition of new conditions or new modules, but are less rigorously statistically independent than independent components.

Conditions can have different degrees of complexity, where the complexity of a condition is defined as the number of raw sensory inputs that contribute (directly or via intermediate conditions, and including duplicates) to the condition. Conditions with different degrees of complexity are useful for determining different types of behaviours. For example, relatively simple conditions may be useful for discriminating between features (e.g. wings, heads, legs etc.), more complex conditions between types of objects (e.g. birds, cats, trees etc.), and yet more complex conditions between types of groups of object (e.g. something chasing something, something climbing something, etc.) as illustrated in figure 1.

Condition definition and detection is performed by devices, and these devices must be arranged in layers. This layering is required to maintain synchronicity for one input state while detecting conditions at a number of different levels of complexity [Coward 2001]. Higher level modules are made up of groups of devices in one layer, columns across several layers, arrays of parallel columns etc. Resource limitations require mechanisms to allow simultaneous detection of conditions derived from multiple input states within the same physical group of modules, in such a way that the condition detections are kept separate until it is appropriate to detect combination conditions. For example, if conditions recommending behaviours in response to a group of objects are needed, conditions must first be detected within each object separately, and then conditions detected that are combinations of individual object conditions. Resource limitations will tend to require use of the same modules for all the individual object condition detections.

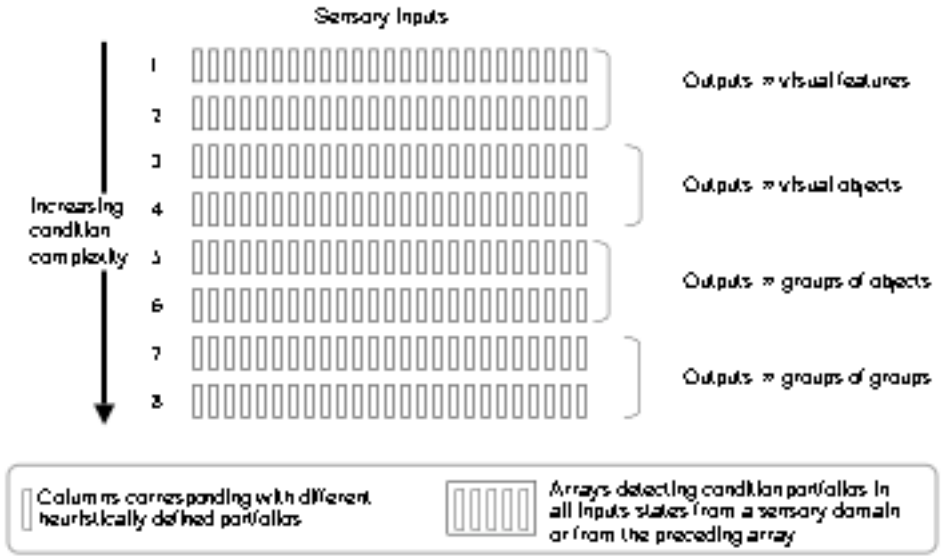


Figure 1. A sequence of arrays of columns detecting conditions on different levels of complexity that can discriminate between different types of cognitive circumstances. Outputs indicating condition detections by one array are communicated to the next array. The ≈ symbol is used to emphasize that the outputs of one column do not correspond exactly with one unambiguous cognitive circumstance, but that the outputs of an array make it possible to discriminate between cognitive circumstances with different behavioural implications. One individual column may have recommendation weights in favour of many different categories (because of some element of similarity between the categories). However, the group of columns producing outputs in response to one instance of a category will be sufficiently similar to the groups for other instances of the same category, and sufficiently different from the groups for instances of other categories that high integrity discrimination can be achieved in the component hierarchy using the recommendation weights associated with each column. In the diagram there are two arrays at different levels of condition complexity providing discrimination between, for example, different types of object. In reality, there could be even more intermediate arrays detecting conditions useful for discriminating between one type of cognitive circumstance.

In order to economize on information recording resources, there must be careful management of when and where new conditions will be recorded. A critical role of the modular hierarchy is to provide this management function. In other words, some condition detections must determine whether or not conditions will be recorded in specific locations. Columns are a key part of this management function. Specifically, a column detects conditions on a number of different levels of complexity, as illustrated in the

three layer column in figure 2. In that column, the top layer detects relatively simple conditions, the middle layer more complex conditions that are combinations of the conditions detected by the top layer, and the bottom layer even more complex conditions that are combinations of the conditions detected by the middle layer. The higher the complexity, the more specific the condition to the input states that contain it: many input states will contain conditions detected by the top layer; fewer will contain conditions detected by the middle layer, and even fewer will contain conditions detected by the bottom layer. If a significant number of conditions detected by the bottom layer, the column will produce an output that will be provided to the component hierarchy. If such an output is present, there is little value in recording any additional conditions. Activity in the bottom layer will therefore inhibit such recording. If there is a need to record conditions (i.e. less that the minimum number of columns is producing an output) there needs to be condition recording in some columns. Conditions must be recorded in columns that have previously recorded similar conditions. If there is strong condition recording in the middle layer of the column, this indicates that the column already contains conditions fairly similar to those in the current input state, and recording in that column would be more appropriate than recording in a column with minimal activity in the middle layer. Activity in the middle layer therefore excites condition recording in the column as a whole. To ensure that condition recording is limited to the most appropriate columns, activity in the middle layer of a column also inhibits recording in any other column, and a competition occurs to determine the most appropriate columns for condition recording. This three layer column illustrates the concept, but although the practical considerations will tend to result in layers and columns, the actual number of layers in a column will depend upon the functional needs of the system [Coward 2005a].

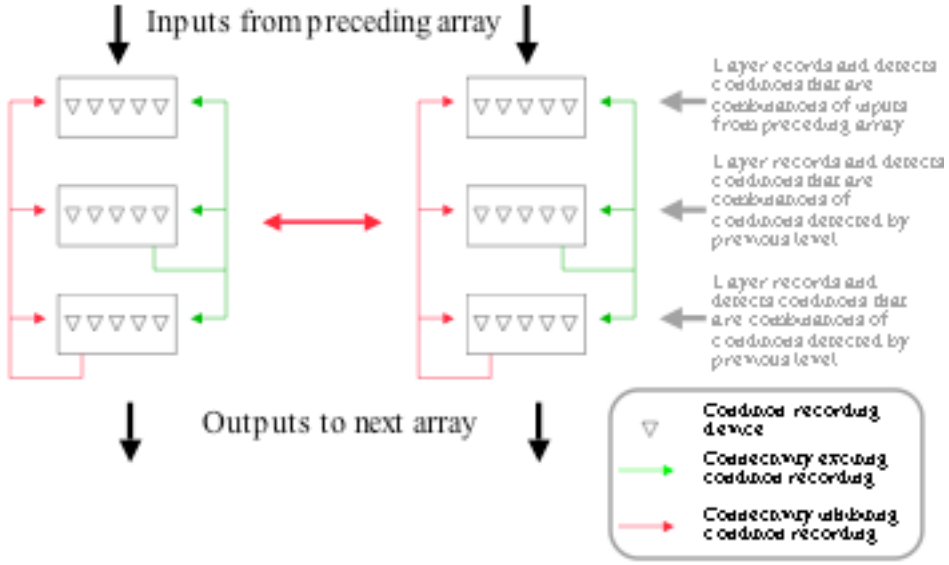


Figure 2 A three layer column. The conditions detected within one layer of a column increase in complexity from top to bottom, where complexity is the number of raw sensory inputs that contribute to a condition either directly or via intermediate conditions. The greater the complexity of the condition, the more specific it is to the sensory input states that contain it. If an input state contains a condition recorded in the bottom layer, the column produces an output and condition recording will in general not be appropriate. Overall activity in the bottom layer therefore inhibits condition recording. If there is activity in the middle layer and none in the bottom layer, this indicates that there is a significant degree of similarity between the current input state and states that have resulted in column outputs in the past. Condition recording could therefore be appropriate. Inhibition between columns selects the columns in which this degree of similarity is greatest.

Direct and Indirect Activation of Information

A module is directly activated by the detection of the presence of some similarity circumstance in current sensory inputs, and such detections have acquired a range of recommendation strengths through past reward feedback. However, the severe limits on module changes mean that some inactive modules may have recommendation strengths relevant to a given sensory input state. For example, if a module is inactive, but has often been active in the past, or has recently been active, or recorded conditions in the past at the same

time as many of the currently active modules, such an inactive module may have relevant recommendation strengths to contribute in the current circumstances. If active modules do not result in a predominant behaviour, such inactive modules could be indirectly activated to expand the range of available behavioural recommendations. Such indirect activations would be appear to the system like pseudosensory experiences generally made up of fragments of many past experiences. Condition recording could occur within such pseudosensory experiences, resulting in a memory trace of "non-real" experiences.

Uncontrolled proliferation of indirect activations would produce chaos. Such indirect activations must therefore be treated as behaviours, and modules have recommendation strengths into appropriate components in favour of indirect activation of other modules on the basis of different types of temporally correlated past activity. The actual activations will be determined by competition between the recommended alternatives. It can be demonstrated [Coward 2005b] that indirect activation on the basis of frequent past simultaneous activity results in the phenomena of semantic memory, indirect activation on the basis of past simultaneous recording results in episodic (and in particular autobiographic) memory, and indirect activation on the basis of recent simultaneous activity results in priming memory.

As an example of semantic memory, consider how the ability to classify different visual images of birds (pigeon, owl, thrush, parrot etc.) as instances of the BIRD category could be supported. Each visual instance will directly activate a different set of columns. However, because of visual similarities, there are some columns that will tend to be activated rather frequently (although probably no one column will always be activated). When the word "bird" is heard, there will be some auditory columns that will tend to be activated rather frequently. Because the word is often heard at the same time as a visual perception of a BIRD instance, the auditory columns will often be active at the same time as the frequently active visual columns. This frequent simultaneous activity means that the auditory columns that are most often active will acquire recommendation strengths in favour of activation of the visual columns that are often active and vice versa. This means that when the word "bird" is heard, it will generate a pseudovisual image of an average bird, at some levels of condition complexity. These levels will generally not include raw visual inputs because there will be considerable differences between different BIRD instance activations at that level. Hence the pseudovisual activation will not be a visual hallucination. When a BIRD instance is seen, it will generate a mental state as if the word "bird" had been heard.

The two key aspects of episodic memory are *firstly* that extensive information can be retrieved about a single unique event, and *secondly* the sequence of occurrences within a long event can be retrieved. As an example of episodic memory, suppose that a novel situation occurred such as watching news of 9/11 on television for the first time. The novelty means a high degree of condition recording. Suppose later the words "nine-eleven", "twin towers" and "airplane" were heard. Each of the words would activate columns on the basis of frequent past simultaneous activity. The result would be an ensemble of active columns that was not the same as those active during the original experience, but there would be some overlap. If then there was indirect activation on the basis of simultaneous past recording, there would be a tendency for the ensemble to evolve towards a closer approximation to the one active

during the original experience. This closer approximation would be experienced as a partial reconstruction of the original experience and would, for example, have recommendation strengths in favour of describing that experience.

Activation of columns on the basis of past activity shortly after currently active columns makes it possible to step through the reconstruction of sequences of experiences in the appropriate order [Coward 2005a, b].

A Biologically Realistic Cognitive Model within the Architectural Bounds

A physiologically realistic cognitive model within these architectural bounds has been described [Coward 2005a]. In this model, the cortex corresponds with the recommendation architecture modular hierarchy. Cortex devices are organized into layers, columns and areas as required by the recommendation architecture bounds. The thalamus, basal ganglia and cerebellum correspond with component hierarchies making behaviour selections determining, respectively, the sensory information which will be allowed to influence behaviour, the general type of behaviour and the specific behaviour. Nuclei within the thalamus and basal ganglia correspond with different more specific behaviour types. The hippocampus corresponds with the recommendation architecture subsystem that manages the assignment of modular hierarchy (cortex) resources. Other brain structures correspond with other recommendation architecture subsystems [Coward 2005a].

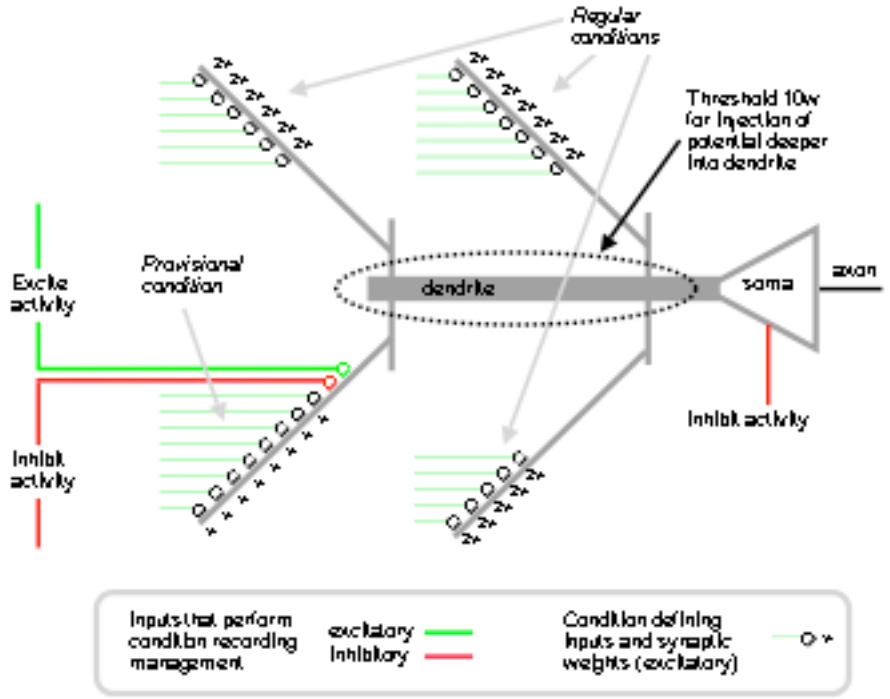


Figure 3. Pyramidal neuron modelled as a condition recording and detection device. The neuron has a dendritic tree made up of a number of branches. There is separate integration of inputs within each branch, then integration across all the branches for which the partial integration exceeds a threshold. Previously recorded conditions correspond with branches in which the total of available synaptic strengths of condition defining inputs exceeds the threshold for contributing to tree integration. A large subset of the inputs to one branch will therefore generate a contribution to dendritic tree integration. A provisional condition provides a framework within which a new condition could be recorded. The condition defining inputs to a provisional condition have total available synaptic strengths less than the threshold for contributing to tree integration. However, inputs that excite condition recording could result in that threshold being exceeded. If the neuron shortly afterwards produces an output action potential, the strengths of the condition defining inputs are increased to the point at which their total synaptic strengths exceed the threshold. In information terms a condition has been recorded.

Pyramidal neurons in the cortex correspond with devices that define and detect conditions. The physiological mechanisms involved can be understood by consideration of figure 3. In the figure, there are a number of groups of synaptic inputs to different pyramidal neuron dendrite branches. A branch integrates the post synaptic potentials from these synapses, and injects potential into the

dendrite if a threshold is exceeded. Further integration occurs within the dendrite as a whole, and then integration within the neuron body (soma) determines the generation of action potentials into the output axon [Sourdet and Debanne, 1999]. In information terms, such an action potential indicates the detection of a significant number of the conditions programmed on the separate dendrite branches.

Three of the branches illustrated in figure 3 correspond in information terms with recorded conditions. The synaptic weights of the condition defining inputs are positive and their total exceeds the threshold for injecting potential further into the dendrite. If a large proportion of these inputs is active, the detection of the condition defined by the group of inputs will be signalled to the dendrite and potentially to the neuron soma.

It would be impractical to create the connectivity defining a new condition at the instant the new condition was required. Hence provisional conditions must be defined in advance, providing the framework for definition of a regular condition, generally using a subset of the available input connectivity. A provisional condition is defined by set of inputs to the branch on the lower left of figure 3. The total weights of the condition defining inputs to this provisional condition are insufficient alone to result in potential injection deeper into the dendrite. In other words, the condition will not be detected on the basis of these inputs alone: no condition has been programmed. However, there are also excitatory and inhibitory inputs that manage the condition defining process. These condition definition management inputs are ultimately derived from column activities as illustrated in figure 2. If a significant proportion of the condition defining inputs is active, inputs exciting condition recording are also active, and inputs inhibiting condition recording are inactive, the total postsynaptic potential may be sufficient to inject potential into the dendrite. If shortly afterwards the neuron produces an output action potential, an action potential also backpropagates into all the recently active dendrite branches. This backpropagating action potential increases the postsynaptic strengths of all recently active synapses (i.e. the well known long term potentiation (LTP) mechanism [Bi and Poo, 1998]). The effect will be that the recently active subset of the condition defining inputs may acquire enough postsynaptic strengths to inject potential deeper into the dendrite independent of the state of the condition recording management inputs: in information terms a condition has been recorded.

An important issue is how provisional condition defining inputs are selected. The selection could be random, but the probability that a behaviourally useful condition can be increased by biasing the random selection in favour of inputs often active in the past at the same time as the target neuron. Such a bias could be achieved by taking the brain off line, and performing an approximate fast rerun of past experience, establishing provisional connectivity between neurons often active at the same time during this rerun. Coward [1990] suggested that sleep is the off-line period during which provisional conditions are configured, and REM sleep provides the rapid rerun information. Simulations have demonstrated that such a process reduces resource requirements (by making conditions more relevant) by about 20% [Coward 2000]. Effectively, resources for recording memories of future experiences are configured to be as appropriate as possible during sleep. Note that this proposal is radically different from suggestions that recent memories are consolidated during

sleep [e.g. McClelland, McNaughton and O'Reilly, 1995]. The advance configuration process will have no effect on prior memories. It is analogous with formatting a storage disk in electronic systems.

Pyramidal neurons are organized into layers, with columns of neurons penetrating several layers. These columns are organized into arrays in which the same input space is available to all the columns in the array. In the recommendation architecture model, a significant proportion of the organization and interconnectivity within and between columns and arrays is to manage the circumstances under which conditions will be recorded. Note that the recommendation architecture bounds do not specify the number of layers in a column, only the different functions that must be performed. A way in which observed inter and intra layer column connectivity in the Macaque cortex can be mapped to recommendation architecture requirements is described in Coward [2005a].

The receptive fields of pyramidal neurons in columns in the visual cortex increase in complexity across successive column areas. The expected cognitive ambiguity is present: even in the TE visual area that discriminates between visual objects the receptive fields do not correspond with objects, but with ambiguous forms that will be detected in some but not all visual inputs [Tanaka 2003].

In order to detect conditions within a group of objects, it is necessary for conditions to be detected simultaneously within the individual objects in arrays 1 through 4 in figure 1, then combination conditions detected in the following arrays. However, it is important that the conditions detected within different objects are kept separate. In other words, in arrays 1 through 4 it is important not to detect conditions containing information derived from multiple objects. As described in detail in Coward [2004] the requirement to keep independent populations of condition detections active without interference within the same physical groups of neurons can be achieved by a frequency modulation mechanism applied to leaky integrator neurons. To simplify the discussion, partial integration across subsections of the dendritic tree will be omitted.

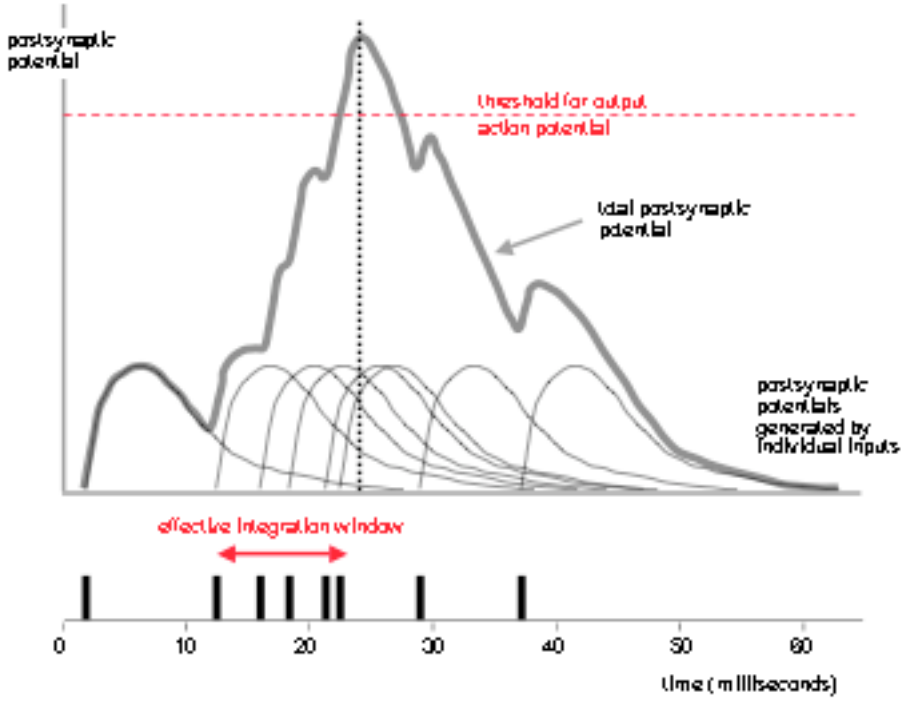


Figure 4. The leaky integrator neuron model. The postsynaptic potential resulting from one input action potential rises rapidly and then decays. There must therefore be a number of input action potentials within a relatively short period of time to exceed the threshold for an output. This period of time can be regarded as an integration window that will in practice be rather shorter than the total period over which one input contributes potential. For simplicity, in the figure all the input action potentials contribute the same postsynaptic potential.

In a leaky integrator neuron model, inputs and outputs are action potential voltage spikes. An input spike injects a potential into its target neuron that initially increases rapidly, then decays exponentially with some time constant. The total postsynaptic potential is therefore the sum of the effects of recent input action potentials as illustrated in figure 4. Because of the decay in postsynaptic potential, if two action potentials arrive too far apart in time, they will not reinforce each other significantly. Hence the concept of an integration window is useful: if a number of spikes arrive within a limited period of time, the threshold will be exceeded. The integration window is in principle no larger that the total period during which an input contributes postsynaptic potential, but in

practice is significantly shorter because the contribution of a spike is very small towards the end of this period as illustrated in figure 4.

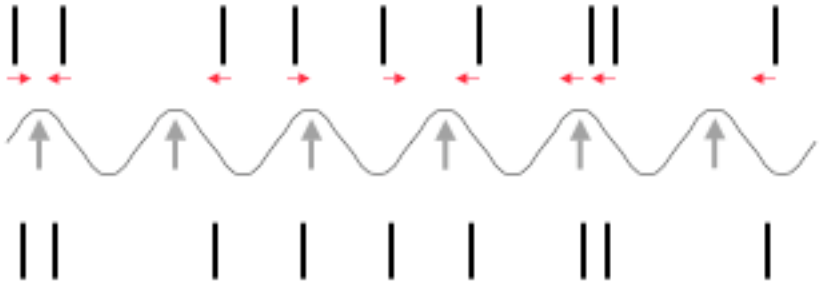


Figure 5. An irregular sequence of output action potentials (top) from a neuron is frequency modulated by an imposed signal (middle) resulting in a modulated spike sequence (bottom) that is still irregular, but has a tendency for spikes to occur more often close to peaks in the modulation signal.

The existence of an effective integration window makes it possible for a frequency modulation mechanism to support multiple independent activations in the same physical set of devices. The meaning of frequency modulation is illustrated in figure 5. An unmodulated sequence of action potentials coming from a neuron is illustrated. Such sequences are in general not regular. The sequence is frequency modulated if each action potential in the sequence is shifted slightly towards the nearest peak in a modulation signal. The resultant sequence still appears irregular, but there is a higher probability of an action potential close to the peaks in the modulation signal. The modulation signal would be directed to a set of devices and each device would have its outputs modulated with the same phase. If the same average output rate was needed, the threshold of each device would be lowered slightly. In figure 5 the modulation signal is illustrated as a continuous sine wave, but in practice the signal could be a sequence of action potentials directed to the appropriate target devices.

If a subset of the currently active neurons in one layer were given such a frequency modulation and other neurons were not given the modulation, the effect would be that groups of outputs from those neurons would be much more likely to arrive within the integration window of neurons in the next layer, and therefore generate outputs from those neurons. Those outputs would also tend to be in phase with each other and the bias would propagate to the next layer and so on. For example, if a subset of the neurons responding to visual inputs could be given such a bias, and subsets corresponded with an object in the visual field (for example, by

targetting all the neurons with receptive fields within some closed boundary with the modulation signal), the effect would be that conditions within the one object would tend to be detected rather than within other objects. This mechanism provided a physiological model for the attention function [Coward 2004].

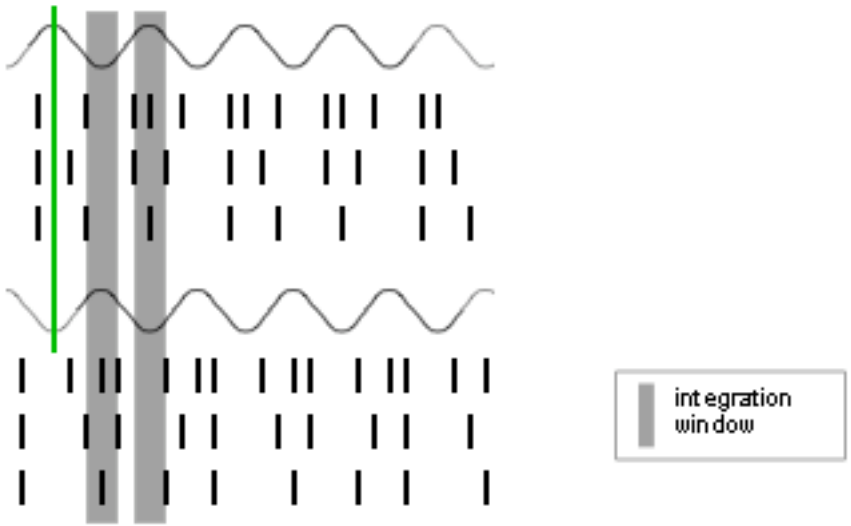


Figure 6. Two groups of three neuron outputs, modulated with different phases of the same frequency, tend to have most of their output action potentials within different integration windows.

Furthermore, if different subsets were modulated with different phases of the same frequency, then provided that the integration windows did not overlap, the effect would be that outputs from the different subsets would not contribute to integration outside their own subset as illustrated in figure 6. The effect would be that two independent populations of condition detections would be present, corresponding with, for example, two different objects.

The ratio of the time between peaks in the modulation signal to the integration time gives a rough indication of the number of different objects that can be processed separately in the same physical resources. If the gamma band (~ 40 Hz) EEG signal is interpreted as the modulation signal, and an integration time ~ 8 milliseconds is assumed, then this ratio indicates a limit of 3 - 4 objects can be processed simultaneously in the same physical neuron resources. This limit can provide an account for working memory observations [Coward 2004; 2005b].

As discussed earlier, individual columns will not correspond exactly with specific cognitive circumstances. Rather, each column will have recommendation strengths in competition in favour of a wide range of different behaviours appropriate in response to different cognitive circumstances, which may contain conditions recorded within the column. Some of the behaviours that may be recommended by a column are listed in table 1.

Recommendation strengths are instantiated in the connection strengths of the outputs from cortex columns into neurons in different nuclei within subcortical structures such as the thalamus and basal ganglia. Behaviour types A and B ultimately result in physical muscle movements outside the brain. However, the other types result in different information management behaviours within the brain that are important intermediate steps in the course of generating behaviour.

Consider first attention behaviours (type C). Columns detecting conditions correlating with the presence of closed boundaries at different places in the visual field all have recommendation strengths in favour of shifting the attention domain to correspond with their boundary. Acceptance of one such recommendation may require eye movements, but the key result of acceptance is that all the sensory inputs derived from the area within the closed boundary are modulated with the same phase, and conditions within the selected object are therefore preferentially detected. Sequences of attention behaviours exist for learned cognitive processes. For example, in processing an arithmetic equation, attention shifts in a particular sequence between different sub-objects within the equation [Coward 2005a].

Information availability recommendation types (D) are required to manage how long a group of condition detections will be allowed to influence behaviour selection, and also to manage when the outputs from several separate active column populations (at different phases) in one array will be synchronized and released to the next layer.

Indirect activation recommendations (E, F, and G) can expand the population of condition detections available to influence behaviour. The conditions detected within the visual objects currently the focus of attention will have recommendation strengths in favour of externally directed behaviours in response to the object. However, there may be other conditions which are not currently being detected but which may have appropriate recommendation strengths for current circumstances. For example, columns which are currently inactive (i.e. not detecting conditions) but which have often been active in the past at the same time as currently active columns may have relevant recommendation strengths in some circumstances.

A. Motor Behaviour Management

1. Perform a general sequence of motor behaviours
2. Perform a specific sequence of motor behaviours
3. Perform an individual motor behaviour

B. Speech Behaviour Management

1. Generate a sound
2. Speak a word
3. Say a phrase
4. Express meaning verbally by a sentence

C. Attention management

1. Perform a general sequence of attention and internal activation behaviours
2. Perform a specific sequences of attention and internal activation behaviours
3. Perform an individual attention behaviour

D. Managing Information Availability

1. Prolong the activity of currently active portfolios
2. Shorten the activity of recently activated portfolios
3. Synchronize the activity of several different populations of currently active portfolios in the same array
4. Release outputs of one array to the next array

E. Activation Based on Frequent Past Activity

1. Activate portfolios which have often been active at the same time in the past as the currently active portfolio
2. Activate portfolios which have often been active just after past activity of the currently active portfolio
3. Activate portfolios which have often been active just before past activity of the currently active portfolio

F. Activation Based on Past Correlated Condition Recording

1. Activate portfolios containing conditions recorded at the same time in the past as some conditions in the currently active portfolio
2. Activate portfolios containing conditions recorded just after some conditions in the currently active portfolio
3. Activate portfolios containing conditions recorded just before some conditions in the currently active portfolio

G. Activation Based on Recent Activity

1. Reactivate portfolios which have recently been active
2. Reactivate portfolios which recently recorded conditions

Table 1. Different types of behavioural recommendation strengths that may be possessed to different degrees by even a single column

Columns therefore have recommendation strengths in favour of activating such other columns. Similarly, if two columns record conditions at the same time, each column acquires recommendation strength in favour of activating the other. Such recommendation strengths will in general decay with time unless reinforced by actual use followed by positive consequence feedback. As discussed in Coward [2005b], indirect activation on the basis of frequent past simultaneous activity supports semantic memory, and indirect activation on the basis of past simultaneous condition recording supports episodic memory. Such indirect activations must be recommendations that compete with, for example, externally directed behaviours, otherwise the brain could be swamped with irrelevant information in the course of performing any behaviour.

All of these recommendation strengths are instantiated as weights possessed by column outputs into components corresponding with the behaviours. Components can correspond with types of behaviour, specific behaviours or sequences of specific behaviours. Selection of a type of behaviour will result in a bias in favour of components corresponding with the specific behaviours of the type. Selection of a sequence of behaviours will result in a bias in favour of the components corresponding with the individual behaviours in the appropriate sequence. Such behavioural sequence components can be defined heuristically when a particular behavioural sequence is often invoked. Implementation of the component hierarchy requires devices with input weights that can vary continuously, conceptually similar to the classical perceptron.

Modelling Conscious Phenomena

The cognitive model described in the previous section makes it possible to model conscious phenomena in terms of physiology.

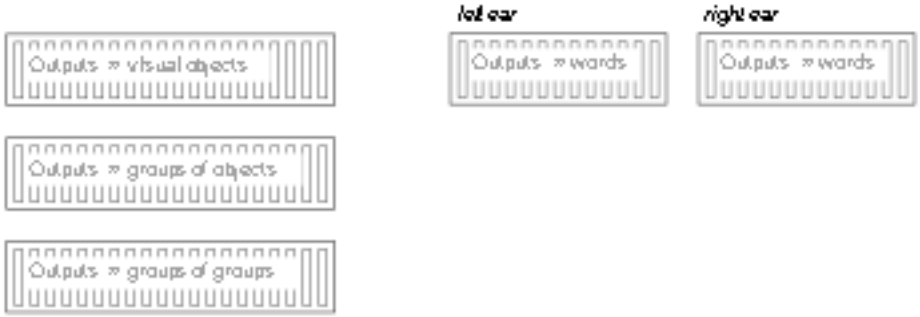


Figure 7. A simplified set of column arrays for the purpose of describing dichotic listening. In this simplified version, an array of columns receives auditory inputs from the right ear and generates outputs that discriminate between different words. Another array of columns performs a similar function for left ear inputs. Three arrays of columns process visual inputs, generating outputs that discriminate between different types of object, different types of groups of object, and different types of groups of groups of objects. The ≈objects array is able to support independent activations detecting conditions in several different objects by maintaining their activity at different phases of frequency modulation. At appropriate points, outputs from these activations are brought into the same phase and released to the ≈group of objects array where they generate condition detections that contain information from all the objects in the group. Similarly for combinations of ≈group of objects activations generating a ≈group of groups activation.

Dichotic Listening

Consider first the dichotic listening phenomenon discussed earlier. A version of the cognitive architecture simplified for the purposes of explanation is illustrated in figure 7. The sequence of arrays illustrated in figure 1 has been compressed to three arrays in

the case of visual processing, and one array for each ear in the case of auditory processing. The columns in the first visual processing array detect conditions which can discriminate between different types of visual object, those in the second array can discriminate between different types of groups of objects and so on. The columns in the array detecting conditions in the auditory information presented to the left ear can discriminate between different words, and similarly for the right ear.

The activation state at the point at which the text switch occurs is illustrated in figure 8. When a word is heard in the left ear, columns detect conditions within the auditory information contained in the word. To simplify speech learning considerably (for a somewhat more realistic discussion see Coward [2005a]), the "auditory" columns activated by hearing a word like "dog" have often been active in the past when visual columns activated by seeing a dog have also been activated, because a teacher has often spoken the word when the learner's attention was directed towards the visual object. The auditory columns have therefore acquired recommendation strengths in favour of activating the visual columns.

Hence the word "dog" results in activation of a visual image of an average dog, although only in arrays detecting conditions at intermediate levels of complexity (i.e. not at the simpler levels where the result would be a visual hallucination, because there is much less consistency in past activity between the auditory columns and the visual columns at these levels). Hearing a phrase like "the black dog" first indirectly activates columns in the top visual array corresponding with "the" and holds them active[2], then columns (at a different modulation phase) corresponding with "black", then columns corresponding with "dog". The outputs from the three column populations in the top array are then synchronized and released to the middle array and conditions detected corresponding with a visual image of the phrase, or group of objects. This population is held active.

[2] The pseudovisual activation in response to hearing the word "the" is determined by the columns that have most often been active in the past when the word has been heard. These columns will modify the activation in response to "dog" and a different way from the modification that would result from hearing "a" shortly before "dog". In this very simplified discussion of speech processing this point will not be discussed.

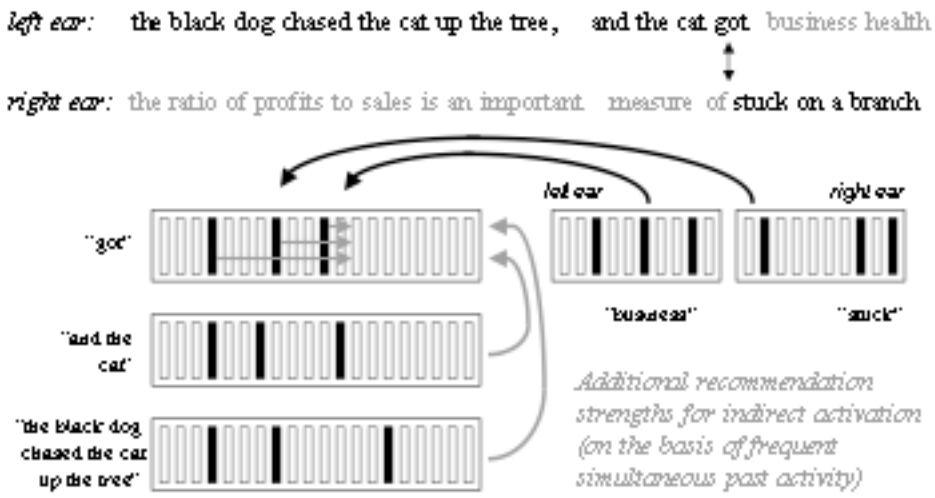


Figure 8. The pattern of column activation during a dichotic listening task, just as the switch of texts between ears occurs.

A similar process then occurs in response to the phrase "chased the cat", resulting in a second activated column population in the middle array. After the phrase "up the tree" there are three separate activated column populations in the middle array, and the outputs from these populations are synchronized and released to the bottom visual array. A population of columns is activated in that bottom array that detects conditions in the group of phrases "the black dog chased the cat up the tree". Hearing the next phrase "and the cat" leads to an activated column population in the middle array, then hearing the word "got" results in a single activated population in the top array. At this point, as illustrated in figure 8, there is one population active in the bottom array, one population in the middle array, and one in the top array.

At each point in this process, there were also recommendation strengths in favour of indirect activation of visual columns corresponding with the words heard in the right ear, but the recommendation strengths of words heard in the left ear were enhanced by the instruction to echo the text heard in that ear.

Consider now the situation when the word "business" is heard in the favoured left ear, and the meaningful continuation "stuck" in the right ear. There are some additional recommendation strengths that must be considered. The visual columns currently active in the three arrays have recommendation strengths in favour of activating portfolios that have often been active at similar times (the same

time or shortly after) in the past. Because "stuck" is the meaningful continuation, these strengths reinforce the indirect activation strengths of the auditory portfolios corresponding with the left ear. These strengths are sufficient to shift the predominant recommendation strengths over in favour of that meaningful continuation. Note that these strengths were always present, but until the switch they reinforced continuation in the targetted ear. The indirect activation recommendation strengths of columns activated by the untargetted right ear were also always present, but did not result in pseudovisual activation and therefore (through condition recording) memory. However, their constant presence results in the switch when they are reinforced by the recommendation strengths of the currently active visual portfolios.

The result is that the unattended words do not generate any pseudovisual activation. Such an activation is required if there is to be any verbal report of the words at the time, and the unattended text is therefore outside of access consciousness. The absence of the activation also means that the unattended text cannot result in any indirect activations which could support stream of consciousness as discussed below. Finally, the absence of an activation means that there is no condition recording which could be the basis for future recall. The model thus provides a complete account for the observations.

The behavioural scenario

In the behavioural scenario discussed earlier, a person is out walking with a companion, and encounters a tree partially blocking the path. One behaviour is simple avoidance: stepping around the tree. A second behaviour is to make the comment "Mind the tree". A third behaviour is to focus attention on the tree and "become aware" of the tree as a tree. The fourth behaviour is to follow a sequence of internal "mental images" ending with the comment "Up in the mountains I saw a whole area covered with trees like it. I wonder what it would be like to have a group of trees like that in my garden".

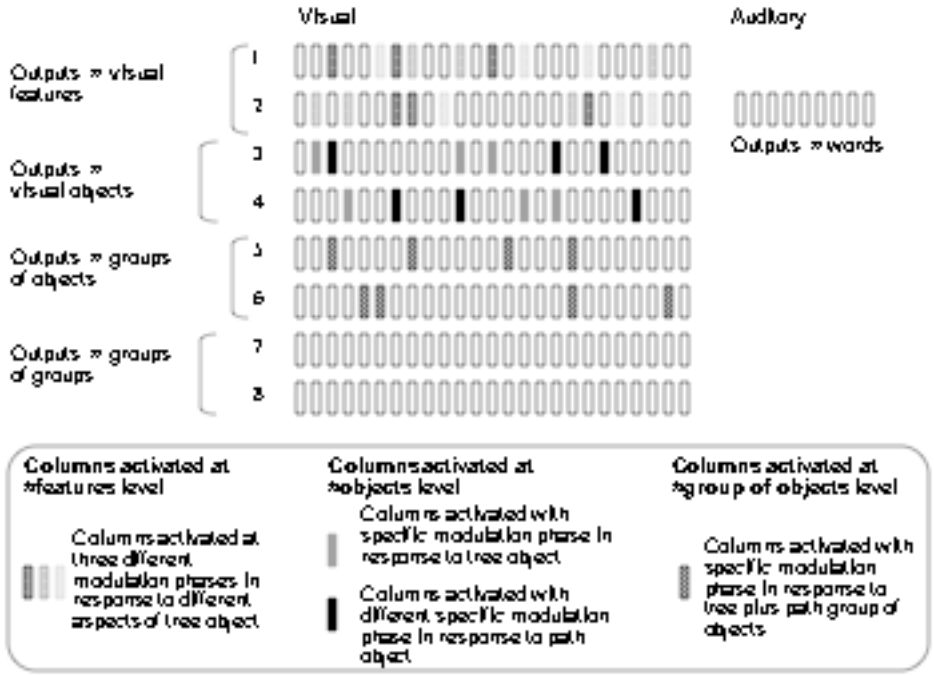


Figure 9. Pattern of column activation supporting motor behaviours for avoiding a tree. In the top two layers, columns are activated by detecting conditions within three different aspects of a tree (for example, the attention domain focussed on the trunk, roots and canopy). Outputs from these two populations are synchronized to the same phase and released to the next layer. In the next two layers, columns are activated by detecting conditions in the tree as a whole, combining conditions detected by the two earlier populations. In addition, another column population was earlier created by a similar attention, synchronization and release process in response to seeing the path. This "path population was generated in the layers 3 and 4, and its activity prolonged at a different phase of frequency modulation from that generated by the tree. Outputs from the two populations in layers 3 and 4 are synchronized and released, generating a population ≈group of objects in layers 5 and 6. This population has recommendation strengths in favour of appropriate motor movements to avoid walking into the tree, and active columns in any layer could also have recommendation strengths in favour of relatively simple verbal behaviours (such as describing what is seen).

To understand the physiological processes supporting this scenario, consider the pattern of activation illustrated in figure 9 for the same architecture as illustrated in figure 1. The figure shows the column activation pattern at the moment when the behaviour of avoiding the tree is initiated.

Somewhat prior to this moment, visual information about the path was being processed. This processing resulted in several populations of active columns in arrays 1 and 2 (generated as a result of attention directed on different parts of the path). These populations were maintained active simultaneously at different phases of frequency modulation, and then the outputs from those populations were synchronized and released to array 3, generating an active column population in arrays 3 and 4. This "path as a whole" population in arrays 3 and 4 was maintained active and the active populations in arrays 1 and 2 extinguished.

Visual attention on the tree then results in activation of three column populations in arrays 1 and 2 as illustrated in figure 9, generated by different parts of the tree (e.g. trunk, crown, roots) and maintained active at different phases of frequency modulation. Array 2 outputs from these populations are synchronized and released to array 3, generating an active population in arrays 3 and 4 which is maintained at a different phase of frequency modulation from the "path as a whole" population. Array 4 outputs from the "path as a whole" and "tree as a whole" populations are synchronized and released to array 5, resulting in an active column population in arrays 5 and 6 at the ≈group of objects level of complexity. The active columns in this population (combined with column activations resulting from propioceptic inputs that are not illustrated) have recommendation strengths in favour of avoidance motor behaviours.

Any of the active columns in this "tree plus path" population also have recommendation strengths in favour of verbal behaviours, and if that type of behaviour is accepted could result in "mind the tree" speech. All of the column activations in figure 9 are the result of detection of conditions directly within current sensory inputs. If there is little novelty in the tree and path objects, there may be little condition recording. Hence motor behaviours are generated, but there may be little future memory of the event.

Now suppose that indirect activation behaviours are encouraged. This will mean that the activated population will be larger (with higher biological cost), and may have overall motor recommendation strengths somewhat less appropriate for the walking motions required. However, the population will include columns that have often been active in the past at the same time as, or that recorded conditions in the past at the same time as, the columns directly activated by current sensory inputs. The result will be additional behavioural recommendations that may be relevant to current circumstances. The third behaviour of "becoming aware" of the tree can be understood as this type of activation. Different subsets of this population will be subsets of populations active during the very large number of past experiences in which trees were present. In general, none of these subsets will be large enough to generate speech behaviour appropriate to the corresponding past experience. The effect will therefore be a much richer mental experience made up of fragments of many past experiences featuring trees. Any one fragment will not have enough speech recommendation strength to describe the past experience, so it will not be possible to describe the experience verbally in detail. The content of the mental experience will be specific to the past experience of the individual.

The behavioural value of such a "conscious" activation is to search a much larger space of possibly appropriate behaviours than is available from conditions actually present in current sensory inputs. The costs are higher biological effort, and the risk of interference

with the most appropriate behaviour in response to the currently perceived objects. Hence selection of "conscious" activation behaviours will depend upon the value of such behaviours compared with the required accuracy of current motor behaviours.

The additional behaviours which could be recommended as a result of a conscious activation of this type could include immediate social communication and planning for future behaviours. The development of recommendation strengths in favour of such behaviours proceeds by a sequence of secondary, tertiary etc. indirect activations.

Within the large initial population of indirectly activated columns, there will be some subgroups that have often been active at the same time in the past, and other subgroups that recorded conditions at the same time in the past. A large subgroup could have a somewhat larger total indirect recommendation strength than any other subgroup. Evolution towards a total population consistent on the basis of past temporally correlated activity would result in a population approximating to the original population during the experience of which the subgroup was a fragment, as discussed earlier for episodic memory retrieval. If at some point such an evolution resulted in a less than minimum activated column population, condition recording could occur to bring the population up to the minimum level. Such condition recording would result in a memory of a non existent circumstance that could be retrieved at some future time. However, in general such memories would not be connected with more raw sensory inputs and could be distinguished from actual experiences on this basis.

At each point in time, some additional columns detecting conditions within the current sensory environment could be incorporated in the ongoing active population and would influence its evolution. For example, sensory input from the walking companion could bias the evolving population towards content of interest to that companion (e.g. by activating columns often active in the past when that companion has been present).

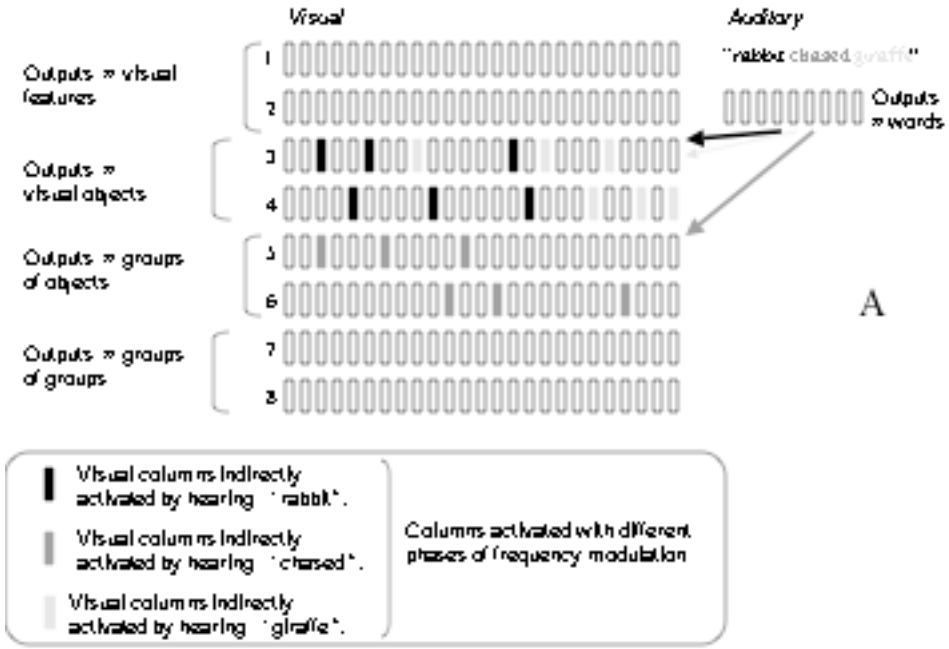
In the specific example, suppose that in the initial indirectly activated population there were columns activated by past perceptions of my garden (because there are trees in that garden), and by past perceptions of hiking in the mountains (because of the presence of trees). If these two fragments expanded, the result could be an image of an area of small but old trees in my garden, supporting the fourth behaviour "Up in the mountains I saw a whole area covered with trees like it. I wonder what it would be like to have a group of trees like that in my garden".

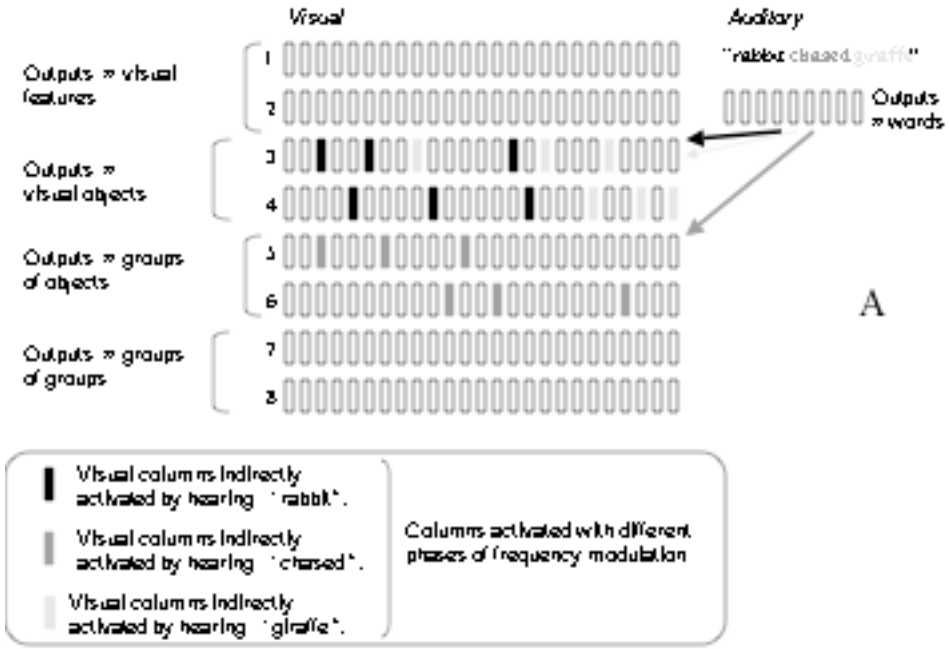
To make the development and impact of the fourth behaviour more clear, consider a much simpler example of the creation of the memory of a non-existent situation. Suppose the words "rabbit chases giraffe" were heard. The three words activate populations of auditory columns that generate pseudovisual activations as illustrated in figure 10A. The words "rabbit" and "giraffe" generate two activated column populations in the ≈objects arrays 3 and 4 at different phases of frequency modulation. The word "chased" generates an active column population in the ≈group of objects arrays 5 and 6, reflecting the frequent activity of the auditory columns at the same time as many different visual perceptions of group of object chase scenes. Then as in figure 10B, the outputs from array 4 are synchronized to the phase of activity in array 5 and released to that array. A combination population is then generated in arrays 5 and

6 that resemble the population that would be activated by actual observation of a rabbit chasing a giraffe[3]. Because such a situation has never been observed before, condition recording will be needed in arrays 5 and 6 to reach the minimum population size for an integrated population.

The activity across arrays 3 to 6 will be similar to the activity that would be generated by actually seeing a rabbit chasing a giraffe. However, there is no activity in layers 1 and 2 close to sensory input that would be present for such an actual observation. The imagined situation is therefore not a visual hallucination. The condition recording makes it possible to recall the imaginary situation in the future.

[3] This description of the handling of "rabbit chased giraffe" in terms of column activations is a simplification which does not address the difference between "rabbit chased giraffe" and "giraffe chased rabbit ". This difference is supported by the existence of two ≈group of objects levels which actually have different information interpretations. Column activations in array 5 can discriminate between situations like "rabbit chased", "chased rabbit", "rabbit held", and "held rabbit" etc. In other words, the word "chased" activates two populations in array 5: a "something chased" population and a "chased something" populations. The word also activates a "something chased something" population in array 6. The "rules of grammar" define that the outputs from the first population activated in array 4 are combined with the "something chased" population in array 5, and the outputs from the second population in array 4 are combined with the "chased something" population in array 5, and the outputs from the two resultant populations are synchronized and combined with the "something chased something" population in array 6. The "rules of grammar" are therefore learned rules for managing indirect population activations.


Visual
Auditory
"rabbit chased giraffe"
Outputs = visual features
Outputs = visual objects
Outputs = groups of objects
Outputs = groups of groups
Outputs = words
1
2
3
4
5
6
7
8
A
Visual columns indirectly activated by hearing "rabbit".
Visual columns indirectly activated by hearing "chased".
Visual columns indirectly activated by hearing "giraffe".
Columns activated with different phases of frequency modulation


Visual
Auditory
"rabbit chased giraffe"
Outputs = visual features
Outputs = visual objects
Outputs = groups of objects
Outputs = groups of groups
Outputs = words
1
2
3
4
5
6
7
8
A
Visual columns indirectly activated by hearing "rabbit".
Visual columns indirectly activated by hearing "chased".
Visual columns indirectly activated by hearing "giraffe".
Columns activated with different phases of frequency modulation

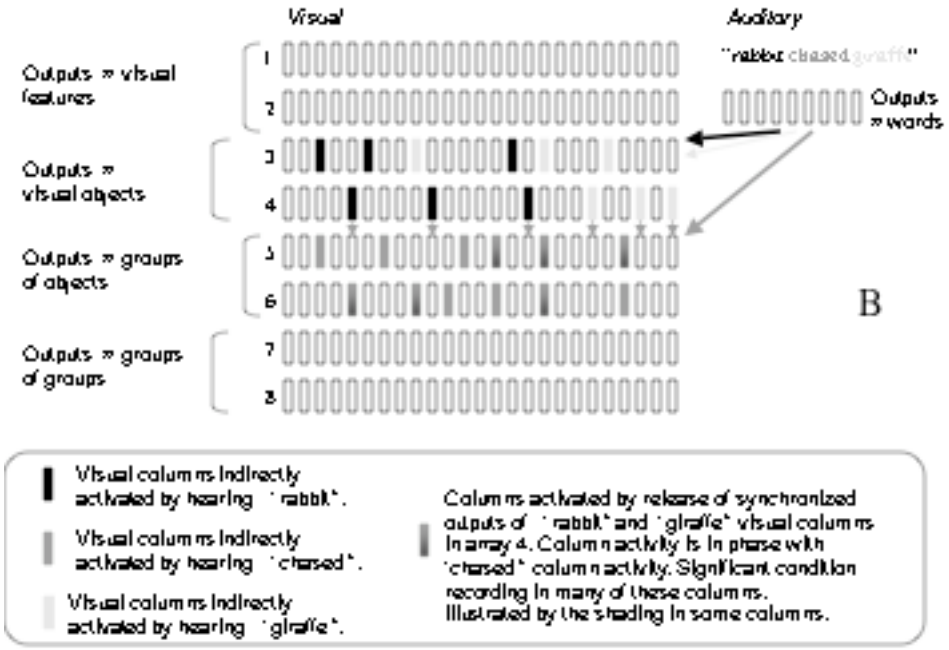


Figure 10. This figure illustrates how spoken words could lead to a pseudovisual experience of a non-existent circumstance, with the possibility of later recall of the experience. In 10A, hearing the words "rabbit chased giraffe" results in a sequence of three active column populations in an auditory column array at a level of complexity that can discriminate between different words. These populations drive the activation of visual columns on the basis of frequent past simultaneous activity. These pseudo visual populations are activated in the level ≈visual objects in the case of "rabbit" and "giraffe" and in the level ≈groups of objects (with probably some ≈objects level activation) in the case of "chase". Active columns in different populations but within the same array produce outputs with different phases of frequency modulation. In 10B, outputs from the different populations in the ≈objects array are synchronized to the same frequency modulation phase and released to the ≈group of objects array, where they combine with the previously activated columns to create an integrated population. Because a rabbit chasing a giraffe has never been seen before, there is some condition recording at this ≈group of objects level which could be the basis for future reactivation of the population.

Speech and the Evolution of Indirectly Activated Populations

The proposed model for consciousness depends upon the evolution of primary active column populations generated from sensory inputs. This evolution is on the basis of past temporally correlated column activity: columns can be indirectly activated if they were

recently active, often active in the past, or recorded conditions at the same time in the past as a significant number of currently active columns. Once a secondary population has been generated on this basis, it in turn can generate a tertiary population and so on. Long sequences of active populations can result without further reference to current sensory inputs, although additional sensory inputs could be incorporated to some degree at any point. The behavioural value of such evolutions is that they may result in end populations with behaviourally useful recommendation strengths. Such behaviourally useful results could be in the areas of planning or complex social behaviours [Coward 2005a].

A problem with this evolution process is that it could drift and become cognitively meaningless, or in other words rarely end with a behaviourally useful population. A second problem is that only 3 - 4 independent populations can be supported at the same time in one array, which could be a limitation if both current sensory inputs and multiple conscious image streams must be processed at the same time.

In the case of drift into cognitive meaninglessness, one way to limit such drift is to make use of capabilities created by speech. As outlined earlier and discussed more fully in Coward [2005a], speech is dependent on the capability of cortical columns to indirectly activate other columns on the basis of frequent past simultaneous activity. For example, auditory columns can indirectly activate visual columns (e.g. hearing the word "bird" activates a pseudovisual image of a bird that is an average over past experience, without the activity close to sensory input that would make it a visual hallucination). In addition, visual columns have recommendation strengths in favour of verbal behaviours (e.g. the pseudovisual bird columns have recommendation strength in favour of saying "bird").

There is a further result of the indirect activation capability. Because the word "bird" is often said after seeing a bird, the visual bird columns are often active just before the auditory "bird" columns, and can therefore acquire recommendation strengths in favour of activating the auditory columns. Such an activation would be the experience of seeing a bird followed by a pseudoauditory experience of hearing the word "bird" (but, again, without the raw auditory sensory experience). Of course, the pseudoauditory column activation would have recommendation strengths in favour of a pseudovisual activation (and so on).

This capability to generate alternating visual and auditory activations can be useful in reducing both of the problems discussed earlier. Firstly, the verbal-visual associations can have the effect of focusing an evolving population. If in a secondary visual population, there are a number of groups of active columns that are subsets of the groups activated by hearing different words, each group will tend to recommend activation of its corresponding auditory columns. Those auditory columns will in turn recommend activation of their corresponding visual set. If one visual subset is somewhat larger than the others, it will tend to grow at the expense of the others by this indirect activation back and forth between visual and auditory. Hence the derived population will tend to converge on cognitively meaningful states.

Furthermore, the sequence of arrays processing auditory information can somewhat increase the number of independent populations that can be supported. Thus somewhat independent secondary populations could be supported and separately processed in

both visual and auditory arrays, although separate processing reduces the cognitive convergence advantage discussed in the previous paragraph.

In practice, a simple motor task may require processing of visual sensory inputs up to the ≈objects levels, leaving the more complex visual levels available for indirect activations (i.e. "thinking"). Required "thinking" at the ≈object level could be handled by the auditory ≈words levels, the results passed to the ≈phrases levels and then generating pseudovisual activity at the ≈group of objects level.

The model for the stream of consciousness is thus that there could at some point in time be a population of columns activated by current sensory input. This population can then evolve by indirect activation, being focussed at certain points to a population with a strong, consistent recommendation strength in favour of verbal expression, such a population being equivalent to a consistent mental image. Between such fixed points there will be no strong, consistent verbal recommendation strength, and the experience will be of a vague and evolving mental state. This sequence of relatively consistent mental images separated by periods of vague evolution is the way stream of consciousness was originally described by James [1892] as substantive "resting places ..... occupied by sensorial imaginations of some sort" separated by transitive "places of flight ..... filled with thoughts of relations, static or dynamic, that for the most part obtain between the matters contemplated in the periods of comparative rest".

The content of consciousness at the fixed points is therefore a mental image, which could be visual or propioceptic or even of feelings, and could be coupled with a mental verbal image. The content between these points is much more vague, and will in general not be fully expressible in speech. Chrisley and Parthemore [2007] has discussed various ways of specifying the content of consciousness. Their alternatives include verbal definition, visual images, and specification by the actions a subject with the experience might be expected to perform. In the proposed model, the content of fixed point activations can be described by all of these specifications. The activated visual columns resemble those that would be activated by direct perception of the corresponding visual object or scene, the activated auditory columns are a verbal description of the object or scene, the auditory columns have a predominant recommendation strength in favour of activating the auditory columns and vice versa, and the activated columns also have recommendation strengths in favour of a range of other behaviours. However, between the fixed points the content does not correspond with well defined cognitive visual or auditory objects, and the predominant recommendation strengths are in favour of indirect activations and not well defined verbal or other cognitive behaviours.

Self Awareness

Self awareness refers to the ability to create internal representations of self, and manipulate those images in a consistent fashion in a manner which makes changes to future behaviours if appropriate. In the resource limited recommendation architecture model the mechanisms used are similar to those discussed for streams of images The development of the capability can be modelled as follows. Suppose that an entity has developed from experience separate arrays of columns that can discriminate between different visual,

auditory and propioceptic experiences. Suppose that the entity is a small boy called John. John hears his parents say the name "John" when they are directing his attention towards himself. Hence auditory columns activated in response to hearing the word are active at the same time as visual columns activated in response to visual input derived from looking at himself, and propioceptic columns indicating the position of his body. The auditory columns can therefore acquire recommendation strengths in favour of activating visual and auditory columns as if he was paying attention to himself. In other words, a pseudoimage of self viewed from within is generated. Furthermore, sometimes his parents may say "boy" when focussing attention on John, sometimes say "boy" when focussing attention on other boys, and sometimes say "John is a boy". This can lead to the auditory columns activated in response to "John" gaining recommendation strengths in favour of activation of visual columns often active in the past when looking at other boys. In other words, the self image activated by the word "John" includes information which can be regarded as viewing self from the outside. This self pseudoimage capability can then participate in stream of consciousness images as described earlier. For further description of the development and use of this self image capability, see Coward [2005a].

Machine Implementations of Required Information Mechanisms

The proposed model for consciousness is dependent upon a number of information mechanisms, including unsupervised organization of experience into column condition groups, association of columns with behaviours using reward feedback, indirect activation of columns on the basis of past temporally correlated activity, and support of different active populations in the same physical resources by different phases of frequency modulation. There have been various electronic implementations of simplified versions of the architecture that demonstrate that the various information mechanisms can be implemented. These implementations have in general used three layer columns as illustrated in figure 2, and have used software (Smalltalk and C++) models for physiological structures.

Electronic simulations have demonstrated that experience can be organized into column modules, where each column detects a gradually expanding similarity space that is relatively orthogonal to the spaces detected by other columns, in such a way that the column ensemble can discriminate between circumstances with behaviourally different implications [Gedeon et al, 1999; Coward et al 2001; Ratnayake et al 2003]. The ability to associate partially ambiguous columns with behaviours using reward feedback and the capability of imitation to improve the efficiency of reward based learning [Coward 2005] has been demonstrated, including the management of behavioural selection by competition between components corresponding with the different behaviours [Coward et al 2004]. It has also been demonstrated that the gradual expansion of column portfolios means that the architecture does not experience catastrophic interference [Coward et al 2004]. Behaviours have included appropriate responses to objects and groups of objects. Simulations have also demonstrated the effectiveness of indirect activation mechanisms in supporting activation of pseudovisual images in response to verbal inputs, and supporting activation of pseudovisual images of objects often present in the past at the same time as currently perceived objects [Coward 2001]. The capability of the frequency modulation mechanism to implement attention

functions at the physiological level has also been demonstrated [Coward 2004]. The use of a sleep-like process for configuration of provisional conditions has also been implemented, with the expected improvement to the behavioural effectiveness of recorded conditions [Coward 2001].

The primary factor limiting realistic modelling of the phenomena of consciousness is the need to organize a huge body of experience in such a way that the information units (column modules) are shared across many experiences but evolved in such a way that the information about earlier experiences can to a considerable degree be reconstructed. The electronic simulations demonstrate "in principle" capabilities, but in practice a realistic consciousness model requires much larger experience profiles and information handling resources. However, electronic experiments to date demonstrate that all the required information mechanisms can be implemented.

These simulations implement the information processes in virtual machine fashion on a conventional computer. The "hardware" of the brain is effective for implementing the types of primitive information processes required for a complex learning system and current computing hardware is much less efficient [Coward 1990; 2005a]. A way in which these primitive information processes could be implemented much more efficiently using custom integrated circuits is described in Coward [1990]. To design and build a conscious machine with full human capabilities would require both the resource driven architecture and efficient implementation technology. The simulations demonstrate that electronic implementations of the types of information processes needed to support conscious phenomena in a resource driven architecture operate as required.

The phenomena of consciousness result from the specific way in which information derived from experience is organized. This organization is modelled on biology and results in various ways in which information can be accessed. These information access routes would not be available, for example, in systems in which the information was organized into rigorously defined cognitive categories. Following biological models for the organization of information derived from experience will therefore be critical for achieving biological like consciousness.

Comparison with other approaches

Since the 1990s, one thread in the effort to understand human consciousness has been design of machines with analogous cognitive capabilities. Efforts have ranged from architectural design on the conceptual level to quantitative modelling of specific phenomena. Such efforts have included the kernel model of Aleksander [2005], the global workspace architecture implementation [Franklin 2003; Shanahan 2006], the virtual machine approach [Sloman and Chrisley 2003], the forward model approach of Cleeremans [2005] and Haikonen's [2003, 2007] neural architecture.

The major differences between the conscious machine architecture proposed in this paper and other approaches is the primary role assigned to resource constraints in determining architecture, together with the requirement for a strategy by which the machine can bootstrap all of a very wide range of capabilities from experience, with limited and plausible a priori knowledge and ongoing

guidance. Prior publications [Coward 2000; 2001] have presented theoretical arguments, supported by experience with the design of the most complex real time electronic control systems and also by comparisons with the mammal brain, indicating that such practical considerations drive the architectural form of any system which must perform a complex combination of behavioural features with sufficiently limited resources.

In general there are many ways in which a given combination of functional features could be implemented. At one extreme, if every feature could be implemented using separate resources, then changes to one feature would have no effect on any other feature. However, if resources are strongly constrained, individual features cannot be supported by separate such resources. Resources will generally be divided into subsystems that perform different types of information processes where the type for one subsystem will be a set of "similar" processes, with two processes being "similar" if they can both be performed on the same "hardware". A system architecture in which there are subsystems corresponding with individual features is in general not practical if resources are sufficiently constrained.

A resource oriented description of how any one feature works will need to refer to sequences of processes in (generally) all of the different subsystems. This makes descriptions of how features operate in terms of the resource architecture more difficult to understand, and confusing for the user of an electronic system who has no interest in system design. Such a user is therefore provided with a description of how the system works (the user manual) that largely ignores the system resource architecture. This description treats features as independent subsystems except when they interact from the point of view of the user. Coward and Sun [2007] have suggested that a number of consciousness models are of this "user manual" type.

However, if a fully featured system was implemented with separate resources assigned to each feature "subsystem" implicit in a user manual, much of the resource sharing would be lost. In this section it will be argued that many of the alternative models do not give adequate consideration to resource limitations and bootstrapping of capabilities, and often assume the existence of resource modules corresponding with features. Such models could be able to implement effectively a small subset of conscious phenomena (or perhaps a very extensive set if "Moore's Law" continues to make increasing levels of information handling resources available), but if as suggested by Coward [2001; 2005a] the human brain has been subject to natural selection pressures on resources and learning they will be less effective as guides to understanding human consciousness.

Haikonen's Neural Architecture

A central feature of Haikonen's cognitive architecture [Haikonen 2003; 2007] is the percept module. This module receives inputs from one of the senses (visual, auditory etc) and generates an output if a specific cognitive feature (or percept) is detected. An object in the environment is represented by the outputs of a specific combination of modules. In other words, objects are represented as signal vectors, with an element in the vector corresponding with the presence or absence of a particular cognitive feature. Haikonen gives the simple example of a cherry which could be represented by vector 100100100, with the three ones indicating the presence of

red, small and sphere, the zeros indicating the absence of green, blue, medium size, large, cube and cylinder. In more fully distributed representations, properties could be represented by specific combinations of elements.

There are some general similarities between percept modules and column modules in the recommendation architecture, but some fundamental differences. In Haikonen's architecture, percept modules form a single level of sensory processing and one module receives information from only one sensory mode. In the recommendation architecture column modules form a hierarchy that detect sensory circumstances on increasing levels of complexity, eventually including multiple sensory modes. A more fundamental difference is that the sensory circumstances of column modules do not correspond with cognitive features. Rather, a sensory circumstance is a similar sensory condition that may be present in many different cognitive features and objects. The outputs of column modules could be regarded as signal vectors, but there will be many different signal vectors corresponding with objects of one cognitive category, and one vector element will be present in objects of many different cognitive categories. Each element has different recommendation weights in favour of behaviours appropriate to the presence of all the different categories in which it might occur, and the predominant weight across all currently present elements will lead to, for example, a verbal behaviour of naming the appropriate category.

The reason for the lack of correlation between column module similarity circumstances and cognitive features is that the vast majority of the similarity circumstances must be bootstrapped from experience with minimal a priori guidance. As argued in Coward [2001], if information handling resources are limited such a bootstrapping process cannot converge similarity circumstances on to cognitive features without excessive interference between early and later learning. Haikonen does not discuss how percept modules can be bootstrapped from experience, but the arguments of Coward [2001] would indicate that a very high level of resources would be required to generate the range of such modules needed to support consciousness.

A second major aspect of Haikonen's model is that associations between representations of different objects from the same sensory mode or between representations of objects from different sensory modes are supported by neuron modules. Neurons in these modules learn associations between a signal and a number of other signals by repeated coincidences, and can therefore associate representations of objects that occur at the same time. Haikonen's model depends upon repetition for learning. Associations or sequences can be learned provided that they are repeated. Haikonen comments that "Instant learning is susceptible to noise. Random coincidences of signals at the moment of association will lead to false associations" [Haikonen 2007, page 39], and his model does not offer a way to support learning of extensively associated information from a single event, as required, for example, to support robust autobiographic memory of the type observed in human beings.

In the recommendation architecture model, associations of different strengths are created between individual column modules that are active at the same time, rather than between signal vectors present at the same time as in the Haikonen model. Such associations between column modules can be regarded as recommendation strengths in favour of activation of one module if the other is active. Learning of one to one associations is less complex than learning associations between complex combinations of modules, and the

predominant recommendation strengths of the activated modules in one signal vector will tend to indirectly activate an associated signal vector. Semantic type memory based on learning by repeated coincidences is therefore also supported, in a manner that will be more resource effective as the number of required associations between signal vectors becomes very large.

However, in the recommendation architecture model, column modules (unlike percept modules) are constantly evolving by gradual expansion of their similarity circumstances. In a novel experience, there will be such expansions in a wide range of modules. These expansions are instant learning. As discussed earlier, indirect activation of a "signal vector" on the basis of past simultaneous expansions provides the means to generate episodic (autobiographical) memories of unique events, a capability apparently unsupported in the Haikonen architecture.

Global Workspace Approach

In the global workspace model as proposed by Baars [1997], there is a large range of special purpose processors that detect sensory features, generate mental images, imagined feelings, ideas etc. These processors all compete, and the content of consciousness is the results of the processor currently winning the competition. However, if resources are limited, the implication of the arguments in Coward [2001] is that such processors would be required to share resources extensively and they would not correspond with resource modules. It is relevant to note that in the human brain it appears that generating images of specific past events uses largely the same brain areas as generating images of imaginary future events [Addis et al 2007], and when future events are imagined, extensive use is made of material from autobiographic memories [Szpunar et al 2007]. Baars does not discuss how an extensive range of special purpose processors could be bootstrapped from experience.

In the recommendation architecture model for consciousness, column modules are common resources that are activated in support of detection of sensory features, generation of memories and imaginary events etc. A column is directly activated by the presence of its similarity circumstance in current sensory inputs. Each column has recommendation strengths in favour of indirect activation of other columns on the basis of past temporally correlated activity and past temporally correlated information recording. The evolution of a group of columns under the control of its own recommendation strengths could generate an autobiographical memory, but a slightly different starting group could generate an image of an imaginary event. The content of consciousness is the group of currently active columns with the strongest consistent set of recommendations, for example in favour of a speech behaviour. Column modules are therefore specified in such a way that they are resources that can be shared across a wide range of different cognitive processes. There is a clear strategy by which an extensive range of column modules could be bootstrapped from experience.

The difference between the models is therefore that the recommendation architecture based model is consistent with the need to share resources and to learn from experience with limited a priori knowledge and guidance. Because in the global workspace model the implication is that resource modules correspond with externally observed features, it is more analogous with a "user manual" type implementation [Coward and Sun 2007].

The IDA implementation of the global workspace model [Franklin 2003] manages just the assignment of billets to sailors, and does not, for example, need to bootstrap natural language understanding. This implementation thus demonstrates that the architectural concept can work for a limited number of features, but does not address the issue of whether large numbers of features can be learned if resources are limited.

Virtual Machine Model

In electronic systems, virtual machine architectures allow multiple complex and independent information processes to proceed without interference in the same hardware system, even though the different processes make use of the same hardware and have different interfaces to that hardware. A commonplace example is the simultaneous support of multiple user processes on a personal computer, such as word processing, graphics design, email, web access etc. in which the operating system creates a virtual machine for each process that appears to have its own memory and processing resources [Smith and Nair, 1995].

The virtual machine approach to consciousness proposed by Sloman and Chrisley [2003] aims to describe the operations of a conscious entity "by (temporarily) ignoring many of the physical differences between systems and focus[ing] on higher level, more abstract commonalities". A conscious entity is viewed as having a number of independent but interacting mental sub-states at each point in time. These sub-states could be complex processes such as believing something, wanting something, trying to solve a problem, enjoying something, having certain concepts etc. Sub-states are persisting processes that can be modelled as virtual machines analogous with those used by software engineers. Each sub-state depends upon sub-mechanisms of the entity such as perception subsystems, action subsystems, long term memory, short term memory, current store of goals and plans, reasoning subsystems etc. However, in the concept of Sloman and Chrisley, "functionally distinct sub-systems [do not] necessarily map onto physically separable sub-systems", so they also appear to be effectively defined as virtual machines.

In their virtual machine approach, one virtual machine may be processing current visual inputs, while other virtual machines may meta-managing the visual processing virtual machine, for example by monitoring intermediate stages in that visual processing. In this model, qualia are what humans or future human-like robots refer to when referring to these meta-management virtual machines.

In a general sense, the processes of the recommendation architecture described in this paper can be interpreted as supporting an interacting set of virtual machines. In this interpretation, a group of columns which evolves under the influence of its own currently predominant behavioural recommendation strengths can be viewed as one virtual machine. If the recommendations are in favour of indirect activation on the basis of past temporally correlated activity, the machine is supporting a semantic memory process; if the recommendations are in favour of indirect activation on the basis of past temporally correlated information recording, an episodic memory process is supported; if the recommendations are in favour of shifting the direction of vision, an attention process is supported; more complex processes are sequences of activations and so on. Multiple virtual machines of this type can be proceeding at

the same time, and the same column could simultaneously participate in a number of different virtual machines on the basis of recommendation strengths in favour of different behaviours.

For example, a planning process could be initiated in the recommendation architecture model by hearing the question "How would you get from point A to point B?" Hearing "point A" and "point B" activates two sets of auditory columns. These sets of auditory columns in turn indirectly activate two sets of visual columns on the basis of frequent past simultaneous activity. These two pseudovisual sets are therefore similar to the sets that would be activated if the actual locations were viewed. Next, the set corresponding with point A indirectly activates another set of columns that recorded information shortly afterwards in the past. This new set would be similar to the one that would be activated by visual input in a location often visited shortly after point A. Such an indirect activation chain could be continued until the result at some point resembled the set of columns corresponding with point B. The chain of indirectly activated column sets is effectively a plan to get from A to B. At each point the visual columns would also have recommendation strengths in favour of motor behaviours often performed to get between two intermediate points, and recommendation strengths in favour of verbally describing those motor behaviours. A verbal description of the route plan could therefore be generated. Each set of activated columns could be viewed as a virtual machine in Sloman and Chrisley's terms, the difference being that it is clear how these "virtual machines" are supported within a resource effective learning architecture.

At any point in the planning process, one set of activated columns could generate two different indirectly activated sets. For example, this could occur by a bias on recommendation strengths of the original set in favour of one type of indirect activation resulting in one indirectly activated set, followed by a bias on the same set of columns in favour of a different type of indirect activation resulting in a second indirectly activated set. The original set would have all the relevant recommendation strengths, the difference would be a bias placed within competition on how heavily different types of recommendation strengths would be weighted. The frequency modulation mechanism discussed earlier means that both of the different sets could be kept active simultaneously in the same column resources. One set could be the next step in the primary planning process, the other could be a meta-management process monitoring the current step. For example, the meta-management process could lead to a verbal description of the content of the current step. A similar account for qualia can therefore be offered as in Sloman and Shipley's model.

There are two major differences between the virtual machine model as proposed by Sloman and Chrisley and the approach in this paper. Firstly the virtual machine model offered by Sloman and Chrisley provides no account for how the system can bootstrap its capabilities from its experience with minimal a priori guidance. Secondly, although Sloman and Chrisley comment that some virtual machine architectures are harder to implement than others, they do not address the issue of finding a virtual machine architecture that is relatively efficient in resource usage. In electronic systems, virtual machine implementations are often relatively inefficient in use of resources. For example, in general two virtual machine applications on a personal computer cannot act freely upon the same information. Often, duplicate copies of the same information in different formats must be stored in memory (such as diagrams in word

processing and graphics processing applications), and the two applications cannot simultaneously make changes to the same information. The greater the limits placed on resource sharing between virtual machines, the greater the resources required.

A machine that has learned to perform a complex combination of conscious and unconscious behaviours is unlikely to be able to afford the resources required for a true virtual machine implementation, and will require far more interaction between separate processes than is allowed by such an architecture.

In a sense, it is the same user manual-system architecture issue between virtual machine and resource driven architectures. Virtual machine architectures break up system functionality into chunks that can be easily and independently designed and used, at a high cost in resources. The huge growth in memory and processing resources has meant that such an approach is feasible for the problem complexity of many current electronic systems, butis less likely to be feasible for a conscious machine with full human like capabilities.

A further issue is that there will be types of communication between different cognitive processes that will be side effects of the need to share resources, but in some cases these communication mechanisms will support valuable cognitive capabilities. These capabilities will emerge naturally in the resource driven architecture but for the virtual machine architecture will require additional special purpose virtual machines. Thus because modules in the resource driven approach are defined to allow resource sharing across many different cognitive processes, they can simultaneously participate in a number of different such processes and support interaction between those processes. Thus in the planning example above, the same column modules could participate in both the primary planning process and any meta-management processes spun off from the primary process.

Simulation Models

A number of authors have suggested that consciousness can be modelled as internal simulation of future possibilities [e.g. Hesslow 2002] and this proposal has been developed into a cognitive architecture based on the global workspace concept by Shanahan [2006].

In Shanahan's architecture, there is an outer loop that links the environment, sensory processing and behavioural selection. This outer loop is purely reactive, and generates motor responses without cognitive processing. There is also an inner loop that performs simulations which can modulate the outer loop. In addition, the inner loop can perform multiple simulations in parallel, with the different simulations competing to be allowed extensive access throughout the system.

The issue with this architectural proposal is that it is a functional architecture that does not address resource conservation issues. For example, there are very likely to be similarities between the information processes required to perceive the environment and those required for imagining the environment. In a resource driven architecture it would therefore be expected that the outer and inner loops in Shanahan's functional architecture will use many of the same resources to support their operation. The subsystems in the functional architecture will therefore not correspond with modules in the system architecture. It is relevant that in the human brain, direct

perception and cognitive processing appear to use substantially overlapping resources [e.g. Simmons et al 2007]. A critical issue is how the capabilities of the resource modules can be learned and organized so that they can support the multiple functions without interference.

From a "user manual" perspective, simulation is a reasonable way to describe some of the column module operations in the resource driven approach. For example, words describing a possible future situation could directly activate auditory columns, and the auditory columns could indirectly activate visual, propioceptic or polymodal columns on the basis of temporally correlated past activity. The resultant activated column population could then evolve through a series of indirectly activated populations, each activated from the preceding population on the basis of temporally correlated past activity. Such a population sequence would be experienced as a pseudoexperience (i.e. a simulation) of an imagined series of situations, with an implicit averaging across past experience being utilized to effectively derive a plausible future pseudoexperience.

Because populations of columns recommending a behaviour are active shortly before populations of columns activated by sensory experience following the behaviour, groups of columns recommending a behaviour can activate a column population approximating to that which would be activated if the behaviour were performed (again, using averages across past experience as the guide). Columns in this post behaviour population could recommend adjustments to recommendation strengths of recently active columns (i.e. implicit reward feedback). Hence the pseudoexperience can directly effect future behavioural probabilities.

The major differences between this understanding of simulation within the resource driven approach and proposed simulation models such as Shanahan's is that the simulation models may describe the functions of the system but will not give insight into how the resources of the system are organized to support these functions. This insight is fundamental to designing a system. Note that although it is possible to improve the resource efficiency of any architecture, as for example the Connor and Shanahan [2007] improvements to their global workspace based architecture, these improvements will not be adequate as resources become more and more constrained.

The Kernel Architecture

In the kernel architecture proposed by Aleksander [2005], there are four primary system functions that are identified as modules: perception; memory; emotion; and action. As discussed earlier, efficiency will require that resources are shared between, for example, perception and memory. With sufficiently strong limitations on resources, this need for sharing coupled with other practical considerations will force resource modules into specific forms that do not correspond with major system functions. Hence the kernel architecture may be a good functional architecture for modelling consciousness, but implementation of a system that could learn a full range of cognitive capabilities including human consciousness using this functional architecture as the resource architecture would require far more resources that a resource constrained architecture approach.

Forward Models

Cleeremans [2005] has proposed that a class of computational models called forward models may provide a good starting point for exploring the properties of mental representations able to support consciousness. In forward models, there are two interconnected networks. One network receives system objectives and a description of the current state of an environment to be controlled as inputs and generates actions as output. The second network (called the forward component) takes the output of the first network and a description of the current state of the environment as inputs and generates a prediction of how the environment will change if the actions were carried out. Once again, this is a functional architecture, which may be useful for exploring the operation of consciousness but does not take account of the need to share resources across functions.

Conclusions

Prior publications have demonstrated that for a complex learning system, as the ratio of learned behaviours to available resources becomes larger, the system architecture will be more and more constrained within a set of architectural bounds. Many different types of system could be designed to learn a given set of behaviours, but resource requirements will tend to become excessive and the learning process more error-prone for systems designed outside these bounds.

The large set of behaviours that make up human consciousness can be expected to require extensive information handling resources. It is therefore probable that in order to design a system which can learn a full range of conscious behaviours, it will be necessary for the system architecture to take account of resource constraints. An architectural concept within the constraints has been described, which includes an architecture, a demonstration that a range of consciousness related processes can be supported by the architecture, and a demonstration that the key information processes can be implemented. It is also demonstrated that limits on changes during learning imposed by resource limitations result in indirect activation processes that can support different types of memory important to consciousness.

Alternative design proposals for a conscious machine do not take into account the architectural bounds imposed by strong resource constraints. Such proposals could implement features equivalent to human consciousness, but such implementations would be more costly in information handling resources and could have difficulties bootstrapping a full range of cognitive capabilities (including human-like consciousness) from experience with limited a priori knowledge and limited ongoing guidance. Because the human brain has been constrained by resource limitations, approaches that do not take account of the effects of such constraints may be less valuable for understanding human consciousness.

References

Aleksander, I. [2005]. The World in My Mind My Mind in the World: Key Mechanisms. Exeter: Imprint Academic.

Aleksander, I. (2005). Machine Consciousness. Progress in Brain Research 150, 99 – 105.

Baars, B. (1997). In the Theatre of Consciousness. Journal of Consciousness Studies, 4,4, 292 - 309.
Bartlett, M. S., Movellan, J. R. and Sejnowski, T. J. (2002). Face Recognition by Independent Components Analysis. IEEE Transactions on Neural Networks 13(6), 1450-1464.
Bi, G-q and Poo, M-m (1998). Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. Journal of Neuroscience 18(24), 10464 - 10472.
Block, N. (1995). On a confusion about a function of consciousness. Brain and Behavioral Sciences 18, 227 - 287.
Chrisley, R. and Parthemore, J. (2007). Synthetic Phenomenology. Journal of Consciousness Studies 7, 44 - 58.
Cleeremans, A. [2005]. Computational correlates of consciousness. Progress in Brain Research 150, 81 - 98.
Connor, D. and Shanahan, M. (2007). A Simulated Global Neuronal Workspace with Stochastic Wiring. AAAI Symposium on Consciousness and Artificial Intelligence: Theoretical foundations and current approaches.
Coward, L. A. (1990). Pattern Thinking, New York: Praeger.
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., Gedeon, T. and Kenworthy, W. (2001). Application of the Recommendation Architecture to Telecommunications Network Management, International Journal of Neural Systems 11(4), 323-327
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.
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 13(2), 268 - 301.
Coward, L. A. (2004). Simulation of a Proposed Binding Model. Brain Inspired Cognitive Systems 2004, L. S. Smith, A. Hussain and I. Aleksander, (editors), University of Stirling: Stirling.
Coward, L. A., Gedeon, T. D. and Ratanayake, U. (2004). Managing Interference between Prior and Later learning. Lecture Notes in Computer Science 3316, 458-464.
Coward, L. A. (2005a). A System Architecture Approach to the Brain: from Neurons to Consciousness. New York: Nova Scientific Publishers.
Coward, L. A. (2005b). Accounting for episodic, semantic and procedural memory in the recommendation architecture cognitive model. Proceedings of the Ninth Neural Computation and Psychology Workshop: Modelling Language, Cognition, and Action.
Coward, L. A., and Gedeon, T. G. (2005). A model for representation of concepts in the brain. Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning. Espoo, Finland.
Coward, L. A. and Sun, R. (2007). Hierarchical approaches to understanding consciousness. Neural Networks 20(9), 947 - 954.
Franklin, S. (2003). IDA: a conscious artifact ? Journal of Consciousness Studies 10(4-5), 47 - 66.
French, R.M. (1999). Catastrophic Forgetting in Connectionist Networks. Trends in Cognitive Science 3(4), 128-135.
Gedeon, T., Coward, L. A., and Bailing, Z. (1999). Results of Simulations of a System with the Recommendation Architecture, Proceedings of the 6th International Conference on Neural Information Processing, Volume I, 78-84.
Grillner, S., Markram, H., De Schutter, E., Silberberg, G. and LeBeau, F. E. N. (2005). Microcircuits in action - from CPGs to neocortex. Trends in Neurosciences 28(10), 525 - 533.
Haikonen, P. O. (2003). The Cognitive Approach to Conscious Machines. Imprint Academic.
Haikonen, P. O. (2007). Robot Brains: Circuits and Systems for Conscious Machines. New York: Wiley.
Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. Trends in Cognitive Science 6(6), 242 - 247.

Hyvärinen, A., Karhunen, J. and Oja, E. (1999). Independent Component Analysis. New York: Wiley.
James, William. (1892). The stream of consciousness. From Psychology (chapter XI). Cleveland & New York, World.
James, W. (1904). Does 'Consciousness' Exist? Journal of Philosophy, Psychology, and Scientific Methods 1, 477 - 491.
McClelland, J.L., McNaughton, B.L. & O'Reilly, R.C. (1995). Why There Are Complementary Learning Systems in the Hippocampus and Neocortex: Insights From the Successes and Failures of Connectionist Models of Learning and Memory. Psychological Review, 102, 419-457.
Ratnayake, U., Gedeon, T. D. (2003) "Extending The Recommendation Architecture Model for Text Mining", International Journal of Knowledge-Based Intelligent Engineering Systems, 7(3), pp. 139-148.
Robins, A. (1995). Catastrophic Forgetting, rehearsal, and pseudorehearsal. Connection Science 7, 123 - 146.
Shadlen, M. N. and Movshon, J. (1999). Synchrony Unbound: a critical evaluation of the temporal binding hypothesis. Neuron 24, 67 - 77.
Shanahan, M. (2006). A cognitive architecture that combines internal simulation with a global workspace. Consciousness and Cogniton 15, 433 - 449.
Shimizu, T. and Bowers, A. N. (1999). Visual circuits of the avian telecephalon: evolutionary implications. Behavioural Brain Research 98, 183 - 191.
Simmons, W. K., Ramjee, V., Beauchamp, M. S., McRae, K., Martin, A. and Barsalou, L. W. (2007). A common neural substrate for perceiving and knowing about color. Neuropsychologia 45, 2802-2810.
Sloman, A. and Chrisley, R. (2003). Virtual machines and consciousness. Journal of Consciousness Studies 10(4-5), 133 - 172.
Smith, J. E. and Nair, R. (1995). The architecture of virtual machines. Computer 38(5), 32 - 38.
Sourdet, V. and Debanne, D. (1999). The role of dendritic filtering in associative long-term synaptic plasticity. Learning and Memory 6, 422 - 447.
Szpunar, K. K. and McDermott, K. B. (2007). Episodic future thought and its relation to remembering: Evidence from ratings of subjective experience. Consciousness and Cognition (In Press)
Tanaka, K. (2003). Columns for complex visual object features in the inferotemporal cortex: clustering of cells with similar but slightly different stimulus selectivities. Cerebral Cortex 13, 90 - 99.
Treisman, A.M. (1960). Contextual cues in selective listening. Quarterly Journal of Experimental Psychology 12, 242-248.
Von der Malsburg, C. (1981). The Correlation Theory of Brain Function. Internal Report 81-2, Max Planck Institute for Biophysical Chemistry, Gottingen, Germany.
Von der Malsburg, C. (1999). The What and Why of Binding: the modeler's perspective. Neuron 24, 95 - 104.