

Hierarchical Approaches to Understanding Consciousness

L. Andrew Coward
Department of Computer Science
Australian National University
Canberra, ACT 0200, Australia

and

Ron Sun
Cognitive Science Department
Rensselaer Polytechnic Institute
110 Eighth Street, Carnegie 302A
Troy, New York 12180, USA

Abstract

There has been much discussion of what a scientific theory of consciousness would look like, and even whether such a theory is possible. Some common misunderstandings of the nature of theories (e.g., in the physical sciences) have confused the discussion of theories concerning consciousness. Theories in the physical sciences establish hierarchies of descriptions that relate high-level descriptions of macro-level phenomena to detailed-level descriptions at a micro level. Detailed descriptions are usually more accurate but information-dense and therefore often beyond human comprehensibility (unless limited to tiny segments of a macro-level phenomenon). High-level descriptions are usually much less information-dense but more approximate. The ability to map between levels of description, and in particular the understanding of when a shift from a higher-level to a more detailed description is needed to achieve a desired degree of accuracy, is fundamental to an effective theory in any field. The *form* of such a theory of consciousness is sketched, and the limitations of some alternative approaches described.

Introduction

There are a range of possible approaches aimed at understanding consciousness. At one extreme, there has been debate over whether a scientific theory of consciousness is possible in principle [e.g. Chalmers 1995]. Another type of approach has been the creation of models that describe cognitive operations regarded as being conscious, but with limited consideration for how physiology could support the processes required by the model. Computer implementations of some of these models have been attempted [e.g. Franklin et al 1999, Sun 1999]. Yet another approach has been to look for physiological activity that discriminates between conscious and unconscious states, the so-called "neural correlates of consciousness" [e.g. Rees et al, 2002, Crick et al 1998)]. Lamme [2006] claimed that neural activity occurred in different structures depending on the type of conscious or unconscious behaviour, and argued that the presence of neural activity should be part of the definition of whether consciousness was present, independent of cognitive measures. Yet another approach has been to claim that understanding of consciousness is only possible with reference to quantum mechanics [Penrose 1994]. Furthermore, there have been extensive debates over what phenomena should actually be labeled "conscious" [e.g. Block 1995]. The end result has been considerable meta-theoretical confusion.

For several centuries, the physical sciences have been regarded as the paradigm for valid scientific theories. We believe that much of the confusion over scientific understanding of

consciousness derives from misunderstanding of what theories in the physical sciences actually deliver, and once these misunderstandings are cleared away, the form which a theory of consciousness must take becomes clearer. There are some good questions about whether a scientific theory of consciousness is possible, but these questions are in fact more empirical than philosophical, and relate to whether the brain is organized in such a way that understanding can occur within the limits of human information handling capabilities.

There are some significant similarities between theories in the physical sciences and the theoretical techniques by which the understanding of a complex computational system is achieved and maintained. Although there are minimal direct resemblances between brains and complex computational systems, natural selection pressures on brains have tended to have architectural effects which make these techniques relevant and applicable (Coward 2005), thus providing a basis for a scientific theory of consciousness. The general form which such a theory would take can therefore be sketched.

Theories in the physical sciences

The ability of theories in the physical sciences to accurately, quantitatively predict the outcome of experiments has made physics the paradigm for a good scientific theory. In quantum mechanics, the extreme accuracy of such predictions has contrasted with the abstract and sometimes counterintuitive nature of the concepts used. This contrast has sharpened the philosophical debate between those who believe that the predictive power of a scientific theory provides insight into the underlying causal structure of reality (or "natural laws"), and those who would argue that such theories simply provide effective but ultimately approximate descriptions of reality (i.e., realists versus empiricists). In the neurosciences, the counterintuitive nature of quantum mechanics has perhaps influenced the development of the view that "folk psychology", or everyday understanding of human psychology, will have no place in a genuine science of the brain [Churchland 1989].

There is an aspect of the practice of the physical sciences that is often missing both from the philosophy of science and from thinking about the ultimate form of a genuine brain science. This aspect is the existence and use of hierarchies of description, which make it possible to describe causal (and other key) relationships within a phenomenon on many different levels of detail. The higher levels are more approximate than lower (more detailed) levels, but all the different levels of detail are essential to create a comprehensible scientific understanding of the phenomenon. An indispensable part of a scientific theory is the ability to map between different levels of description, including rules to indicate when a transition to a deeper (lower) level is required to achieve a desired degree of accuracy.

In order to illustrate this aspect of the physical sciences, consider how the science of matter can be applied to designing and adjusting the refining process for crude oil. Some key elements of the five-level description hierarchy are summarized in table 1.

	Entities	Number of types of entities	Examples of entity types
Quantum mechanics	Individual elementary "particles"	< 10	Electron, proton, photon
Atomic theory	Individual atoms	~ 100	Carbon atom, hydrogen atom, sulphur atom
Molecular theory	Individual molecules	All possible combinations of atoms	Benzene molecule, isobutene molecule
Chemical theory	Chemicals = large numbers of almost identical molecules	All possible collections of identical molecules in all possible temperatures and states	Different masses of benzene, at different temperatures, in solid, liquid or gas state etc.

Materials theory	Materials = mixtures of chemicals	All possible collections of chemicals	Crude oil
------------------	-----------------------------------	---------------------------------------	-----------

Table 1. A physical sciences hierarchy of descriptions relevant for understanding the refining process for crude oil. Descriptions at a detailed level (such as quantum mechanics) use fewer different types of entities with relatively simple interactions between them, are more mathematically exact, but have an information density that makes them incomprehensible if applied to a complete macro phenomenon. Descriptions at a higher level have more types of entities with complex interactions between them that can be approximated by a simple description, have a much more comprehensible (lower) information density, but are more approximate. However, their accuracy is adequate for many practical purposes, and there are clear rules indicating when a more detailed level is required for adequate descriptive accuracy.

At the materials level, it can be understood that crude oils vary from light to heavy (referring to density and viscosity), and from sour or sweet (reflecting lower temperatures at which corrosion occurs with sour crudes). Even at this level, such understanding can be applied to adjusting the refining process. At the chemical level, the knowledge that crude oils are a mixture of chemicals such as butane (a paraffin), benzene (an aromatic), cyclohexane (a naphthene), butene (an alkene), butadiene (an alkyne), and many other chemicals of these types provides deeper understanding of the differences between light and heavy crudes which may be required to design the process. At the molecular level, such knowledge as that paraffins have the general structure C_nH_{2n+2} while aromatics have the general structure C_6H_6 -hydrocarbon branch(es) is necessary for some process design aspects. At the atomic level, the knowledge that the presence of sulphur atoms causes reactions with iron alloys at lower temperatures is also necessary to address adjustments for sour oils. At the quantum mechanical level, the knowledge that some electrons in the benzene ring are delocalized to the point that they are shared across all the carbon atoms in the ring is also necessary to understand the relative stability of such rings.

Note, however, that any attempt to describe a complete oil refining process in quantum mechanical terms would need to follow the behaviour of perhaps 10^{30} electrons, protons, photons etc. and such a description would be completely incomprehensible to human intellect. In practice, most of the thinking about the process occurs at the materials and chemicals level, but with the awareness of when the descriptions at these levels will become inaccurate and more detailed levels will be required, and of how to shift to more detailed levels. More detailed descriptions can only cover very tiny segments of the entire process (the interactions of a few molecules, for example), but there are also rules (such as statistical mechanics) for how to scale up.

Consider some key properties of this hierarchy. At the most detailed level, there are often relatively few possible types of entity (electron, proton, photon etc.) and relatively few types of interactions between them. At higher levels, there are often many more possible entities (e.g., all possible combinations of entities at more detailed levels) and the interaction between two entities can be very complex (e.g., the sum of all the interactions between their constituent entities). To be useful, higher-level entities must be defined in such a way that (1) one type is present in a range of phenomena, (2) in any one phenomenon the number of higher-level entities is small, and (3) the interaction between two entities depends only on their types. With these conditions, a description of a phenomenon using only the higher-level entities and their interactions has a much lower information density than a description that keeps track of all the detailed entities. The conditions will in general only be met approximately, and the higher-level description will therefore be approximate (in some sense). The key requirement is therefore to define higher-level entities in such a way that the

approximation is adequate for some purposes (e.g., some practical purposes). The route to higher precision is to shift to a description in terms of more detailed entities and their interactions, but at this level of detail only descriptions of small segments of an overall phenomenon will be comprehensible (Coward and Sun 2004).

There are numerous examples of such hierarchies in the physical sciences. A description of the geology of the Earth in terms of quantum mechanics would be incomprehensible. To understand the gross surface features, the appropriate entities are continental plates floating and moving on the mantle. To understand mineral deposits, a more detailed description in terms of fluid mechanics and the ways different minerals solidify and separate with cooling is required. A crystallographic level of description is needed to understand the types and properties of different minerals, shifting to the molecular level to understand why certain chemicals have certain crystalline forms, and to quantum mechanics to understand the electron diffraction results used to determine the details of those forms. The higher levels of description are approximate compared with the precision achievable with quantum mechanics, but they are simpler and useful for their own purposes.

It is relevant to note that even research at the borders of quantum mechanics starts in "purely classical language that ignores quantum probabilities, wave functions and so forth subsequently overlaying quantum concepts upon a classical framework" [Greene 1999]. Thus research at the most detailed level begins at a higher level known to be "incorrect", but with a clear understanding of the points at which a less accurate ("incorrect") description must be mapped into a more detailed, more accurate level, although this "incorrectness" is sometimes a source of discomfort to physicists [Greene 1999].

What are the implications of these considerations for cognitive science and neurosciences? Imagine a situation in which individual neurons could be modeled computationally in complete physiological detail. Suppose that such neuron models could be connected together in a way that replicated the connectivity in the human brain. Then suppose that when the complete model was run, it exhibited higher cognitive functionalities and capabilities. The problem would be that such a model would simply be another incomprehensible cognitive system. The requirement for an effective theory, we believe, is to find a hierarchy of descriptions such that small elements of a psychological description can be mapped into a more detailed level, small elements of that more detailed level into a yet more detailed level, and so on, down to descriptions at the neuron level or below [Sun et al. 2005]. It is accepted that descriptions at higher levels would be often more approximate, but the approximations would be well understood (relatively speaking) and there often would be rules identifying when and how to shift to a more detailed level to achieve a required level of accuracy.

The argument that a science of consciousness may be impossible in principle starts from the position that subjective mental states may not be describable. For example, it may not be possible to describe what it is like to experience the colour red, and the experience may be different for every individual. From the point of view of hierarchies of description, the problem reduces to one of generating a more detailed level of description for the phenomenon that observing the colour red sometimes generates a mental state which can cause, for example, general verbal responses such as "I am having a strong experience of the colour red" but cannot cause verbal responses which describe the content of the experience in detail. It may turn out that the mental states of seeing red at a more detailed level of description in different individuals may be different. Mental states under some other circumstances may be able to generate more detailed verbal responses (in contrast to observing the colour red). This is what a scientific theory of consciousness can deliver.

However, a critical question remains: can such a hierarchy of descriptions be defined for the human mind/brain? It is possible to generate descriptions of cognitive processes at a high

level and achieve a degree of quantitative accuracy in areas such as: semantic memory response times [Rips, Shoben and Smith 1973], or numbers of objects that can be held in working memory [McCarthy and Warrington 1990]. At the other extreme, it is possible to create descriptions of neurons and groups of neurons which can also achieve a degree of quantitative accuracy in areas like modeling the EEG [Benuskova et al, 2001]. The issue is whether it is possible to create the several intermediate levels of description which will be required to link these two extremes (and maybe even beyond).

Support for the existence of such a hierarchy in the human mind/brain can be argued from an analogy with the design of the complex computational systems [Coward 2001]. If a system must perform a large number of different behaviours with limited resources, it will tend to be constrained, within specific architectural forms, by a number of practical considerations. For computational systems, these considerations include the need to change and add system features/functionalities without undesirable side effects on other features/functionalities (modifiability), the need to build the system without excessive errors (constructability), and the need to diagnose and correct damage and component failures (repairability). Such practical considerations constrain system architectures into the form of a modular hierarchy with very similar properties to the hierarchies of description used in the physical sciences [Coward 2005]. The word "module" is used to describe units in these hierarchies, but there are major qualitative differences between these modules [see e.g. Kamel, 1987; Coward 2005] and modules as defined in the cognitive sciences [e.g. Fodor 1983; Hirschfeld and Gelman 1994; Karmiloff-Smith 1992]. In particular, the primary driving force on module definition here is the need to conserve resources. Modules are therefore groups of physical resources customized to perform sets of similar information handling processes relatively efficiently. There may be some lesser degree of similarity between processes performed by a group of different modules. If the degree of similarity is large enough, some resource sharing may be possible between the modules. Such resource sharing defines an intermediate-level module, and even higher-level modules can be defined by resource sharing across intermediate-level modules and so on. Resource economy thus dictates the existence of a modular hierarchy.

A change to a system feature or functionality will in general require changes to at least one module at a certain level. The need to make such changes without creating excessive undesirable side effects on other system features/functionality means that information exchange between modules at the same level must be minimized as far as possible (i.e. most interactions must be within a module and a relatively small proportion between modules). The need to diagnose and repair failures leads to a similar requirement for minimization of information exchange. The need for a simple system construction process making as few errors as possible means that modules at the same level will be fairly comparable to each other. The result of the combination of resource constraints, modifiability, repairability, and constructability is thus a modular hierarchy. This modular hierarchy, in effect, can function as a hierarchy of descriptions, in which high-level descriptions of a feature/functionality can be created based only on high-level modules, but with the possibility of shifting to descriptions based on more detailed modules (for example, for greater accuracy), along with a way of making a shift amongst descriptions at different levels (Coward and Sun 2004).

Similar processes may be required by a number of different system features/functionalities, and one system feature/functionality will require processes performed by many different modules. Modules may therefore not correspond directly with features/functionalities (and are therefore qualitatively different from the kinds of modules as often defined in cognitive science). (In computational systems, this lack of correspondence gives rise to the well known disconnect between the user manual and the system architecture for the same system.)

In the case of the mind/brain, it is plausible that natural selection pressures would have favoured mind/brains that can perform a large set of functionalities with few resources. Resource constraints therefore likely exist. Natural selection will also favour mind/brains that can learn with the least accidental interference between prior and later learning, that can be built from DNA "blueprints" with few errors, and that can recover from component failures and damage. There are therefore pressures exerted by natural selection on mind/brains that are analogous with the resource limitation, modifiability, constructability, and repairability considerations for computational systems, making it probable that analogous modular hierarchies in human mind/brains do exist, and furthermore, such modular hierarchies can in effect operate as hierarchies of description as mentioned before (see also Coward and Sun 2004, Sun et al 2005, Sun 2002).

Modular hierarchies in the brain

What would indicate the existence of a (resource-based) modular hierarchy in the mind/brain analogous with those observed in complex computational systems? Firstly, a separation of the brain into a number of major physiological structures, with a tendency for much more connectivity within a structure than between structures, and further separations within each major structure on the same basis, and so on. Secondly, some general similarities between peer structures, but significant detailed differences. Thirdly, lack in general of direct correspondence between performance of a behavioural or cognitive function and modules on any level. Any cognitive process would normally require the participation of many structures, and with few exceptions any one structure would participate in many functionalities.

The human mind/ brain does appear to exhibit this type of organization. At the highest level, there is a separation between cortex, basal nuclei, thalamus, hypothalamus etc. The cortex is separated into Brodmann areas which have some general similarities but detailed cytoarchitectural, neurochemical and connectivity differences [e.g. Fatterpekar, 2002]. All cortex areas are separated into microcolumns [e.g. Mountcastle 2003], and microcolumns are separated into layers. Again, there are general similarities but detailed differences in these substructures. Subcortical nuclei like the striatum are separated into subnuclei like the putamen, the ventromedial caudate and the dorsolateral caudate etc. on the basis of connectivity patterns [e.g. Alexander et al. 1986], and at a more detailed level into a patch and matrix structure, again on the basis of connectivity and neurochemical patterns [Goldman-Rakic, 1982]. On every level, these modules have a significant degree of similarity but can be distinguished on the basis of cytoarchitecture, neurochemistry, and/or connectivity. In general, these modules have more internal connectivity than external.

As observed by many fMRI and PET studies, any one cognitive process involves participation by many different physiological structures. For example, imagined navigation involves a wide range of brain structures including "occipitotemporal areas, medial parietal cortex, posterior cingulate cortex, parahippocampal gyrus" [Maguire et al 1997], the hippocampus proper [Maguire et al 1997; Voermans, 2004] and the caudate nucleus [Voermans, 2004]. Even when a structure appears to be specialized for one type of stimuli, such as the fusiform gyrus for face processing [Poirtois et al. 2005], the relative activities of the fusiform gyrus are different during different types of face related processing [Zeineh et al. 2003]. Other structures are also involved, including extensive regions of the ventral extrastriate temporal cortex, the lateral occipital cortex, and left inferior frontal gyrus [Poirtois et al. 2005] and the prefrontal cortex and hippocampus [Zeineh et al. 2003]. The fusiform gyrus is also sometimes involved in non-face processing tasks, for example arithmetic processing [Dehaene et al. 2004]. It is of particular significance that cortical areas specific to face processing are recruited for certain other types of visual expertise. Experts in

distinguishing between different car models or different species of birds show a degree of recruitment of specific face processing areas that is proportional to their degree of expertise [Gauthier et al, 2000]. Thus these areas appear to be specialized in a particular type of information processing (distinguishing between large numbers of different but visually similar objects) rather than a particular stimulus type. At the microcolumn level in the cortex, columns do not correspond with high-level cognitive functionalities or high-level cognitive categories. For example, even in the visual area TE, which plays an important role in visual object recognition, columns correspond with abstract visual shapes not clear object categories, and one column may be activated in response to objects of different types [Tanaka, 2003].

Thus physiological evidence supports the theoretical arguments that the brain has a modular hierarchical structure, which could be used as the basis for a hierarchical scientific theory of consciousness.

General form of a scientific theory of consciousness

Given the understanding of the form of theories in the physical sciences and what such theories actually deliver, and the arguments in favour of modular hierarchies in the human mind/brain, the general *form* of a scientific theory of consciousness may be sketched.

Firstly, it would involve causal descriptions, at a psychological level, of some phenomena generally regarded as being conscious.

To make some causal relationships implicit in, for example, the Block definition of consciousness [1995] explicit, consider the following scenario 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 in a way which minimizes the stress on the sore ankle. A second behaviour is uttering a warning "mind the tree" to the companion. A third behaviour is to make the comment "Do you see that tree? It's a Hemlock ". The first behaviour can generally be generated unconsciously. Sensory input from the tree and the sore ankle etc. generates some mental state internal to the brain which leads to avoidance behavior but does not lead to higher cognitive functions or verbal report. The third behaviour falls within the definition of access consciousness. The internal mental state in response to the tree (or "representation") leads to both cognitive processing and a verbal report. The second behaviour is of interest because it appears to indicate that simple verbal warnings can be initiated by an unconscious activation without a great deal of cognitive processing. The entities and causal relationships suggested by these scenarios could be visual input from a tree, pain input from the ankle, "unconscious" mental activation, "conscious" mental activation, avoidance behaviour, verbal warning behaviour, higher cognitive behaviour (including associative thinking, and verbal reports of associative thinking). Sensory input can cause an unconscious and/or a conscious activation. An unconscious activation can cause avoidance behaviour and simple verbal warnings. A conscious activation can cause avoidance behaviour and simple verbal warnings, and can also cause higher cognitive processing and complex verbal reports of that processing. A conscious activation can also lead to explicit memory (which is less often the case for an unconscious activation).

Secondly, a theory would involve identification of the different types of processing performed by different physiological structures such as the cortex, thalamus, basal nuclei, etc, and the differences between the processing performed by substructures of these structures. The expectation would be that there would be more similarity between the types of processing performed by different cortex modules than between a cortex module and a thalamus module for example. The causal relationships at the psychological level would then be mapped into sequences of combinations of such processes. The descriptions at the highest level would be

approximate, but a comprehensible description of an end-to-end conscious process would be possible. This description could be mapped through a series of more accurate, more detailed levels down to neuronal processes, although the information density would increase rapidly, and only short segments of such descriptions would be comprehensible.

Thirdly, an important issue is whether it is possible to create the needed hierarchy of intermediate-level descriptions. The earlier discussion indicates that the combination of resource constraint, modifiability, repairability, and constructability requirements tends to result in a modular hierarchy with the right properties to support such a description hierarchy (Coward and Sun 2004, Sun et al 2005). A description of a conscious behaviour in terms of high-level modules would be comprehensible but approximate for a macro behaviour. However, there would exist a path for describing various small components of the behaviour to any desired degree of detail, with the use of the lower-level modules at the various intermediate levels of the description hierarchy.

The critical point is that some degree of inaccuracy will be inherent in the higher levels of descriptions of consciousness, but this is not necessarily a failure of the science. For one thing, it is present also in the physical sciences.

Approaches to understanding consciousness based on delineation and separation of subsystems that perform different types of information processes (thus forming intermediate-level descriptions) are in actuality possible. One example is the CLARION model, which separates the information processes of the mind/brain into two types with localist and distributed representations respectively and consequently different learning and decision processes are involved with these two types of representation [Sun 1999, Sun 2002]. Further divisions of subsystems are then made, leading up to a (partial) hierarchy. Another is the recommendation architecture, which is divided into two subsystems, one defining and detecting conditions in the information available to the system, the other interpreting conditions as behavioural recommendations, and adjusting recommendation weights on the basis of rewards [Coward, 2001; 2005]. Both models can provide descriptions of “conscious” processes at levels intermediate between high-level psychological descriptions and detailed descriptions in terms of neuronal processes (Sun et al 2005).

Issues with different approaches

A scientific theory of consciousness must be based on an understanding of what a scientific theory should be like and what it actually delivers. The present paper has proposed that a scientific theory of consciousness will have some critical qualitative characteristics, based on the properties of theories in the physical sciences (and the ways in which understanding is possible in complex computational systems).

Firstly, it will be made up of a hierarchy of causal descriptions. At a high (e.g., psychological) level there will be many different types of descriptive entities, a relatively low information density in the descriptions, and a relatively high degree of approximation. At a detailed (e.g., physiological) level there will be few different types of descriptive entities, a higher level of information density, and a much lower degree of approximation. There will be a number of intermediate levels of description with intermediate numbers of types of entities, information densities and degrees of approximation.

This hierarchy of description is necessary if understanding of a very complex system is to be possible within human mental capabilities. Each level can describe a cognitive phenomenon in its own terms, but the differences in information density mean that although a description of a complete psychological phenomenon at the highest level would be comprehensible, only descriptions of small segments of that phenomenon will be comprehensible at more detailed levels. It must be possible to map descriptions between

levels, and there must be a clear understanding of when translation to a deeper level is necessary to achieve a required degree of accuracy.

Secondly, a number of practical needs/constraints tend to result in the resources of the brain being organized into a modular hierarchy, and modules at different levels in this hierarchy have the properties of entities at different levels in a hierarchy of descriptions. In complex computational systems it is this modular hierarchy that makes human understanding possible, and an equivalent hierarchy in the brain would enable some scientific understanding of consciousness. Resource constraints mean that modules will not in general correspond with features as perceived by an outside observer.

The question of the possibility of such a theory centres around whether an appropriate hierarchy of descriptions can be constructed within the limits of human intellectual capabilities. The analogies with the structure of complex computational systems indicate that such a hierarchy of descriptions of consciousness may be possible.

From this perspective, there are some common problems with some existing theories. One is that the approximate nature of higher-level descriptions is not recognized, and another is that attempts are made to match modules with superficial cognitive features rather than deeper (resource-driven) types of information processes. A few very brief comments about some existing approaches would be as follows, in relation to our discussions above:

(1) The approximation inherent in higher-level descriptions means that the wholesale exclusion of approaches like "folk psychology" is not necessarily appropriate, because such approximate descriptions might sometimes be, in some ways, analogous with the higher levels of description in the physical sciences.

(2) The issue with computational/mathematical modeling without reference to lower-level (e.g., physiological) structures at all is that such modeling is analogous to trying to directly implement the user manual for a complex computational system. Such an implementation would be possible in principle, given unlimited information handling resources, but the result would sometimes bear little resemblance to the actual system architecture, and any system features/functionalities not explicitly addressed in the user manual ("accidental capabilities") would in general not be present in the implemented user manual.

For example, in the global workspace model (Baars 1997), the contents of consciousness are the contents of a global workspace located in the primary sensory projection areas of the cortex (e.g. VI for the visual sense). A set of input processors generate visual images, inner speech etc. which compete through structures like the brainstem reticular formation and the nucleus reticularis for access to the global workspace. The contents of the global workspace are distributed widely to unconscious specialty processors including perceptual analyzers, output systems, action systems, syntax systems, and planning and control systems. In Baars' model, "the overall function of consciousness is to provide very widespread access to unconscious brain knowledge including autobiographical memory ... ; the lexicon of natural language ; automatic routines that control actions ; and, by way of sensory feedback, even the detailed firing of neurons and neuronal populations".

A problem with Baars' model is that it is essentially a user manual type description which utilizes components that correspond with cognitive functions (like perceptual analyzers, output systems, action systems, planning and control systems). Such models may be helpful for organizing easy-to-understand descriptions of the phenomena of consciousness at the psychological level, but will have little relationship with how those phenomena are supported in a system architecture which is organized to make effective use of information handling resources.

(3) The limitation to searching for some physiological activity that correlates with consciousness is that it can be anticipated that any one physiological structure will participate

in a wide range of functionalities (as discussed earlier). It is likely not the activity of one structure that will correlate with consciousness, but a specific combination and sequencing of module activities. Such an activity combination and/or sequencing would define the presence of conscious behaviour, but unlike the proposal of Lamme [2006], there might be consistency between psychological and physiological measures.

(4) The invocation of quantum mechanics is almost equivalent to an argument that no hierarchy of descriptions exists that can bridge the gap. Given that biochemistry may be understood at a level of description higher than quantum mechanics, there seems to be no clear reason why understanding of conscious phenomena would require descriptions exclusively at such a low level, except in the same sense that all sciences will require quantum mechanics as the degree of required precision increases.

(5) How does this relate to the so called “hard problem”: “Why it is that when our cognitive systems engage in visual and auditory information processing, we have visual or auditory experience: the quality of deep blue, the sensation of middle C? How can we explain why there is something it is like to entertain a mental image, or to experience an emotion?” [Chambers 1995]. This is not exactly a scientific question within the current scope of science, although a scientific understanding will probably reduce the sense of mystery around the questions by giving a better perspective on the issue.

For example, an experience of the colour deep blue at the phenomenological level would at the physiological level be the state of a very large number of neurons. Each of these neurons will have had a developmental process unique to the individual having the experience, because of the unique sensory experience of that individual. Many of these neurons will have behavioural implications (such as saying “that is deep blue”), but because of resource sharing will typically have many behavioural implications, and again all of these behavioural implications will be specific to the individual. These behavioural implications could include accessing a range of memories or emotional states. Which behavioural implications will be followed will depend on the exact nature of the sensory stimulus, the prior state of psychological/physiological activity, and past sensory history (as recorded by the neurons). Thus a physiological understanding implies that the response to experiencing deep blue will be individual-specific and could lead to a wide range of (again individual-specific) behavioural responses (with the actual response being very sensitive to small variations in inputs, etc.).

Although the present approach does not yet answer the question “Why is it like what it is?”, it does provide another way of looking at and understanding the way in which conscious experience is defined and evolves. This is what a scientific theory currently delivers.

Conclusions

An argument has been made here that a scientific theory of consciousness analogous with theories in the physical sciences is likely to be possible, but the form of that science may be different from some existing (and often widespread) expectations because of the misunderstanding of what a theory in the physical sciences actually delivers. The rough general form of an eventual theory of consciousness has been outlined (to a very limited extent, of course), and issues with a few of the many approaches to understanding consciousness briefly identified.

References

Alexander, G. E., DeLong, M. R. and Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience* 9, 357 - 381.

- Baars, B. (1997). In the Theatre of Consciousness. *Journal of Consciousness Studies*, 4, 4, 292 - 309.
- Benuskova L, Kanich M and Krakovska A. (2001) Piriform cortex model of EEG has random underlying dynamics. In: Proc. World Congress on Neuroinformatics. F. Rattay (Ed), ARGESIM/ASIM-Verlag, Vienna, 287-292. ISBN 3-901608-20-6.
- Block, N. (1995). On a confusion about a function of consciousness. *Brain and Behavioral Sciences* 18, 227 - 287.
- Chalmers, D. J. 1995. Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2, 200-19.
- Churchland, P. (1989). *A Neurocomputational Perspective*. Cambridge, MA: MIT Press.
- 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. (2005). *A System Architecture Approach to the Brain: from Neurons to Consciousness*. New York: Nova Science Publishers.
- 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.
- Crick, F. C. and Koch, C. (1998). Consciousness and neuroscience. *Cerebral Cortex* 8, 97-107.
- Dehaene, S., Molko, N., Cohen, L. and Wilson, A. J. (2004). Arithmetic and the brain. *Current Opinion in Neurobiology* 14, 218-224.
- Fatterpekar, G. M., Naidich, T. P., Delman, B. N., Aguinaldo, J. G., Gultekin, S. H., Sherwood, C. C., Hof, P. R., Drayer, B. P. and Fayad, Z. A. (2002). Cytoarchitecture of the Human Cerebral Cortex: MR Microscopy of Excised Specimens at 9.4 Tesla. *American Journal of Neuroradiology* 23, 1313-1321.
- Fodor, J. A. (1983). *Modularity of Mind*. Cambridge, MA: MIT Press.
- Franklin, S. and Graesser, A. (1999). A Software Agent Model of Consciousness. *Consciousness and Cognition* 8, 285 - 305.
- Goldman-Rakic, P. S. (1982). Cytoarchitectonic heterogeneity of the primate neostriatum: subdivision into island and matrix cellular compartments. *Journal of Comparative Neurology* 205, 398 - 413.
- Greene, G. (1999). *The Elegant Universe*. Norton.
- Gauthier, I., Skudlarski, P., Gore, J. C. and Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature* 3(2), 191- 197.
- Hirschfeld, L. and Gelman, S. (1994). *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge University Press.
- Karmiloff-Smith, A. (1992). *Beyond Modularity*. MIT Press.
- Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Science* 10(11), 494 – 501.
- Maguire, E. A., Frackowiak, R. S. J. and Frith, C. D. (1997). Recalling routes around London: Activation of the right hippocampus in taxi drivers. *Journal of Neuroscience* 17(18), 7103 - 7110.
- McCarthy, R. A. and Warrington, E. K. (1990). *Cognitive Neuropsychology: A Clinical Introduction*, (Chapter 13: Short-Term Memory). San Diego: Academic Press.
- Mountcastle, V. B. (2003). *Cerebral Cortex* 13, 2 - 4.
- Penrose, R. (1994). *Shadows of the Mind*. Oxford University Press. Oxford, UK.

- Pourtois, G., Schwartz, S., Seghier, M. L., Lazeyras, F. and Vuilleumier, P. (2005). Portraits or People? Distinct Representations of Face Identity in the Human Visual Cortex. *Journal of Cognitive Neuroscience* 17:7, 1043-1057.
- Rees, G., Krieman, G. et al. (2002). Neural Correlates of Consciousness in Humans. *Nature Reviews* 3, 261 - 270.
- Rips, L., Shoben, J. and Smith, E. Semantic Distance and Verification of Semantic Relations. *Journal of Verbal Learning and Verbal Behaviour* 12, 1 - 20.
- Stensaas, S. S., Eddington, D. K. and Dobelle, W. H. (1974). The topography and variability of the primary visual cortex in man. *Journal of Neurosurgery* 40, 747 – 755.
- Sun, R. (1999). Accounting for the computational basis of consciousness: A connectionist approach. *Consciousness and Cognition* 8, 529-565.
- Sun, R. (2002). *Duality of the Mind*. Lawrence Erlbaum Associates: Mahwah, NJ.
- Sun, R., Coward, L. A. and Zenzen, M. J. (2005). On Levels of Cognitive Modeling. *Philosophical Psychology* 18(5), 613 - 637.
- 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.
- Voermans, N. C. et al. (2004). Interaction between the human hippocampus and the caudate nucleus during route recognition. *Neuron* 43, 427 - 435.
- Zeineh, M. M., Engel, S. A., Thompson, P. M. and Bookheimer, S. Y. (2003). Dynamics of the hippocampus during encoding and retrieval of face-name pairs. *Science* 299, 577 - 580.