seminars in THE NEUROSCIENCES, Vol 2, 1990 : pp 263-275

Towards a neurobiological theory of consciousness

Francis Crick and Christof Koch

Visual awareness is a favorable form of consciousness to study neurobiologically. We propose that it takes two forms: a very fast form, linked to iconic memory, that may be difficult to study; and a somewhat slower one involving visual attention and short-term memory. In the slower form an attentional mechanism transiently binds together all those neurons whose activity relates to the relevant features of a single visual object. We suggest this is done by generating coherent semi-synchronous oscillations, probably in the 40-70 Hz range. These oscillations then activate a transient short-term (working) memory. We outline several lines of experimental work that might advance the understanding of the neural mechanisms involved. The neural basis of very short-term memory especially needs more experimental study.

Key words: consciousness / attention / binding / 40 Hz oscillations / short-term memory

IT IS REMARKABLE that most of the work in both cognitive science and the neurosciences makes no reference to consciousness (or 'awareness'), especially as many would regard consciousness as the major puzzle confronting the neural view of the mind and indeed at the present time it appears deeply mysterious to many people. This attitude is partly a legacy of behaviorism and partly because most workers in these areas cannot see any useful way of approaching the problem. In the last few years several books have appeared[1-4] that address the question directly but most of these[1-3] have been written largely from a functional standpoint and so have said rather little about neurons and other machinery of the brain.

We suggest that the time is now ripe for an attack on the neural basis of consciousness. Moreover, we believe that the problem of consciousness can, in the long run, be solved only by explanations at the neural level. Arguments at the cognitive level are undoubtedly important but we doubt whether they will, by themselves, ever be sufficiently compelling to explain consciousness in a convincing manner. Attempting to infer the internal structure of a very complex system using a 'black-box' approach (i.e. manipulating the input variables while observing the output of the system) will never lead to unique answers. In short, such methods are not by themselves powerful enough ever to solve a problem, though they are good enough to suggest tentative solutions.

Our basic idea is that consciousness depends crucially on some form of rather short-term memory and also on some form of serial attentional mechanism. This attentional mechanism helps sets of the relevant neurons to fire in a coherent semi-oscillatory way, probably at a frequency in the 40-70 Hz range, so that a temporary global unity is imposed on neurons in many different parts of the brain. These oscillations then activate short-term (working) memory. In the later parts of the paper we shall be mainly concerned with visual awareness.

Before approaching the problem in detail, it seems sensible to describe our general approach to consciousness and to decide what aspects of it are best left on one side.

Prolegomenon to the study of consciousness

We make two basic assumptions. The first is that there is something that requires a scientific explanation. There is general agreement that we are not conscious of all the processes going on in our heads, though exactly which might be a matter of dispute. While we are aware of many of the results of perceptual and memory processes, we have only limited access to the processes that produce this awareness (e.g. "How did I come up with the first name of my grandfather?"). In fact, some psychologists[5] have argued that we have only very limited introspective access to the origins of even higher order cognitive processes. It seems probable, however, that at any one moment some active neuronal processes correlate with consciousness, while others do not. What are the differences between them?

From the Salk Institute, 10010 N Torrey Pines Road, La Jolla, CA 92037, USA, and Computation and Neural Systems 216-76, California Institute of Technology, Pasadena, CA 91125, USA

1044-5765/90/0204-0003$5.00/0

The second assumption is tentative: that all the different aspects of consciousness, for example pain and visual awareness, employ a basic common mechanism or perhaps a few such mechanisms. If we understand the mechanisms for one aspect, we will have gone most of the way to understanding them all. Paradoxically, consciousness appears to be so odd and, at first sight, so difficult to understand that only a rather special explanation is likely to work. The general nature of consciousness may be easier to discover than more mundane operations, such as shape-from-shading, that could, in principle, be explained in many different ways. Whether this is really true remains to be seen.

The following topics will be left on one side or our attitude to them stated without further discussion, for experience has shown that otherwise much time can be frittered away in fruitless argument about them.

1. Everyone has a rough idea of what is meant by consciousness. We feel that it is better to avoid a precise definition of consciousness because of the dangers of premature definition. Until we understand the problem much better, any attempt at a formal definition is likely to be either misleading or overly restrictive, or both.

2. Arguments about what consciousness is for are probably premature, although such an approach may give a few hints about its nature. It is, after all, a bit surprising that one should worry too much about the function of something when we are rather vague about what it is.

3. We shall assume that some species of animals, and in particular the higher mammals, possess some of the essential features of consciousness, but not necessarily all. For this reason, appropriate experiments on such animals may be relevant to finding the mechanisms underlying consciousness.

3.1. From this it follows that a language system (of the type found in humans) is not essential for consciousness. That is, one can have the key features of consciousness without language. This is not to say that language may not enrich consciousness considerably.

3.2. We consider that it is not profitable at this stage to argue about whether 'lower' animals, such as octopus, *Drosophila* or nematodes, are conscious. It is probable, though, that consciousness correlates to some extent with the degree of complexity of any nervous system.

4. There are many forms of consciousness, such as those associated with seeing, thinking, emotion, pain and so on. We shall assume that self-consciousness, that is the self-referential aspect of consciousness, is merely a special case of consciousness and is better left on one side for the moment. Volition and intentionality will also be disregarded and also various rather unusual states, such as the hypnotic state, lucid dreaming and sleep walking, unless they turn out to have special features that make them experimentally advantageous.

5. No neural theory of consciousness will explain everything about consciousness, at least not initially. We will first attempt to construct a rough scaffold, explaining some of the dominant features and hope that such an attempt will lead to more inclusive and refined models.

6. There is also the problem of qualia.[6] Some argue that certain aspects of consciousness (such as whether the red I see is the same as the red you see), being essentially private, cannot in principle be addressed by any objective, scientific study. We feel that this difficult issue is, for the moment, best left on one side. Our present belief is that it *may* prove possible, in the fullness of time, to make it plausible that you see red as I do (assuming that psychophysical tests suggest you do). To decide whether one *can* make a plausible case or not we shall need to know the exact neural correlate in a human brain of seeing red. Whatever the outcome, we believe that any adequate theory of consciousness should explain how we see color at all.

This outlines the framework within which we will address the problem of consciousness. We next ask what psychology can tell us about the phenomenon.

The cognitive approach

The most effective way to approach the problem of consciousness would be to use the descriptions of psychologists and cognitive scientists and attempt to map different aspects of their models onto what is known about the neuroanatomy and neurophysiology of the brain. Naturally we have attempted to do this but have not found it as useful as one might hope, although such models do point to the importance of attention and short-term memory and suggest that consciousness should have easy access to the higher, planning levels of the motor system.

A major handicap is the pernicious influence of the paradigm of the von Neumann digital computer. It is a painful business to try to translate the various boxes labeled 'files', 'CPU', 'character buffer', and

so on occurring in psychological models, each with its special methods of processing, into the language of neuronal activity and interaction. This is mainly because present-day computers make extensive use of precisely-detailed pulse-coded messages. There is no convincing evidence that the brain uses such a system and much to suggest that it does not. For these reasons we will not attempt to comment in detail on these psychological models but only make rather general remarks.

Johnson-Laird[1] proposes that the brain is a complex hierarchy of largely parallel processors—an idea that is almost certainly along the right lines—and that there is an operating system at the top of the hierarchy. He considers the conscious mind to correspond to this operating system. Thus the mechanism of consciousness expresses the results of some of the computations the brain makes but not the details of how they were done. If there is an operating system of this type it is not easy at the moment to see any particular brain area in which it is located.[7] Johnson-Laird also lays emphasis on self-reflection and self-awareness, topics that we have decided to leave on one side.

An alternative view, due to Jackendoff[2] is that consciousness is not associated with the highest levels in the hierarchy but with the intermediate levels. He arrives at this point of view by considering in detail the language system, the visual system and the music system. His arguments are certainly suggestive, though not completely compelling.

Both authors emphasize the intimate relationship between consciousness and working memory. They also both envisage a serial process (that we may loosely identify as attention) on top of the parallel processes. Thus clearly we shall have to consider the mechanisms of attention and of working memory.

The type of memory that will be of most interest to us is short-term memory, whether it be very short (a fraction of a second; sometimes called 'iconic' memory[8] if it is visual) or short-term memory proper (meaning a few seconds) sometimes called working memory.[9,10] These are so important that we shall take up each of these subjects later. It does not seem essential for consciousness that the brain should be able to put anything into the long-term episodic memory system,[11] since people with certain brain damage (to be discussed below) cannot lay down new episodic memories but appear in all other respects to be fully conscious. Procedural memory,[11] i.e. that part of memory responsible for highly automated procedures, such as typing or swimming, is probably largely unconscious.

Consciousness appears to be a process that is fairly immediate. In other words, it does not take too long to become conscious of, at least, straightforward things. Exactly how long is not clear, but figures somewhere in the 50-250 ms time range might be reasonable, rather than a few seconds.

General neurobiological aspects

When

When is an animal conscious? This turns out not to be a straightforward problem. It seems likely that the essential features of consciousness are probably not usually present in slow-wave sleep, nor under a deep anaesthetic. Rapid Eye Movement (REM) or dreaming sleep is another matter. It seems to us that a limited form of consciousness often occurs in REM sleep. Although cognition is not completely normal and memories cannot be put into the long-term store, nevertheless dreams seem to have some of the attributes of consciousness. Whether there are any experimental advantages in studying REM sleep to help us understand consciousness remains to be seen. At the moment we do not see any.

The really difficult case is that of rather light anaesthesia or states induced by modern receptor-specific anaesthetic agents.[12] It seems likely that in some such states the brain is at least partly conscious. This would not matter were it not for the large amount of experimentation on mammals in these states. We return to this problem later.

Where

Where in the brain are the neural correlates of consciousness? One of the traditional answers—that consciousness depends on the reticular activating system in the midbrain—is misleading. Certainly the relevant parts of the brain need to be activated in various ways, but this is a little like believing that the characteristic part of a television set is its electrical supply. It is more likely that the operations corresponding to consciousness occur mainly (though not exclusively) in the neocortex and probably also in the paleocortex, associated with the olfactory system, since local damage to the cerebral cortex often removes particular aspects of consciousness (as, for example, in face agnosia).

The hippocampal system (the allocortex) is a little more complicated. A person with complete bilateral damage to the hippocampal system and to all the higher order association cortices that are connected with it, appears to have most aspects of consciousness[13] (including short-term conjunctions over at least three modalities) so the hippocampal system is unlikely to be essential. But it could be argued that what goes on in the normal hippocampus does indeed reach consciousness, so it may be important to consider in detail the inputs and outputs of the hippocampal system.

Structures in the midbrain or hindbrain, such as the cerebellum, are probably not essential for consciousness. It remains to be seen whether certain other structures, such as the thalamus, the basal ganglia and the claustrum, all intimately associated with the neocortex, are closely involved in consciousness. We shall include these structures together with the cerebral cortex as 'the cortical system'.

Split brains

The study of persons with 'split brains' gives us information about some of the pathways that are (or are not) involved in consciousness. It is a well-established fact[14] that for persons whose corpus callosum—the massive fiber bundle connecting the two cortical hemispheres—has been cut, the left-hand side of the brain (for right-handed people) is not aware of the activity in the visual system taking place on the right side, whereas in a normal person it is. (We leave on one side the somewhat controversial matter as to whether the right side, on its own, is as conscious as the left side.) This strongly suggests that some of the information associated with consciousness can traverse the normal corpus callosum. It also suggests that such information, with the exception of some emotional states, cannot be transmitted from one side of the cortex to the other by the subcortical pathways that remain intact in this operation.

Blindsight

It now seems generally agreed that blindsight is a genuine phenomenon.[15] Certain people with cortical blindness can point fairly accurately to the position of objects in their blind visual field, while denying that they see anything. It was previously suspected that the neural pathway was subcortical, through the superior colliculus, but this has recently been brought into question.[16] The motor output appears to be voluntary and the subject is certainly aware what movements he is making. It is now an urgent matter to decide experimentally, by comparative work on humans and monkeys, exactly which neural pathways are used in blindsight, since this information may suggest which neural pathways are used for consciousness and which not.

We shall not discuss here other cases in which people respond to a stimulus such as an auditory or visual cue, but claim to be unaware of the stimulus (for example, subliminal perception and priming,[17,18] and also galvanic skin responses[19]). Such stimuli may affect ongoing mental processes, including higher order processes, without being registered in working or long-term memory.

Neuronal firing

What is the general character of neural behaviors associated, or not, with consciousness? We think it plausible that consciousness in some sense requires neuronal activity, in which we include not only neurons that fire action potentials but also non-spiking neurons such as amacrine cells.

Our basic hypothesis at the neural level is that it is useful to think of consciousness as being correlated with a special type of activity of perhaps a subset of neurons in the cortical system. Consciousness can undoubtedly take different forms, depending on which parts of the cortex are involved, but we hypothesize that there is one basic mechanism (or a few such) underlying them all. At any moment consciousness corresponds to a particular type of activity in a transient set of neurons that are a subset of a much larger set of potential candidates. The problem at the neural level then becomes:

1. Where are these neurons in the brain?
2. Are they of any particular neuronal type?
3. What is special (if anything) about their connections?
4. What is special (if anything) about the way they are firing?

Visual awareness

At this point we propose to make a somewhat arbitrary personal choice. Since we hypothesize that there is a basic mechanism for consciousness that is rather

similar in different parts of the brain (and, in particular, in different parts of the neocortex), we propose that the visual system is the most favorable for an initial experimental approach. The visual system has several well-known advantages as an experimental system for investigating the neuronal basis of consciousness. Unlike language, it is fairly similar in man and the higher primates. There is already much experimental work on it, both by psychophysicists and by neuroscientists. Moreover we believe that it will be easier to study than the more complex aspects of consciousness associated with self-awareness. From now on, then, we shall discuss not consciousness in all its aspects but visual awareness and only the awareness that does not involve the laying down of long-term episodic memory. Johnson-Laird[1] has called this 'bare awareness', which does indeed convey the idea that it is simpler than self-awareness. Although here we mainly consider visual perception, many of the processes we discuss are also likely to be used in visual inagery.

The mammalian visual system

The visual system of mammals is complex and we only give an outline description. In the macaque monkey there are many distinct visual areas in the neocortex, perhaps as many as two dozen.[20] One reason for these multiple areas is that to handle all activity in one single very large neural net, with everything connected to everything else, would make the brain both cumbersome and prohibitively large. Loosely speaking these areas are connected in several hierarchies, with the processing within any one area being largely parallel (see Figure 1). There are also many back projections (and some cross connections) between cortical areas, and also from the cortex back to the thalamus, the functions of all of which are unknown.

The dominant output of the retina projects—through the lateral geniculate nucleus of the thalamus—to the primary visual cortex at the back of the brain. This area has neurons responding to fairly simple features (such as an oriented edge) occupying a tiny part of the visual field.

The neurons in the higher cortical areas respond to more complex features, e.g. aspects of faces, and their receptive fields cover much larger areas. The different cortical areas respond, in general, to different features. For example, the neurons in area MT are mostly interested in motion and depth, those in area V4 in color and shape, and those in 7a in position in space relative to the head or the body. *So far no single area has been found whose neurons correspond to everything we see.* How is it, then, that we seem to have a single coherent visual picture of the scene before us?

The obvious suggestion is that at any one time the relevant neurons in many cortical areas cooperate together to form some sort of global activity and that this global activity corresponds to visual awareness. This is the hypothesis that we shall try to fill out with psychological and neurobiological details.

Convergence zones

Which particular cortical areas contribute directly to visual awareness? For example, is the first visual area, V1, closely involved? Damasio[7] has also suggested that many cortical areas take part at any moment at several different levels. It is probably significant that Damasio arrived at this idea largely from considering the detailed effects of various types of brain damage, evidence rarely used by other theorists. He suggests that the synchronous activity in the various cortical areas is coordinated by feedback projections from 'convergence zones'.

What exactly convergence zones means in neural terms is somewhat obscure. We suggest that convergence zones may mainly refer to the neurons (or a subset of them) that project 'backwards' (see Figure 1), such as those that project from the second visual area back to the first one. Thus the term zones may turn out to convey a somewhat misleading impression, since such neurons may be sprinkled fairly uniformly through certain layers of the neocortex. Nevertheless the basic idea is a valuable one. It points to the possibility that, in addition, convergence zones may also exist in other places, such as the thalamus or the claustrum. (The claustrum[21] is a thin sheet of cells located just below the neocortex of the insula. It receives axons from almost all cortical areas and, in return, projects back to them. Its caudal region is largely visual.)

One would certainly expect areas near to the hippocampus, such as the inferotemporal region, as well as those near to the higher levels of the motor system to be closely involved in visual awareness. How far awareness is directly correlated with every neocortical level of the visual system remains uncertain, as is the involvement of the thalamus, although certain parts of it, such as its reticular nucleus[22] and the pulvinar[23] may be involved with attention.

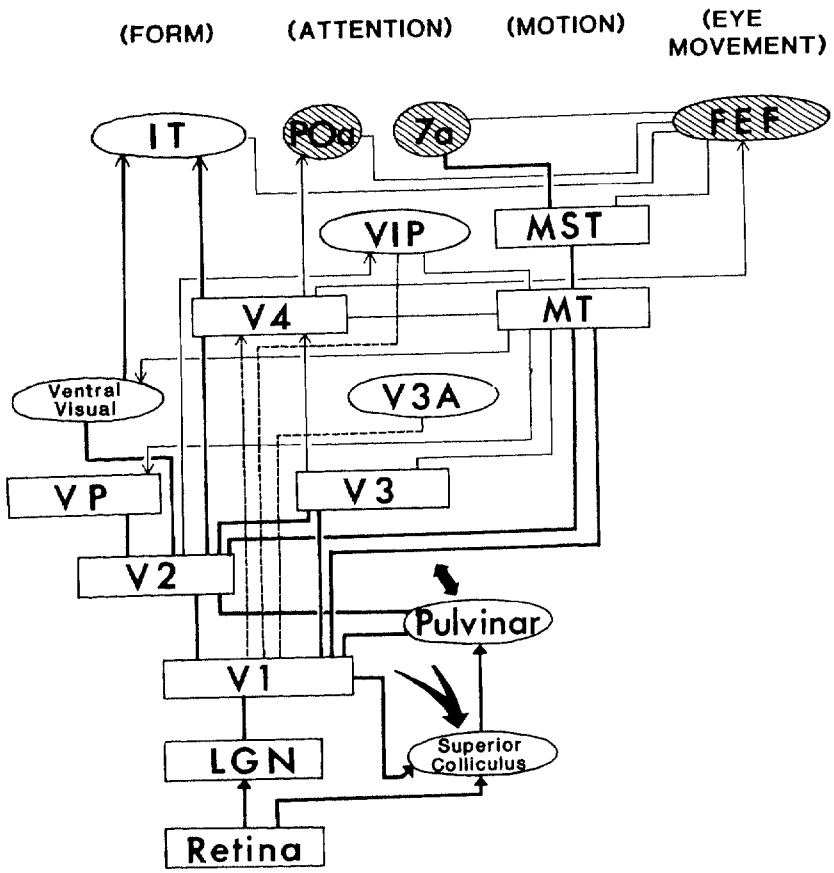

Figure 1. Some of the major cortical and sub-cortical visual areas in the macaque monkey. Major pathways are indicated by bold lines, minor pathways by thin lines; dashed lines weak or uncertain. The shaded ellipsoids correspond to higher-level visual-motor areas. Most cortical areas project back to the superior colliculus as well as to the different maps in the pulvinar, which is part of the thalamus. Furthermore, every forward projection has usually an equally strong, if not stronger, back-projection associated with it (not shown). These feedback pathways usually terminate outside cortical layer IV, while the forward projection terminates most densely in this layer. Areas related to the control and expression of visual attention include the superior colliculus and the pulvinar, as well as areas POa and 7a, part of the posterior parietal cortex. Abbreviations: FEF—frontal eye fields; IT—inferior temporal lobe; LGN—lateral geniculate nucleus; MT, MST—motion-processing areas; V1—primary visual area; V2-4—higher-order visual processing areas; VIP—ventral interparietal area; VP—ventral posterior. The figure is from D. Van Essen and J. Maunsell, personal communication. For more information, see ref 20.

Binding and selective visual attention

The binding problem

There are an almost infinite number of possible, different objects that we are capable of seeing. There cannot be a single neuron, often referred to as a grandmother cell, for each one. The combinatorial possibilities for representing so many objects at all different values of depth, motion, color, orientation and spatial location are simply too staggering. This does not preclude the existence of sets of somewhat specialized neurons responding to very specific and ecological highly significant objects, such as the neurons responding to aspects of faces in infero-temporal cortex of primates.[24]

It seems likely that at any moment any individual object in the visual field is represented by the firing of a *set* of neurons. This would not cause any special problem if the members of the set were in close proximity (implying that they probably interact somewhat), received somewhat similar inputs and projected to somewhat similar places. But because any object will have different characteristics (form,

color, motion, orientation) that are processed in several different visual areas, it is highly reasonable to assume that seeing any one object often involves neurons in many different visual areas. The problem of how these neurons temporarily become active as a unit is often described as 'the binding problem'. As an object seen is often also heard, smelt or felt, this binding must also occur across different sensory modalities.

Our experience of perceptual unity thus suggests that the brain in some way binds together, in a mutually coherent way, all those neurons actively responding to different aspects of a perceived object. In other words, if you are currently paying attention to a friend discussing some point with you, neurons in area MT that respond to the motion of his face, neurons in area V4 that respond to its hue, neurons in auditory cortex that respond to the words coming from his face and possibly the memory traces associated with recognition of the face all have to be 'bound' together, to carry a common label identifying them as neurons that jointly generate the perception of that specific face.

Types of binding

Binding can be of several types. In a sense a neuron responding to an oriented line can be considered to be binding a set of points. The inputs to such a neuron are probably determined by genes and by developmental processes that have evolved due to the experience of our distant ancestors. Other forms of binding, such as that required for the recognition of familiar objects such as the letters of a well-known alphabet, may be acquired by frequently repeated experience; that is, by being overlearnt. This probably implies that many of the neurons involved have as a result become strongly connected together. (Recall that most cortical neurons have many thousands of connections and that initially many of these may be weak.) Both these types of binding are likely to have a large but limited capacity. These are the types of binding with which Damasio[7] is mainly concerned.

The binding we are especially concerned with is a third type, being neither epigenetically determined nor overlearnt. It applies particularly to objects whose exact combination of features may be quite novel to us. The neurons actively involved are unlikely all to be strongly connected together, at least in most cases. This binding must arise rapidly. By its very nature it is largely transitory and must have an almost unlimited potential capacity, although its capacity at any one time may be limited. If a particular stimulus is repeated frequently, this third type of transient binding may eventually build up the second, overlearned type of binding.

The role of attention[25-28]

This form of transient binding probably depends on a serial attentional mechanism, sometimes called the spotlight of attention.[25,28] It is thought by some to concentrate on one place in the visual field after another, possibly moving every 60 ms or so. It is faster than eye-movements (another, slower, form of attention), and can work across different spatial scales. We suggest (the idea goes back to the last century[29]) that what reaches visual awareness is usually the result of this attentional step. In other words, *that awareness and attention are intimately bound together.* Note that although the results of attention are postulated to reach consciousness, the attentional mechanisms themselves are probably largely unconscious.

Short-term memory

It has long been argued that awareness is not only associated with attention but also with some form of short-term memory. We have to distinguish two forms, iconic and working memory. Iconic memory is similar to a sensory input buffer, storing information mainly at a pre-categorical level, i.e. in terms of simple visual primitives such as orientation or movement, although it may also involve more complex features, such as familiar words. It has a very large capacity but decays very quickly, perhaps in half a second or less.[8] Working memory, in contrast, may last for seconds;[9,10] it seems to have a rather limited capacity (a figure often quoted is seven items[30]), is different for different modalities and seems to use a much more abstract representation (post-categorical memory). It can be prolonged by rehearsal, as when one rehearses a telephone number. We suspect that it is this form of memory that is strongly activated in the binding process.

Neural basis of short-term memory

The neural basis of these two forms of memory is much understudied. Three broad types of mechanisms

might underlie either iconic or working memory, either singly or in combination:

(i) the strength of certain synapses is temporarily increased or decreased;
(ii) a neuron keeps on firing for some time due mainly to its intrinsic biophysical properties;
(iii) a neuron keeps on firing mainly due to extrinsic circuit properties ('reverberatory circuits').

Intracellular recordings show that the cell bodies of retinal ganglion cells remain depolarized for 200 ms or longer following a very brief flash of light within their receptive field.[31] How neurons further in the system express a transient memory remains to be seen. In a memory task in which a trained monkey had to discriminate and retain individual features of compound stimuli, about 15% of the neurons in inferotemporal cortex continued to fire for 10-20 s after the stimulus.[32] The mechanism producing (and terminating) this sustained firing is unknown. It is not yet known whether this firing is oscillatory or not. It may not have to be oscillatory, since a task can be remembered well enough without one being vividly conscious of it all the time.

A phenomenon that may correspond to the first mechanism was discovered many years ago at the neuromuscular junction and called post-tetanic potentiation. This is not to be confused with the much studied long-term potentiation (LTP)[33] that has a longer time course. The terminology for these short-lasting changes has now become more complicated,[34,35] with effects variously labelled facilitation (with two time constants of decay, of 50 ms and 300 ms), augmentation (7 s) and potentiation (30 s to minutes). There is also depression (5 s). We shall refer to all these collectively as 'short-term synaptic modification'. Note that the decay times are in general in the same range as the times required for the decay of working memory. Short-term modification has also been observed in the hippocampus and, in one case, in the neocortex. These temporary changes in synaptic strength can be surprisingly large (see figures 11 and 13 in ref 33).

Hebbian or non-Hebbian?

The causes of the various aspects of short-term modification are not fully understood but many of them seem to result from calcium accumulation in the presynaptic terminal[34,35] and are not influenced by activity in the postsynaptic cell. For this reason they are probably all non-Hebbian[36] in character. ('Hebbian' means that the alteration in synaptic strength depends on activity, at about the same time, in both the presynaptic and the postsynaptic cell.) Non-Hebbian implies that the change is not (to a first approximation) just at an individual synapse but at all the synapses associated with that particular presynaptic cell, so that the 'unit of change' is the presynaptic cell, not a single synapse. Such changes may form part of the neuronal basis of (short-term) priming.[17]

Von der Malsburg[37] in 1981 postulated on theoretical grounds a fast (fraction of a second) mechanism for increased synaptic efficiency, having a broadly Hebbian[36] character. So far this transient Hebbian alteration to synaptic strength has not been observed, although if it existed it would provide very considerable theoretical advantages.

We know almost nothing concerning the anatomical localization of iconic and working memory, except that (as discussed above[13]) the hippocampal system need not be present in order to remember for short times. Both iconic and working memory are likely to be distributed throughout the appropriate cortical areas, with auditory events transiently stored in auditory cortices, visual events in the visual cortices, and so on. No case of a person who is conscious but has lost all forms of short-term memory has been reported.

Neuronal oscillations

It has often been hypothesized, in particular by von der Malsburg,[37] that in neural terms binding means the temporarily correlated firing of the neurons involved. In other words, neurons in different parts of cortex responding to the currently perceived object fire action potentials at about the same time. The main theoretical requirement is for correlated firing. It now seems theoretically plausible that this is most easily implemented using oscillations. There are many oscillations known in the cortex, among them the α-rhythm and the θ-rhythm of the hippocampus. Experimental work on 'γ-oscillations' in mammalian olfactory cortex[38,39] and in the visual cortex[40-46] indicates that an important mechanism may involve neurons firing in semisynchrony at a frequency in the 40-70 Hz range (at least in the cat) (see Figure 2). In the visual cortex, only a proportion of the active neurons fire in this way. Most of the neurons that

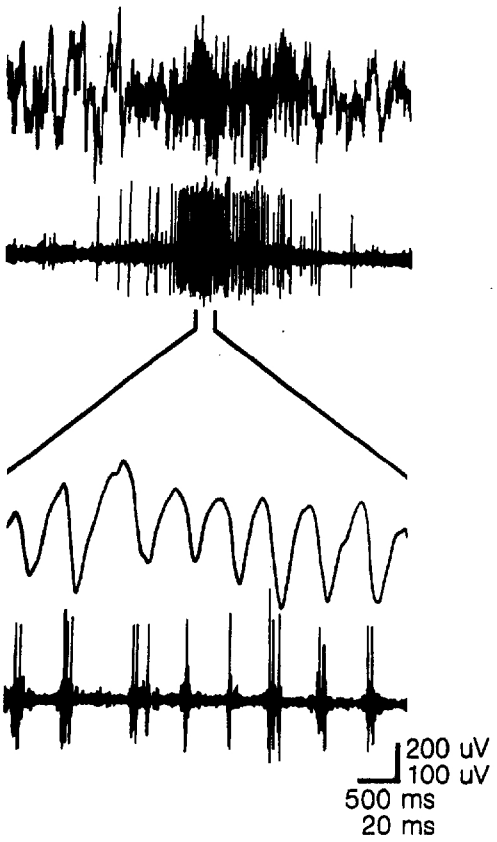

Figure 2. The local field potential (upper trace) and multiunit activity (second trace) recorded from an electrode in the first visual area of an adult cat in response to an optimally oriented light bar moving across the receptive field. The third and fourth traces show a corresponding part of the first and second ones (respectively) on an expanded time scale. Note the presence of rhythmic oscillations (35-45 Hz). The multiunit activity occurs near the peak negativity of the local field potential. Reproduced, by permission, from ref 42.

do oscillate appear to be complex cells rather than simple cells, so that very few are found in cortical layer IV.[45]

It has been shown[41] in cat primary visual cortex that in many cases two neurons some distance apart, but both responding to a particular moving bar, can show frequency-locked oscillations with no measurable phase shift in their cross-correlations. More tellingly, the activity of two relevant neurons, some 7 mm apart, showed[41] more highly correlated oscillations (at about 50 Hz) to a single moving bar than did the same two neurons stimulated by two shorter, separate, moving bars. That is, the correlation was much greater in response to an obviously single object than to a pair of somewhat distinct objects. The activity of the relevant neurons in both the first and second visual areas has also been shown to be correlated.[44] We shall refer to these neuronal oscillations as the 40 Hz oscillations, though the frequency on any one occasion is not very precise and can vary between 35 and 75 Hz.

Most of these experiments have been done on lightly-anaesthetized cats whose exact state of awareness is unknown but as similar oscillations have also been seen in neurons in alert cats,[46] they are unlikely to be merely an artifact of the anaesthetic.

A sketch of a theory

The role of the oscillations in binding may be as follows. Objects in the visual field give rise to appropriate responses in the appropriate cortical areas. Visual attention now has to select one of these objects at a particular location. Exactly how this works is not yet known. One possible solution[25,28,47] is that the brain has some sort of topographic saliency map that codes for the conspicuousness of locations in the visual field in terms of generalized center-surround operations. This map would derive its input from the individual feature maps and give a very 'biased' view of the visual environment, emphasizing locations where objects differed in some perceptual dimension, i.e. color, motion, depth, from objects at neighboring locations. Where this map might be located is unclear but parts of the thalamus,[22,23] such as the pulvinar, might be involved. (Note that the pulvinar receives a projection from the superior colliculus, an area that could be thought of as having a saliency map for eye movements.)

Once a particular salient location has been selected, probably by a winner-take-all mechanism, the information associated with it must be activated by referring back to the individual feature maps. The various parts of the visual thalamus are well suited to influence the behavior of neurons in those parts of the neocortex from which they receive information and to which they project back. Some such feedback pathways then activate or synchronize the oscillations at the corresponding locations in areas such as V1, V4 and MT, so that a coherent set of features is bound. Of course, the system had to decide, using the categorical knowledge stored in the connections, exactly which neurons must oscillate together to produce a veridical representation of the object being attended to. This is probably not an easy task. These

are the problems that concern much theoretical work in cognitive science and computational neuroscience. It is well known that making sense of our perceptual inputs is an 'ill-posed' problem and much 'computation' must be done to produce veridical solutions.

It seems likely that, for one reason or another, certain neurons in the cortical areas involved tend to oscillate at around 40 Hz. This could make it easier to activate such oscillations quickly. The effects of the 40 Hz oscillations is probably greater than the same amount of random firing for two reasons: first, because spikes arriving at a neuron simultaneously will produce a larger effect at the soma; and second, because the 40 Hz oscillations may promote 40 Hz 'resonances' elsewhere.

To be effective the phases of the relevant neurons must be synchronized. A recent mathematical analysis[48] shows that a set of oscillating neurons can very quickly be frequency- as well as phase-locked if there is a centralized feedback unit. On the other hand phase-locking is much slower and more difficult to achieve if the interactions have to pass over long chains of interneuronal connections.

As there has probably been strong evolutionary pressure to produce appropriate phase-locking as quickly as possible, it may be a general principle that all phase-locking interactions use only pathways with as few links as possible; these may be within a cortical area, between cortical areas (in adjacent levels in the hierarchy), or to and from one or more global coordinating regions, such as the thalamus or the claustrum.

We suggest that one of the functions of consciousness is to present the result of various underlying computations and that this involves an attentional mechanism that temporarily binds the relevant neurons together by synchronizing their spikes in 40 Hz oscillations. These oscillations do not themselves encode additional information, except in so far as they join together some of the existing information into a coherent percept. We shall call this form of awareness 'working awareness'. We further postulate that *objects for which the binding problem has been solved are placed into working memory*. In other words, some or all of the properties associated with the attended location would automatically be remembered for a short time. One very attractive possibility, with no experimental support, is that the 40 Hz oscillations themselves preferentially activate the mechanisms underlying working memory and possibly other longer-term forms of memory as well.

The likelihood that only a few simultaneous, distinct oscillations can exist happily together might explain,[49] in a very natural way, the well-known limited capacity of the attentional system. The 40 Hz oscillations may also be an important element in laying down and improving stored, categorical information, because at any moment they represent the brain's consensus as to what it is seeing.

Fleeting awareness

Can a spotlight of attention, moving over the visual field from one 'salient' location to the next, explain the perceptual richness of our environment? Would such a mechanism not lead to a sort of 'tunnel vision', in which the currently attended location appears in vivid detail with its associated perceptual attributes while everything else is invisible? We suggest, very tentatively, that this richness may be mediated by another form of awareness that is very transient, being associated with iconic memory and having a very large capacity at any one time. This form, that we propose to call 'fleeting awareness', we expect not to solve the *ad hoc* binding problem (as working awareness does) but to embody 'features' that are bound only epigenetically or by overlearning. Attention can then focus on a subset of relevant items within iconic memory for further processing. Because fleeting awareness is very transient it may be especially difficult to study. Whether it really exists remains to be seen.

Experimental problems

To understand working awareness it seems clear that experimental studies of three somewhat different phenomena are needed. The first is the 40 Hz oscillations, and indeed all other oscillations and any other kinds of coherent firing. In what cortical areas are oscillations found, especially in alert animals? What is their natural history: when do they occur? How long do they take to set up? How long does any one of them last? Are there, simultaneously, several different frequencies and/or phases? Are they indeed mainly correlated with objects? Do they solve the well-known figure-ground problem? Do they ever occur in the thalamus or the basal ganglia? Is the claustrum[21] involved—it is ideally placed to help synchronize neurons in different places?

The second is attention. Are any of the oscillations associated with it? How does the mechanism work

in neural terms? It is already known that neurons in area V4[50] and in the parietal region[51] in the macaque monkey are influenced by attention. Evidence on humans with brain damage implicate parts of the thalamus in certain aspects of attention,[22] as do certain experiments on monkeys,[23] but we need to know much more about how attention works neurobiologically.

The third is the neural basis of both iconic and working memory. These are neglected subjects that urgently demand an experimental attack. If there are reverberating circuits, where are they? Under what circumstances do they oscillate? Are there special mechanisms within certain neurons that make it easy for them to continue firing? How widely is the short-term synaptic modification mechanism distributed? What are its characteristics in different places? Is there any sign of a Hebbian form of modification? What exactly are the biochemical and biophysical mechanisms[34,35] underlying it? In addition we need to relate the neural activity to psychological observations.

We plan to discuss these experimental questions more fully in a later paper and to offer some tentative answers based on more detailed theoretical arguments.

Some experimental approaches

There are at least three possible methods of directly approaching working awareness. One is to exploit the process known as rivalry, an example of which is the two alternative ways of viewing the well-known Necker cube drawing. Here the visual input is constant but the percept varies from one interpretation to the other. A form of rivalry that is easier to study neurophysiologically is binocular rivalry, where one eye is given one constant visual input and the other a constant but rivalrous one. An example studied by John Allman and his colleagues (personal communication) involved projecting a small horizontal grating into one eye of a macaque monkey and a small vertical one into the other eye so that their visual fields overlap. In the same circumstances people see first one grating and then the other, the two percepts alternating every second or so. Detailed psychophysical studies on both monkeys and humans, where the monkeys signalled the change by pressing one of two keys, give fairly similar results, so it is reasonable to assume that the monkey has an awareness of the changing percepts. Most neurons in the first visual area fire in response to the constant visual input and do not change but the firing of others tends to change with the monkey's apparent percept. A similar experiment involving horizontal gratings moving either upwards (in one eye) or downwards (in the other) produces a similar variety of responses in area MT of the monkey.[52,53]

The second approach,[54] using cats, is to let the animal view the same set of visual signals first while it is awake and then again when it is in slow-wave sleep. The neurons studied were mainly in the first visual area. The general result is that the response of any given neuron is broadly the same in the two conditions but neurons in the lower layers are often markedly less active in slow-wave sleep. This was confirmed by experiments in which activity was monitored histochemically using deoxyglucose. (Note that all this was done before the 40 Hz oscillations were discovered.) Both these types of experiments are obviously in their infancy but they certainly suggest that visual awareness can be directly approached by suitably designed experiments.

A third approach to working awareness would be to study the effect of anaesthetics on awareness and recall in humans and on neuronal responses in monkey under similar conditions. Surprisingly, no reliable means exists today to test whether a patient is unconscious during modern anaesthetic procedures,[12] which usually includes muscle relaxants to block voluntary and involuntary movements. Even the EEG is no reliable indicator of depth of anaesthesia, as it is often not distinguishable from that of a sleeping person. It has been claimed, however, that the disappearance of a 40 Hz wave in the auditory evoked potential correlates with loss of consciousness under anesthesia.[55] Four to six cycles of such an oscillation, with a period of about 25 ms, can be seen following stimulation by a single or a series of sharp clicks.[56] More work on this is obviously very desirable.

Much useful information can also be obtained by careful studies on humans with brain damage who have selective impairment of different aspects of visual perception[7,57] and especially of attention and binding.

Conclusion

Our tentative theory, most of the elements of which have already been proposed by others, is a program for research rather than a detailed model.

What are the essential features of visual awareness? The first requirement is for a form of running short-term memory, an idea that is certainly more than 100 years old (see ref 29). We postulate two distinct forms of this running memory, a very

transient iconic one that records built-in visual features and a working memory that lasts for a somewhat longer time and can also store combinations of features. We presume that parts of iconic memory form the basis of working memory.

The information about a single object is distributed about the brain. There has, therefore, to be a way of imposing a temporary unity on the activities of all the neurons that are relevant at that moment. (Incidentally we see no reason at all why this global unity should require fancy quantum effects.) The achievement of this unity may be assisted by a fast attentional mechanism, the exact nature of which is not yet understood. This mechanism is postulated to concentate on one object at a time, choosing by a winner-take-all process the next object that appears to it to be the most salient. The required unity takes the form of the relevant neurons firing together in semi-synchrony, probably at a frequency in the 40-70 Hz range. We tentatively suggest that this activates the appropriate parts of the working-memory system. The neural basis of this memory system is at the moment obscure, although the transient alteration of synaptic strengths is one likely mechanism.

There is also much neural activity in the visual system that does not reach full awareness. Much of this corresponds to the computations needed to arrive at the best interpretation of all the incoming information that is compatible with the stored, categorical information acquired in the past. It is this 'best interpretation' of which we become aware.

Why, then, is consciousness so mysterious? A striking feature of our visual awareness (and of consciousness in general) is that it is very rich in information, even if much of it is retained for only a rather brief time. Not only can the system switch rapidly from one object to another, but in addition it can handle a very large amount of information in a coherent way at a *single moment*. We believe it is mainly these two abilities, combined with the very transient memory systems involved, that has made it appear so strange. We have no experience (apart from the very limited view provided by our own introspection) of machines having complex, rapidly changing and highly parallel activity of this type. When we can both construct such machines and understand their detailed behavior, much of the mystery of consciousness may disappear.

Acknowledgements

We thank John Allman, Bernard Baars, Patricia Churchland, Paul Churchland, Antonio Damasio, Charles Gray, Bela Julesz, Dan Kammen, Georg Kreisel, and Leslie Orgel for helpful comments on earlier versions of the manuscript and thank Jennifer Altman especially for extensive improvements to the sumitted manuscript.

F.C. is supported by the J. W. Kieckhefer Foundation. C.K. is supported by the Air Force Office of Scientific Research, a Presidential Young Investigator Award from NSF and by the James S. McDonnell Foundation.

References

1. Johnson-Laird PN (1988) The Computer and the Mind. Harvard Univ Press, Cambridge, MA
2. Jackendoff R (1987) Consciousness and the Computational Mind. MIT Press, Cambridge, MA
3. Baars BJ (1988) A Cognitive Theory of Consciousness. Cambridge University Press, Cambridge, UK
4. Edelman GR (1989) The Remembered Present. Basic Books, New York
5. Nisbett RE, Wilson TD (1977) Telling more than we can know: verbal reports on mental processes. Psychol Rev 84:231-259
6. Churchland PM (1985) Reductionism, qualia and the direct introspection of brain states. J Philos 82:8-28
7. Damasio AR (1989) The brain binds entities and events by multiregional activation from convergence zones. Neural Computat 1:123-132
8. Coltheart M (1983) Iconic memory. Philos Trans R Soc Lond B 302:283-294
9. Baddeley A (1986) Working Memory. Oxford University Press, Oxford
10. Phillips WA (1983) Short-term visual memory. Philos Trans R Soc Lond B 302:295-309
11. Tulving E (1985) Memory and consciousness. Can J Psychol 26:1-12
12. Kulli J, Koch C (1990) Does anesthesia cause loss of consciousness? Trends Neurosci, in press
13. Damasio AR, Eslinger PJ, Damasio H, Van Hoesen GH, Cornell S (1985) Multimodal amnesic syndrome following bilateral temporal and basal forebrain damage. Arch Neurol 42:252-259
14. Geschwind N, Galaburda AM (1986) Cerebral Lateralization. MIT Press, Cambridge, MA
15. Weiskrantz L (1986) Blindsight. Oxford University Press, Oxford
16. Stoerig P, Cowey A (1989) Wavelength sensitivity in blindsight. Nature 342:916-918
17. Tulving E, Schacter DL (1990) Priming and human memory systems. Science 247:301-306
18. Kihlstrom JF (1987) The cognitive unconscious. Science 237:1445-1452
19. Tranel D, Damasio AR (1985) Knowledge without awareness: an autonomic index of facial recognition by prosopagnosics. Science 228:1453-1454
20. Van Essen D (1985) Functional organization of primate visual cortex, in Cerebral Cortex, vol 3, Visual Cortex (Peters A, Jones EG, eds), pp 259-329. Plenum Press, New York
21. Sherk H (1986) The claustrum and the cerebral cortex, in Cerebral Cortex, vol 5, Sensory-motor Areas and Aspects

of Cortical Connectivity (Jones EG, Peters A, eds), pp 467-499. Plenum Press, New York
22. Rafal RD, Posner MI (1987) Deficits in human visual spatial attention following thalamic lesions. Proc Natl Acad Sci USA 84:7349-7353
23. Petersen SE, Robinson DL, Morris JD (1987) Contributions of the pulvinar to visual spatial attention. Neuropsychologia 25:97-105
24. Desimone R, Ungerleider LG (1989) Neural mechanisms of visual processing in monkeys, in Handbook of Neuropsychology, vol 2 (Damasio AR, ed), pp 267-299. Elsevier, Amsterdam
25. Treisman A (1988) Features and objects: the fourteenth Bartlett Memorial Lecture. Q J Exp Psychol 40A:201-237
26. Julesz B (1981) Textons, the elements of texture perception, and their interactions. Nature 290:91-97
27. Posner MI (1986) Chronometric Explorations of Mind. Oxford University Press, Oxford, UK
28. Wolfe JM, Cave KR, Franzel SL (1989) Guided search: an alternative to the feature integration model for visual search. J Exp Psychol 15:419-433
29. James W (1981) The Principles of Psychology. Harvard University Press, Cambridge, MA
30. Miller GA (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol Rev 63:81-97
31. Baylor DA, Fettiplace R (1979) Synaptic drive and impulse generation in ganglion cells of turtle retina. J Physiol Lond 288:107-127
32. Fuster JM, Jervey JP (1981) Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. Science 212:952-955
33. Brown TH, Ganong AH, Kairiss EW, Keenan CL, Kelso SR (1989) Long-term potentiation in two synaptic systems of the hippocampal brain slice, in Neural Models of Plasticity (Byrne JH, Berry WO eds), pp 266-306. Academic Press, New York
34. Magleby KL (1987) Short-term changes in synaptic efficacy, in Synaptic Function (Edelman GM, Gall WE, Cowan WM, eds), pp 21-56. Wiley, New York
35. Zucker RS (1989) Short-term synaptic plasticity. Annu Rev Neurosci 12:13-31
36. Brown TH, Kairiss EW, Keenan CL (1990) Hebbian synapses: biophysical mechanisms and algorithms. Annu Rev Neurosci 13:475-511
37. Von der Malsburg C, Schneider W (1986) A neural cocktail-party processor. Biol Cybern 54:29-40
38. Freeman WJ (1978) Spatial properties of an EEG event in the olfactory bulb and cortex. Electroencephalogr and Clin Neurophysiol 44:586-605
39. Wilson MA, Bower JM (1990) Cortical oscillations and temporal interactions in a computer simulation of piriform cortex. J Neurophysiol, in press
40. Freeman WJ, van Dijk BW (1987) Spatial patterns of visual cortical fast EEG during conditioned reflex in a rhesus monkey. Brain Res 422:267-276
41. Gray CM, König P, Engel AK, Singer W (1989) Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. Nature 338:334-337
42. Gray CM, Singer W (1989) Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. Proc Natl Acad Sci USA 86:1698-1702
43. Eckhorn R, Reitboeck HJ (1989) Stimulus-specific synchronizations in cat visual cortex and their possible role in visual pattern recognition, in Synergetics of Cognition, Springer Series in Synergetics, vol 43 (Haken H, ed), pp 99-111. Springer, Heidelberg
44. Eckhorn R, Bauer R, Jordan W, Brosch M, Kruse W, Munk M, Reitboeck HJ (1988) Coherent oscillations: a mechanism of feature linking in the visual cortex? Biol Cybern 60:121-130
45. Gray CM, Engel AK, König P, Singer W (1990) Stimulus-dependent neuronal oscillations in cat visual cortex. Receptive field properties and feature dependence. Eur J Neurosci 2:607-619
46. Gray CM, Raether A, Singer W (1989) Stimulus-specific intercolumnar interactions of oscillatory neuronal responses in the visual cortex of alert cats. Soc Neurosci Abstr 15:320.4
47. Koch C, Ullman S (1985) Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiol 4:219-227
48. Kammen DM, Holmes PJ, Koch C (1989) Cortical architecture and oscillations in neuronal networks: feedback versus local coupling, in Models of Brain Function (Cotterill RMJ, ed), pp 273-884. Cambridge University Press, Cambridge
49. Stryker MP (1989) Is grandmother an oscillation? Nature 338:297-298
50. Spitzer H, Desimone R, Moran J (1988) Increased attention enhances both behavior and neuronal performance. Science 240:338-240
51. Andersen RA (1987) Inferior parietal lobule function in spatial perception and visuomotor integration, in Handbook of Physiology: The Nervous System V (Mountcastle VB, Plum FS, Geiger SR, eds), pp 483-518. American Physiol Society, Bethesda, MD
52. Logothetis NK, Schall JD (1989) Neuronal correlates of subjective visual perception. Science 245:761-763
53. Myerson J, Miezin F, Allman J (1981) Binocular rivalry in macaque monkeys and humans: a comparative study in perception. Behav Anal Lett 1:149-159
54. Livingstone MS, Hubel DH (1981) Effects of sleep and arousal on the processing of visual information in the cat. Science 291:554-561
55. Madler C, Pöppel E (1987) Auditory evoked potentials indicate the loss of neuronal oscillations during general anaesthesia. Naturwiss 74:S.42
56. Galambos R, Makeig S, Talmachoff PJ (1981) A 40-Hz auditory potential recorded from the human scalp. Proc Natl Acad Sci USA 78:2643-2647
57. Marcel AJ (1983) Conscious and unconscious perception: an approach to the relations between phenomenal experience and perceptual processes. Cog Psych 15: 238-300