Computational models of association cortex

Thomas Gisiger*, Stanislas Dehaene† and Jean-Pierre Changeux‡

Recent computational models, or mathematical realizations of neurobiological theories, are providing insights into the organization and workings of the association cortex. Such models concern the construction of cortical maps, the neural basis of cognitive functions such as visual perception, reward-motivated learning and some aspects of consciousness.

Addresses
*‡CNRS UA D1284 – 'Neurobiologie Moléculaire', Institut Pasteur, 28 rue du Dr Roux, 75724 Paris, Cédex 15, France
†Inserm Unit 334, Service Hospitalier Frédéric Joliot, CEA/DRM/DSV, 4 Place du général Leclerc, F-91401 Orsay, Cédex, France

Current Opinion in Neurobiology 2000, **10**:250–259

0959-4388/00/$ – see front matter

Abbreviations

LTD	long-term depression
LTP	long-term potentiation
NCC	neuronal correlate of consciousness
NMDA	*N*-methyl-D-aspartate
NO	nitric oxide
V1	primary visual cortex

Introduction

The development of cognitive activities during mammalian evolution is paralleled by a major expansion in both the size and complexity of the association cortices, which interconnect and unite primary sensory and motor areas. In order to understand how the organization of the association cortex — in particular, the prefrontal cortex — is related to its functions, models and their mathematical formalizations have become essential tools for the evaluation and organization of the increasing amount of experimental data from the neurosciences and cognitive psychology. Such models can be used to test neural hypotheses and to propose new theories [1•,2•]. In this review, we will examine computational models featuring characteristic aspects of the association cortex with the general aim of establishing a causal relationship between its neuronal organization and its functions.

The models used to investigate the association cortex are artificial networks of formal neurons, which are mathematical idealizations of the real cells, connected together by the functional equivalent of synapses. Depending on the models, the elementary units of the networks can be simple input–output binary units, such as the McCulloch–Pitts neurons [3], or physiologically more realistic integrate-and-fire-neurons [2•,4], to mention just two. It is also possible to use as units whole groups [5] or clusters [6] of neurons that are equivalent, for instance, to the ocular dominance columns found in the visual cortex and which are composed of roughly a hundred neurons densely interconnected by mostly excitatory projections. The main constraint one should try to impose on these networks is their 'neurorealism'; in other words, one should ensure that they reproduce biological reality to a sufficient extent, and thus can be experimentally tested.

Another important aspect of neural network modeling is the introduction of elementary learning mechanisms — these can be formally implemented by algorithms expressing the regulation of synaptic strength as a function of experience. An example of such a learning procedure formally relies on the output of the network to back-propagate deep into the layers of the network and make corrections to the strength of all the connections (see [2•]). This so-called 'back-propagation algorithm', introduced in connectionist networks in which computational efficiency often supersedes biological realism, has nevertheless found support in observations of small arrays of hippocampal cells *in vitro* by Fitzsimonds *et al.* [7].

A more locally acting learning rule, introduced by Hebb [8], proposes that the strength of the synapse between two neurons increases when presynaptic and postsynaptic activity coincide within a short time window. Similarly, synaptic strength may decrease if both neurons consistently fail to fire together. This mechanism therefore favors and stabilizes frequently occurring activity while removing counter-productive circuits. Two main biological implementations of this Hebbian learning have been proposed. The first makes use of the experimental phenomena of long-term potentiation (LTP) and long-term depression (LTD), which take place in NMDA-rich synapses [2•]. The other makes use of the allosteric properties of a large body of non-NMDA neurotransmitter receptors [9], the archetypal example of which is the nicotinic acetylcholine receptor. This macromolecule has the property of existing in at least four different states, each possessing its own characteristics and each accessible via discrete conformational transitions (see [9] for details). As the proportion of receptors in each conformational state changes with neuronal activity, so do the properties of the synapse. This therefore enables a form of plasticity and consequently implements some type of Hebbian learning. This molecular mechanism is a critical component of the 'synaptic triad', introduced initially as a theoretical construction in [6], where the synaptic strength of a synapse between two neurons is enhanced by the activity of a third neuron. This original device was shown to be able to recognize and produce time sequences [6]. Anatomical evidence supporting the existence of synaptic triads involving dopaminergic terminals has been recently reported in monkey cortex [10].

We end this introduction by mentioning another important aspect of modeling neural networks: the global mode of operation of the organism within its environment. Some models have deliberately abandoned the traditional

input–output information-processing scheme currently used in cybernetics in favor of a more realistic projective style; here, the formal organism constantly tests its environment by the constant production of hypotheses or pre-representations which are then internally compared and evaluated by the organism against the outside world [6,11–13].

Here, we will first describe the case of cortical mapping and information processing in the visual system; then we will introduce models of cognitive learning, including reward mechanisms, self-evaluation and strategy building. We will end with a consideration of neurally plausible formal theories of behavioral awareness or consciousness.

From cortical organization to perception

In this section, we present two approaches to the study of the association cortex. The aim of the first approach is to account for the development of the neuronal architecture of the association cortex by introducing local rules of synaptic epigenesis and segregation. The second approach concerns information processing in the visual system using models whose overall structure, although inspired by actual anatomical and physiological data, is artificially implemented by the modeler. The gap between these two approaches for studying the same system — though each seems quite promising in its own right — illustrates their respective limitations: models of cortical maps reproduce certain aspects of the spatial structure of areas such as the primary visual cortex very well, but their information processing capacity is almost nonexistent; models of visual perception, on the other hand, reproduce the neuronal architecture of cortical maps in a rather crude way.

The genesis of cortical maps

The association cortex comprises short-distance excitatory and inhibitory connections, as well as longer-range projections to functionally related areas. The models examined in this review represent the dynamics of fully developed networks of neurons and implement both types of connections. The growth and organization of these fully developed networks takes place in two phases: a growth phase, which leads to a large excess of connections; and a pruning phase, where redundant connections are removed. This process has been accounted for by a number of mechanisms, such as the selection of synapses at sensitive periods of development ([14]; see also [5,15]). In this framework, a role has been suggested for certain short-lived diffusible substances such as nitric oxide (NO) in the retrograde stabilization of active synapses and the weakening, or removal, of silent ones [16]. Also, trophic factors, such as brain-derived neurotrophic factor or nerve growth factor, have been shown experimentally to contribute to activity-induced segregation of neuron groups with a lesser initial redundancy in connections. Montague *et al.* [17] have proposed a model implementing a situation in which axons grow, sprout branches, and make synapses that are then either strengthened or eliminated via a retrograde (NO-type) messenger as a result of network activity. Simulations show that, in this formal framework, the proposed mechanism reproduces the segregation of neurons into ocular dominance columns, or into barrel-like structures (and can even reproduce the experimental effects on the somatosensory cortex of the plucking or taping together of whiskers), as well as reproducing the formation of reciprocal cortical connectivity. Input characteristics play an important role in determining which structure emerges. Yet, the actual contribution of NO to the genesis of cortical maps remains to be experimentally specified.

Work has also been undertaken in order to better understand the local sensitivity to stimulus features exhibited by the neural columns of the primary visual cortex (V1). Models implementing learning rules at the synaptic scale have been proposed to simulate the emergence of this columnar sensitivity, as the system is exposed to different types of visual stimuli. The visual topographic mappings obtained by simulations seem to depend on the choice of the learning rules. Among these rules, the Bienenstock–Cooper–Munro (BCM) algorithm seems the most promising; for instance, it can reproduce different degrees of ocular dominance and orientation selectivity (see [18] for details), and is supported by experimental evidence [19].

On a larger scale, Durbin and Mitchison [20] have proposed a dimension-reducing mapping model that implements possible mechanisms that shape the global characteristics of cortical maps. In the case of area V1 of the visual cortex, neuronal columns show well-defined preferences for distinct stimulus characteristics such as retinotopic position, orientation, degree of orientation tuning, ocular dominance, and so on. The organization of area V1 therefore implies that during development a mapping is established between the cortex — which is functionally equivalent to a two-dimensional sheet (i.e. because of its columnar structure, its reponse to stimuli does not change as a function of cortical depth) — and the parameter space of visual stimuli — of dimension at least equal to four, given the stimulus characteristics above. The reduction in the number of dimensions from the parameter space to the cortex implies that the requirement for neighboring neuron columns to code for similar stimuli cannot be met everywhere. To study further the constraints imposed on cortical maps, Durbin and Mitchison [20] introduced into their model a self-organizing algorithm that takes into account two effects: competitive interaction between units, which leads (via Hebbian-type learning) to the dominance of those columns with the strongest initial response to the input; and strengthening of units when neighboring clusters fire. The latter condition implements the principle that computations in the cortex (such as the sharpening of orientation tuning) involve connections between neurons coding for qualitatively similar parameter values. Simulations produce cortical maps roughly similar to those observed experimentally in area V1 [21], with smooth patches separated by regions of sudden jumps in stimuli response (see [20] for details). This model was further

tested by Goodhill *et al.* [22], who studied the effects on the mapping of the roughly ellipsoid shape of the cortex and of the higher density of retinal points in the fovea.

Re-entrance and the binding problem

Models that address the computational aspects of the visual cortex, and more precisely the possible roles of particular types of connectivity, have also been proposed. Using a large-scale simulation, Sporns *et al.* [23] found that reciprocal connections, or 'reentry', can establish coherent oscillatory activity in well-defined neuronal groups. They later extended this model to the processes involved at the early stages of the visual pathways, adding the property of fast synaptic plasticity (i.e. with time constants of the order of 100 ms) to reentrant connections [24]. This new feature allows the activity of neuron groups responding to parts of moving objects to synchronize rapidly with near-zero phase lags. As a result, the network is able to segregate two coherently moving parts in stimuli, such as a figure moving over a background or two figures moving relative to each other. This model illustrates the possible role of synchrony in coherent object formation, as studied empirically by Singer [25•].

Tononi *et al.* [26] have also presented evidence that a model implementing reentrant connectivity, fast synaptic plasticity and a biologically realistic architecture can solve the binding problem for stimuli consisting of simple geometric shapes. Structurally, the model simulates nine visual areas, which are divided between a primary/secondary region, and a higher associative region with a dense array of reentrant connections. Functionally, it is composed of three processing streams that converge in the higher processing areas. These streams, which correspond to motion, color and form, allow stimuli to be decomposed into these three features. A motor region enables the network to 'point' by foveation to a region within its visual field, and a reward or saliency system reinforces activity by long-term synaptic plasticity via Hebbian-type learning (see the section below on Cognitive learning). Tononi *et al.* [26] show that although stimuli are decomposed according to highly specialized streams in multiple areas, their features are correctly bound together, as the model is able to differentiate between several objects present in its visual field. Further, by asking the system to point to an object in its visual field and applying positive reinforcement whenever it chooses the right object, the network may be taught to recognize a given object (see [26] and references therein for relevant experimental data on visual processing).

Binocularity and perception

Other models address rivalry in binocular vision, a problem closer to perception than to vision [27]. When presented with a different stimulus at each eye, a subject can only see one of them at a time (rivalry) and, after a while, the two images start to alternate in the subject's perception (perceptual alternation). This phenomenon may depend on the competition between monocular neurons in the primary cortex. However, a model proposed by Lumer [28•] illustrates another view according to which rivalry arises by competition between the interpretations of stimuli in cortical areas higher than V1. Another key element of the dynamics of this network is the relative timing of the neuronal activities that are 'solving' each interpretation; this timing alters the outcome of the competition. It also enables the system to differentiate between conflicting stimuli and congruent ones (see [28•] for details). As noted by Lumer, even though the model can exhibit some sort of alternance for particular initial conditions, it is not consistent (nor does it reproduce the observed frequency) and so should not be perceived as actually reproducing the phenomenon of perceptual alternation. Dayan [29] showed, in a similar model, that the introduction of an oscillatory mechanism implementing a form of fatigue allows the reproduction of alternance between percepts.

Finally, and higher still in terms of cerebral functions, a model related to the perception of stimuli and processing of information in the thalamocortical system has been proposed by Lumer *et al.* [30,31]. It implements, for both the primary and secondary visual pathways, the connections between the cortex (represented by the supragranular, infragranular and IV layers), the reticular nucleus and the thalamus. Each area is modeled using neuro-realistic data for connections and ionic channels, and integrate-and-fire neurons. The resulting network reacts to stimuli in a non-linear fashion. For stimulus input intensities below a certain threshold (defined by the parameters of the model), the system only exhibits low-intensity, background activity. As the intensity of the input rises beyond the threshold, however, the dynamics of the system experience what the authors call a 'phase transition': the activity of the system changes suddenly to a stable oscillatory mode with a frequency of about 20–60 Hz and a large amplitude (compared to background activity).

Cognitive learning

In this section, we present formal efforts to build neuro-realistic architectures that are able to perform tasks devised to target specific cognitive functions. These models make specific predictions about the relevant neural processes that can, in principle, be compared with data from electrophysiological recordings and brain imaging studies in normal and lesioned subjects.

In the following section, we assume that living organisms adopt the projective style [12,13,32,33] mentioned in the introduction. Spontaneous activity and reward processes are critical ingredients in such exploratory-motivated behavior as reaching goals and learning new tasks. This aspect is traditionally absent from formal models of neural network dynamics. A significant step forward in this area of research was the initial formalization of reinforcement learning by Sutton and Barto (see [34••] and references therein). In their model, the network is not explicitly taught what to do, but instead receives a signal from the

exterior (such as an outside observer) that evaluates its performance. If this reward signal is negative, then the network must generate new solutions to the problem at hand until the reward becomes positive. The reward signal therefore either stabilizes the correct behavior or destabilizes the incorrect ones. However, the authors did not propose any biological implementations for this model.

Reward-motivated learning

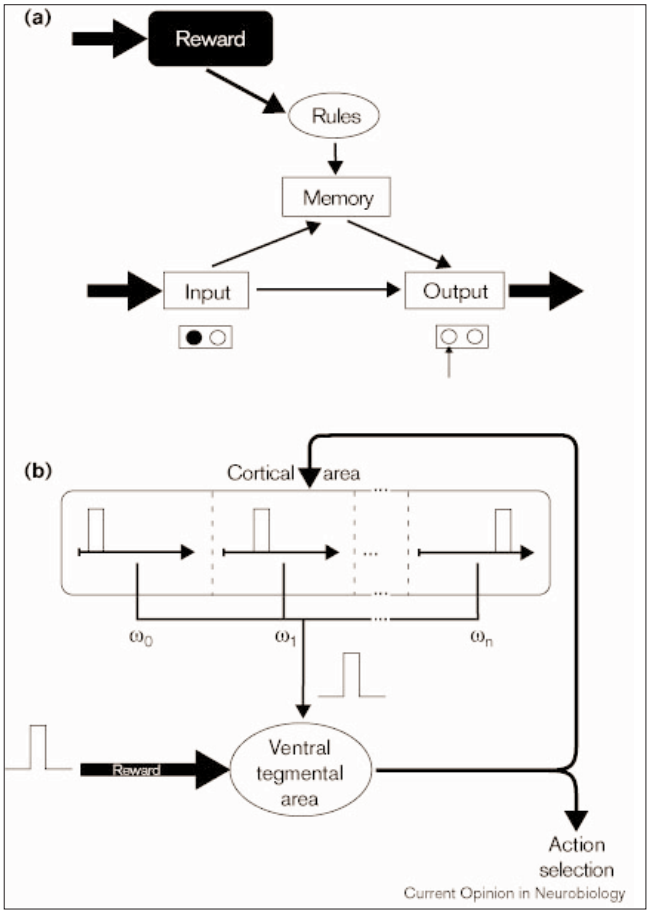
Delayed-response tasks were initially designed to test the acquisition of cognitive patterns that selectively engage the prefrontal cortex in higher vertebrates, including humans.

Dehaene and Changeux [12] have proposed a plausible neural architecture capable of performing such tasks (see also the subsequent work of Reeke *et al.* [35]). The delayed-response test proceeds in three steps. First, the subject is shown a cue object, followed by a waiting period of variable length. Second, two objects are then presented simultaneously to the subject, who chooses one. Finally, reward (either positive or negative) is provided, which serves to evaluate the performance through release of neuromodulator substances, such as dopamine, serotonin or acetylcholine. The rule defining the correct choice may vary during the test.

The artificial organism [12] comprises two levels (see Figure 1a). Level 1 is a visuo-motor loop comprising representations of visual areas and the premotor cortex. Level 2 contains memory and rule-coding units implementing functions of the prefrontal cortex or related areas. Rule-coding clusters play a key role in the dynamics, selecting the feature (e.g. color, position, shape) of the cue maintained 'on-line' by the memory unit during the delay period and then using the feature to choose an object. Because of lateral inhibition, only one rule-coding unit can be active at a time. The network receives an evaluation of its performance in relation to the outside world via a reward signal that causes changes in synaptic efficacy via allosteric Hebbian learning. Negative reward acts to lower the activity of the dominant rule-coding cluster and to raise that of the others, making the organism ready for a new behavioral rule.

Level 1 of the network can perform the task correctly if only one rule is used in the test. However, when this constraint is relaxed, the organism persists with the old rule, even though it receives negative reward at each trial. This behavior is typical of patients with frontal lesions. With the addition of level 2, however, the network passes the test and displays performances similar to normal subjects. The rule-coding layer plays the role of a generator of diversity, permitting the organism to switch rules when needed (see Figure 2 for a possible implementation). The network does not learn rules case-by-case. Instead, the rule-coding neurons signal the expectancy entertained by the organism at a given time (see [12] for further details), and define the content of its short-term memory during the delay period.

Figure 1

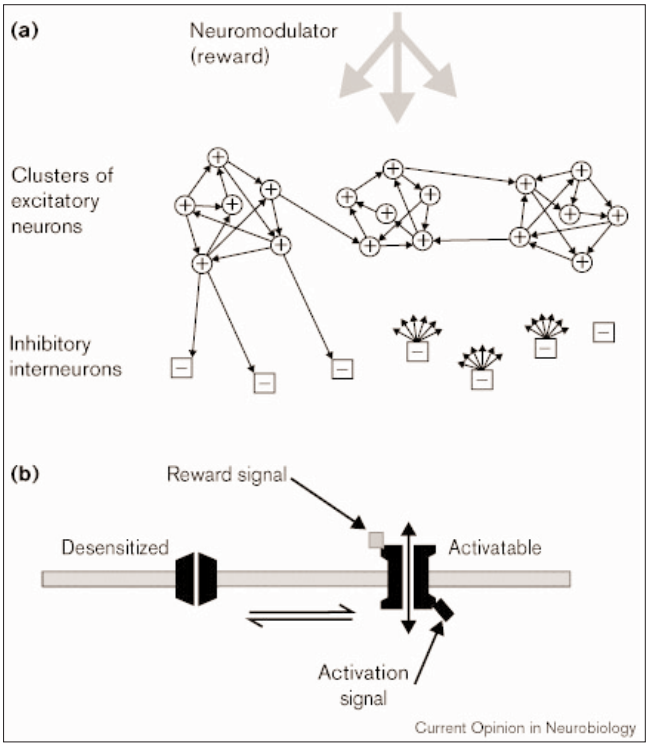


Roles of reward (or saliency) signals. **(a)** In the delayed-response task, reward acts on rule-codings clusters, which filter the feature of the cue object that will be kept on-line during the waiting period. Reproduced from [43]. **(b)** Reward takes the form of a signal released in a certain amount and at a given time. The internal representation of the reward signal is coded in 'n' cortical areas of contribution weighted by 'synaptic efficacities' $\omega_0,\omega_1,...,\omega_n$. Both reward and expectation signals converge to an equivalent of the ventral tegmental area, as learning modifies the ω_i to minimize the discrepancies between the two. The output of this area adequately reproduces dopamine release during learning and performance of the tasks. Modified from [37].

Electrophysiological recordings in the monkey have revealed short-term memory cells whose behavior resembles those of the memory units (see [10]). Neurons analogous to the postulated rule-coding cells have also been described (e.g. by Asaad *et al.* [36•]).

The Dehaene–Changeux model illustrates the principle of cognitive learning by production and selection of pre-representations — rules in this case ([13]; see also [35]). Networks with this type of architecture can also pass other tests targeting prefrontal functions, such as the Wisconsin card sorting test [11]. Here, objects are cards representing simple shapes that the subject must classify according to criteria of color, number or type of forms represented on the cards, which vary during the test.

Figure 2



Possible implementation of a neuronal generator of diversity.
(a) Clusters, consisting of strongly interconnected excitatory neurons, inhibit each other via long-range inhibitory interneurons. Reward enters the system by diffusion, modulating the activity of all neurons.
(b) Possible molecular implementation of synaptic strength modulation by allosteric transition of a receptor molecule. When a positive reward and a postsynaptic activation signal reach the receptor simultaneously, the conformational equilibrium switches from a desensitized state to an activatable state, thus increasing the synaptic efficacy of the clusters. At variance with this, negative reward destabilizes the system and allows the generator of diversity to test, at random, new alternative clusters of activity until a positive reward is received. Modified from [11].

Besides motivating an organism to perform simple tasks, reward mechanisms also seem to play a role in situations where immediate goals are absent. This observation is supported by experiments where for instance a monkey has to wait for a cue before touching a lever in order to receive a reward (see [37] for a review). Before and during training, recordings show that most of the dopamine neurons in the ventral tegmental area fire after reward delivery. However, when the task is learnt, dopamine neuron responses shift from the time of the reward presentation to that of the cue. Dopamine neuron activity, therefore, codes for the discrepancy between the time of presentation and amount of reward expected, and their actual realizations.

Montague and colleagues [38,39] have modeled these findings using the temporal difference algorithm proposed a decade earlier by Sutton and Barto (see [34••] and references therein). Roughly, this algorithm gives representations of the stimulus and of the time of reward presentation. The difference between this prediction and the subsequent reality is then computed as an error signal, which must be minimized using synaptic plasticity (see Figure 1b). Following this learning process, the firing of the module representing the ventral tegmental area takes place after the cue, not when the reward is received. However, modeling of this process is constrained by the lack of reliable information about how the brain stores temporal information, even though such storage does take place. Yet the model of Montague and colleagues still leads to testable predictions [37]. If the task is now modified to include multiple cues, the dopamine release should be strongest right after the earliest consistent cue. Furthermore, the model predicts that if one cue is unexpectedly removed from the task, dopamine release should be minimal at the time of the absent cue, not after the reward. Suri and Shultz [40•] also extended this model to a spatial delayed response task.

Auto-evaluation and hierarchical problem solving

Another element of neuronal architecture was introduced by Dehaene and Changeux [11] in their model of the Wisconsin card-sorting task. It consists of an auto-evaluation loop, which takes effect when the organism receives negative reward. This loop short-circuits the sensory–motor loop and allows the network to test internally generated 'intentions' by comparing them with memories of previous attempts (see [11] for details). Dehaene and Changeux [41,42] made use of this internal regulatory loop in two other models; in the first, the formal organism acquires elementary numerical abilities [41]; in the second, it solves the Tower of London test [42,43]. One of the key elements of the first network [41] is a numerosity-detector device, which is able to estimate the number of objects (here, one dimensional 'blobs' of various sizes) which form on a simulated retina. The model is able to compare two numerosities, as well as to develop by itself the concepts of 'larger' compared to 'smaller', and of 'more' compared to 'less'. Experimental evidence has recently confirmed that this ability develops spontaneously in animals [44].

In the case of the Tower of London task [42,43], the network has to move beads on rods according to well-defined rules from an initial position to a pre-specified goal [45]. Solving the problem requires planning, because several intermediate moves are sometimes necessary to reach the goal. This strategy building is known to involve the prefrontal cortex. The network which solves the task [42,43] consists of two main components: a descending planning system, and an ascending evaluation system. In the descending planning system, plans unfold in a hierarchy of three levels: a motor level, which commands moves of the beads; an operation level, which implements single moves (by pointing to a bead and then pointing to the position where it must go); and finally a plan level, which generates whole series of moves. This series of moves is not random, as it is internally tested, using an auto-evaluation loop, by

the ascending evaluation system to ensure that it brings the system closer to its goal. The performances of the resulting network reproduce those of normal subjects, and specific lesions of the architecture of the organism produce patterns of errors typical of patients with frontal lesions (see [42,43] for details).

The problem of consciousness

The previous sections dealt with the interactions between the association cortex and sensory or motor areas. These interactions naturally lead to such issues as perception, awareness, decision-making and consciousness. Already several of the cognitive learning tasks mentioned in the previous section are often referred to as requiring consciousness, or 'conscious thinking'. However, the scientific investigation of consciousness is still in its infancy, and it is much too early to predict which approach will be the most successful. Here, we will restrict ourselves to describing a few attempts to formally implement limited aspects of consciousness.

A major step forward in this investigation would be the identification of what Crick and Koch call "the neuronal correlate of consciousness" (NCC) [46•]. This correlate is defined as a neural architecture or an assembly of neurons in which information is represented (if it is represented) in the conscious space.

In order to make the problem more tractable, Crick and Koch [46•] hypothesize that the many aspects of consciousness, such as pain, emotion, thought, vision and self-awareness, all employ similar mechanisms. Therefore, one can choose the more readily experimentally accessible problem of visual consciousness in the hope of generalizing its solution to the other types of consciousness. The authors suggest that visual consciousness in humans serves to produce the best interpretation of a given stimulus or scene, and to make it available to the parts of the brain that contemplate and plan voluntary motor actions. When visually aware of an object, the brain constructs an explicit, multilevel, symbolic interpretation in terms of representations of the object's features. Each of these feature representations is coded by a localized, homogenous group of neurons, and might itself be distributed over different parts of the visual system. Both attention and some form of short-term memory seem likely to be involved in this first step. Studying these mechanisms might help to locate the NCC, which should, according to Crick and Koch [46•], receive visual information and transmit it, without recoding it, to parts of the brain that plan voluntary action. Investigations of the connections between the highest levels of the visual hierarchy and the premotor and prefrontal cortices are therefore indicated. Crick and Koch [46•] have tentatively suggested that the NCC could be confined to layers 5 and 6 of the cortex.

Dynamic core model

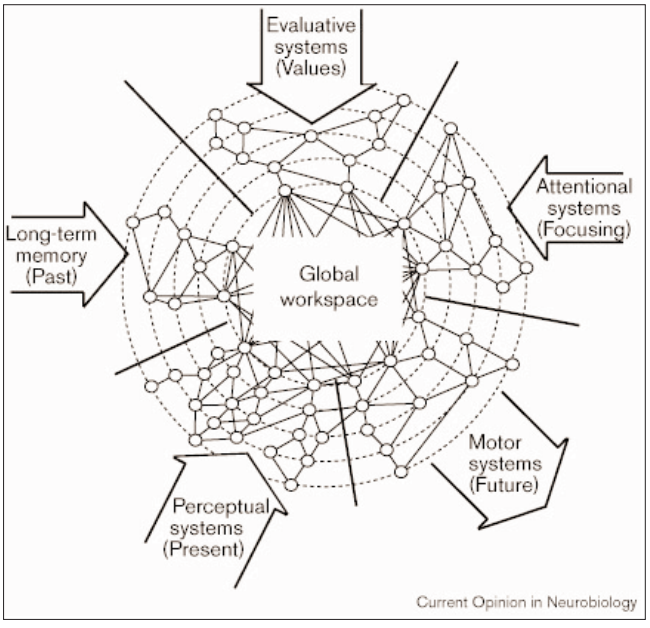
Tononi and Edelman [47••,48] have followed a more theoretical and speculative approach in their investigations of consciousness. They propose that the neural correlate or substrate of consciousness is an ever-changing ensemble of neurons, called the 'dynamic core', which is defined (but not formalized) as a subset of neurons that interact more strongly (by over an order of magnitude) with each other than with other neurons. Neurons that can be part of this subset are the corticothalamic neurons ([49••]; see also [50] for a review of early work on the role of thalamocortical loops in consciousness) as well as others from neighboring regions. They envision that the recruiting of neurons into the dynamic core takes place through the phase transition mentioned earlier (see section From cortical organization to perception). As presented above, Lumer *et al.* [30,31] observed in their simulation of the thalamocortical system the existence of roughly two types of stable activity : a low-intensity background activity, and a large-amplitude oscillatory mode of activity. Neurons are able to shift from one dynamic to the other by what the authors call a 'phase transition'. Tononi and Edelman propose that this same mechanism separates those neurons which participate to consciousness from the others: neurons in the dynamic core would be in the large amplitude, oscillatory mode, while the others would be in a low activity mode. This dynamic core would then originate in the dynamics of brain activity and, more precisely, in the varying strength of connections between neurons.

On the basis of this abstract construction, Tononi and Edelman [47••,48] have attempted to address two key properties of consciousness: integration and differentiation. Integration defines the property of a conscious experience to be unified and therefore not separable into independent components (the authors give the example of the bistability of ambiguous figures and of perceptual rivalry). Differentiation represents the enormous number of states available to consciousness over a short period of time.

Integration, by definition, would be a property of the dynamic core because it is bound together by the interactions between its constituents. It therefore cannot be divided into smaller independent clusters that would each account for a conscious state. In order to test this feature of their theory, Tononi *et al.* [51] introduced the 'cluster index', a mathematical quantity that measures the extent to which a neural process is unified, as opposed to being just a collection of independent processes. Though successfully applied to simulated networks, further assumptions and mathematical developments are required before this index can be reliably used to test for the presence of functional clusters in biological networks. Indeed, the synchronization of cortical and thalamocortical areas that is observed in brain imaging studies — which the authors claim to be indirect evidence for their model [48,51] — can be accounted for by other mechanisms.

In the framework of Tononi and Edelman's model, the question of the differentiation of consciousness is directly related to an estimation of the number of states available to

Figure 3



Schematic representation of the global workspace model. The global workspace is composed of distributed and heavily interconnected neurons with long-range axons. It connects to a set of specialized and modular perceptual, motor, memory, evaluative and attentional processors. Figure reproduced from [53••].

the dynamic core. The authors attempt to answer this difficult question by introducing a quantitative measure of neural complexity that reflects the number of possible states of a system arising from interactions between its constituents [52]. Evaluation of neural complexity for simulated neural networks of diverse connectivity has shown that the value of neural complexity is low for networks of independent or strongly coupled neurons, but that the value is higher when neurons are grouped in strongly connected clusters that interact, in a patchy manner, by projections and reentrant circuits. This latter connectivity is typical of the brain and also of corticothalamic circuits. Therefore, the proposed dynamic core should have access to a large number of possible states and be very differentiated.

Finally, according to this theory, consciousness can be seen as the temporal evolution of the dynamic core along a trajectory in the space of possible conscious states, which shifts from state to state as information is integrated in the thalamocortical system.

So far, Tononi and Edelman have approached the dynamic core problem using physical arguments (such as phase transition) and statistical arguments (such as neuronal complexity and cluster index). An objective test of the explanative power of this theory (which the authors do not address) will be its ability to reproduce and account for cognitive aspects of brain activity. The dynamic core hypothesis also does not describe the extent to which processing can unfold unconsciously.

Global workspace model

In parallel with the modeling of Tononi, Edelman and colleagues [47••,48], Dehaene, Kerszberg and Changeux [53••] have more concretely tackled the issue of consciousness by a neuronal implementation of the global workspace that was proposed, on psychological grounds, by Baars [54] and others (see Figure 3). To simplify, they distinguish two main computational spaces in the brain. First, there is a processing network, which consists of a set of functionally specialized, parallel processors that are capable of a large amount of encapsulated non-conscious processing. The processors range from primary sensory processors (e.g. V1) or unimodal processors that combine multiple inputs within a sensory modality, up to heteromodal processors (e.g. 'mirror' neurons; see [55]) that extract highly processed and categorical information. Each processor mobilizes topologically distinct cortical domains with local or medium range connections.

The second computational space is the global workspace, consisting of a distributed set of cortical neurons with long-range horizontal excitatory projections. They receive and send back connections to homologous neurons in cortical areas enriched in this particular population of neurons. Pyramidal neurons of layers 2 and 3 are known to extend long-range cortico-cortical axons (including callosal axons, which cross the mid-line). It is thus proposed that the extent to which a given area contributes to the global workspace would be simply related to the fraction of pyramidal neurons comprising layers 2 and 3 (this fraction is particularly elevated in dorsolateral and inferoparietal cortices). In addition, these cortical neurons are reciprocally connected with layer 5 neurons, thus establishing self-sustained vertical circuits with thalamic nuclei.

The global workspace neurons are postulated to be the seat of a 'brain-scale' activity that can be assumed to represent or to index the content of consciousness. This activity is realized by the spontaneous and coherent firing of a set of neurons from the workspace, together with active top-down amplification (or extinction) of active processor neurons belonging to one of five large classes: perception, motor programming, long-term memory, evaluation, and attention circuits. Only one such active representation can occupy the workspace at a given time (i.e. reproducing the property of integration of consciousness), where it can either remain and resist changes or be spontaneously replaced by another. Access to evaluative circuits can mobilize an auto-evaluation loop that enables new representations to be generated and tested (i.e. differentiation). Finally, perceptual areas enable the outside world to influence workspace representations, while motor circuits give the brain the ability to act or to communicate these representations in gestures and words. Attention circuits also allow the workspace to amplify or attenuate

contributions from other specialized processors, or even to mobilize its circuits independently from the external world. Arousal and emotional signals can globally control workspace neurons, thus controlling the balance between conscious and non-conscious states.

The global workspace formalism is especially suited to the implementation of the spatio-temporal dynamics of the control function of consciousness; this control function may correspond to the supervisory attentional system of Shallice [56] or to the central executive of Baddeley [57]. With its connections to functionally specialized areas of the brain, the workspace is able to take over when the system faces unusual or effortful tasks and to test hypotheses in a projective style. After a trial-and-error selection period, a learning phase takes place that progressively transfers the task to specialized processors. The activity of the global workspace thus becomes free for other tasks, until another unusual problem presents itself.

To test this model, Dehaene *et al.* [53••] simulated its architectural dynamics in a computer-based neural network, and submitted it to a test involving the learning of new skills, specifically performance of the Stroop test. This simple test requires the subject to react correctly to different conflicting stimuli. The authors showed that the network was able to perform the Stroop test, activating the neurons of the global workspace every time unusual or error-prone situations arose, and relying on the specialized processors in the other cases. The model predicts defined spatio-temporal activation patterns of brain imaging — particularly the observed contribution of dorsolateral prefrontal and anterior cingulate cortex in the performance of effortful tasks (see e.g. [58]) and the correlation of prefrontal cortex activation with that of distant areas, specifically during periods of conscious learning [59].

We end this section by conjecturing on how consciousness and conscious thinking might be implemented within the theoretical frame of the workspace. Being able to acquire or generate and hold representations, the workspace can, in principle, carry out computations on representations, evaluating and linking them into a temporal chain using short-term memory. Representations must follow each other in a coherent and meaningful manner when compared to the exterior world and past events. The problem of elucidating the mechanisms that orchestrate this flow of 'mental objects' should be addressed by any reliable theory of consciousness.

Conclusions

The aim of this review was first to briefly summarize recent attempts to construct formal models of the association cortex, starting from elementary building blocks at the level of the synapse, and using neuro-realistic learning algorithms and formal neural architectures. A second aim was to illustrate that several of these models account for cognitive functions of the brain, ranging from sensory binding and perception to simple generalization, reward-motivated behavior and exploratory strategy — finally building up to primary consciousness.

Although the models reviewed here are fairly diverse in architecture and dynamics, a few concepts and ideas about brain development and cognitive functions emerge. The first is the notion of nested processing of information: this takes place, at the small scale, in the segregation of neurons into groups by selective stabilization of synapses; at the larger scale, in the reentrant connections that are crucial for solving the binding problem during perception; and at a larger scale still, in the evaluation, amplification or dissipation of representations in the conscious space. A second important concept is the contribution of reward processes, which complement perception as a means of interaction with the outside world and play a key role in cognitive learning by reinforcement. The third and last important concept is that of an internal 'generator of diversity' which, within a selectionist framework, acts in a top-down manner, providing to the organism the 'forward' competence to act creatively and to evolve within its environment.

The further implementation of these notions and their experimental testing in future studies might render more tractable the still poorly accessible aspects of secondary (or even higher) language-based types of consciousness.

Acknowledgements

We thank R Klink, M Kerszberg and J-P Bourgeois for useful discussions, and R Miles for critical reading of the manuscript. This work was supported by the Collège de France, the Centre National de la Recherche Scientifique, the Institut National de la Santé et de la Recherche Médicale, and the McDonnell Foundation. T Gisiger has a postdoctoral fellowship from the Ministère des Affaires Etrangères du Gouvernement Français as part of an exchange program between France and Canada.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. • Arbib MA, Érdi P, Szentágothai J: *Neural Organization: Structure, Function and Dynamics.* Cambridge, Massachusetts: MIT Press; 1998.
This is an outstanding presentation of anatomical data of the brain as well as of computational models, from the synaptic to network level, proposed to account for some of its functions.

2. • Rolls ET, Treves A: *Neural Networks and Brain Function.* Oxford, UK: Oxford University Press; 1998.
This is an excellent textbook presenting computational models of various areas of the brain.

3. Amit DJ: *Modeling Brain Function: The World of Attractor Neural Networks.* Cambridge, UK: Cambridge University Press; 1989.

4. Koch C, Segev I: *Methods in Neuronal Modeling: From Synapses to Networks.* Edited by Koch C, Segev I. Cambridge, Massachusetts: MIT Press; 1989.

5. Edelman GM: **Group selection and phasic re-entrant signalling: a theory of higher brain function.** In *The Mindful Brain: Cortical Organization and the Group-Selective Theory of Higher Brain Function.* Edited by Edelman GM, Mountcastle VB. Cambridge, Massachusetts: MIT Press; 1978.

6. Dehaene S, Changeux J-P, Nadal J-P: **Neural networks that learn temporal sequences by selection.** *Proc Natl Acad Sci USA* 1987, **84**:2727-2731.

7. Fitzsimonds RM, Song H-j, Poo M-m: **Propagation of activity-dependent synaptic depression in simple neural networks.** *Nature* 1997, **388**:439-448.

8. Hebb DO: *The Organization of Behavior: a Neuropsychological Theory.* New York: Wiley; 1949.

9. Heidmann T, Changeux J-P: **Un modèle moléculaire de régulation d'efficacité d'une synapse chimique au niveau post-synaptique.** *CR Acad Sci III* 1982, **295**:665-670 [Title translation: Molecular model of the regulation of chemical synapse efficiency at the postsynaptic level.]

10. Goldman-Rakic PS: **The 'psychic' neuron of the cerebral cortex.** *Ann NY Acad Sci* 1999, **868**:13-26.

11. Dehaene S, Changeux J-P: **The Wisconsin card sorting test: theoretical analysis and modeling in a neuronal network.** *Cereb Cortex* 1991, **1**:62-79.

12. Dehaene S, Changeux J-P: **A simple model of prefrontal cortex function in delayed-response tasks.** *J Cognitive Neurosci* 1989, **1**:244-261.

13. Changeux J-P, Dehaene S: **Neuronal models of cognitive functions.** *Cognition* 1989, **33**:63-109.

14. Changeux J-P, Courrège P, Danchin A: **A theory of the epigenesis of neuronal networks by selective stabilization of synapses.** *Proc Natl Acad Sci USA* 1973, **70**:2974-2978.

15. Katz LC, Shatz CJ: **Synaptic activity and the construction of cortical circuits.** *Science* 1996, **274**:1133-1138.

16. Gally JA, Montague PR, Reeke GN Jr, Edelman GM: **The NO hypothesis: possible effects of a short-lived rapidly diffusible signal in the development and function of the nervous system.** *Proc Natl Acad Sci USA* 1990, **87**:3547-3551.

17. Montague PR, Gally JA, Edelman ME: **Spatial signaling in the development and function of neural connections.** *Cereb Cortex* 1991, **1**:199-220.

18. Shouval H, Intrator N, Law CC, Cooper LN: **Effect of binocular cortical misalignment on ocular dominance and orientation selectivity.** *Neural Comput* 1996, **8**:1021-1040.

19. Rittenhouse CD, Shouval HZ, Paradiso MA, Bear MF: **Monocular deprivation induces homosynaptic long-term depression in visual cortex.** *Nature* 1999, **397**:347-350.

20. Durbin R, Mitchison G: **A dimension reduction framework for understanding cortical maps.** *Nature* 1990, **343**:644-647.

21. LeVay S, Connolly M, Houde J, Van Essen DC: **The complete pattern of ocular dominance stripes in the striate cortex and visual field of the macaque monkey.** *J Neurosci* 1985, **5**:486-501.

22. Goodhill GJ, Bates KR, Montague PR: **Influences on the global structure of cortical maps.** *Proc R Soc Lond B Biol Sci* 1997, **264**:649-655.

23. Sporns O, Gally JA, Reeke GN Jr, Edelman GM: **Reentrant signaling among simulated neuronal groups leads to coherency in their oscillatory activity.** *Proc Natl Acad Sci USA* 1989, **86**:7265-7269.

24. Sporns O, Tononi G, Edelman GM: **Modeling perceptual grouping and figure-ground segregation by means of active reetrant connections.** *Proc Natl Acad Sci USA* 1991, **88**:129-133.

25. • Singer W: **Neuronal synchrony: a versatile code for the definition of relations?** *Neuron* 1999, **24**:111-125.
This is an excellent review of different aspects of synchrony in the brain, including data analysis and plausible functions in the central nervous system.

26. Tononi G, Sporns O, Edelman GM: **Reentry and the problem of integrating multiple cortical areas: simulation of dynamic integration in the visual system.** *Cereb Cortex* 1992, **2**:310-335.

27. Leopold DA, Logothetis NK: **Multistable phenomena: changing views in perception.** *Trends Cogn Sci* 1999, **3**:254-264.

28. • Lumer ED: **A neural model of binocular integration and rivalry based on the coordination of action-potential timing in primary visual cortex.** *Cereb Cortex* 1998, **8**:553-561.
The authors given an illustration of the possible effect that the synchronization of neuron population activity in the early visual system has on perception in higher areas.

29. Dayan P: **A hierarchical model of binocular rivalry.** *Neural Comput* 1998, **10**:1119-1135.

30. Lumer ED, Edelman GM, Tononi G: **Neural dynamics in a model of the thalamocortical system. 1. Layers, loops and the emergence of fast synchronous rhythms.** *Cereb Cortex* 1997, **7**:207-227.

31. Lumer ED, Edelman GM, Tononi G: **Neural dynamics in a model of the thalamocortical system. 2. The role of neural synchrony tested through perturbations of spike timing.** *Cereb Cortex* 1997, **7**:228-236.

32. Edelman GM: **Neural Darwinism: selection and reentrant signaling in higher brain function.** *Neuron* 1993, **10**:115-125.

33. Berthoz A: *Le Sens du Mouvement.* Paris: Odile Jacob; 1997. [Title translation: The sense of movement.]

34. •• Sutton RS, Barto AG: *Reinforcement Learning: an Introduction.* Cambridge, Massachusetts: MIT Press; 1998.
This is an excellent review of the field of learning by reinforcement which provides a complete description of the mathematical algorithms and illustrates their usefulness by numerous examples of recent work in various domains.

35. Reeke GN Jr, Finkel LH, Sporns O, Edelman GM: **Synthetic neural modeling: a multilevel approach to the analysis of brain complexity.** In *Signal and Sense: Local and Global Order in Perceptual Maps.* Edited by Edelman GM, Gall WE, Cowan WM. New York: Wiley-Liss; 1990:607-707.

36. • Asaad WF, Rainer G, Miller EK: **Neural activity in the primate prefrontal cortex during associative learning.** *Neuron* 1998, **21**:1399-1407.
The authors present important evidence supporting the presence of rule-coding neurons in the prefrontal cortex of the monkey, confirming the theoretical predictions of the model introduced in [12].

37. Schultz W, Dayan P, Montague PR: **A neural substrate of prediction and reward.** *Science* 1997, **275**:1593-1599.

38. Montague PR, Dayan P, Sejnowski TJ: **A framework for mesencephalic dopamine systems based on predictive Hebbian learning.** *J Neurosci* 1996, **16**:1936-1947.

39. Egelman DM, Person C, Montague PR: **A computational role for dopamine delivery in human decision-making.** *J Cogn Neurosci* 1998, **10**:623-630.

40. • Suri RE, Schultz W: **A neural network model with dopamine-like reinforcement signal that learns a spatial delayed-response task.** *Neuroscience* 1999, **91**:871-890.
The authors present an interesting study of the role of dopamine neurons in reward and reward-motivated learning, including both theoretical modeling and experimental investigations of the spatial delayed response task.

41. Dehaene S, Changeux J-P: **Development of elementary numerical abilities: a neuronal model.** *J Cogn Neurosci* 1993, **5**:390-407.

42. Dehaene S, Changeux J-P: **A hierarchical neuronal network for planning behavior.** *Proc Natl Acad Sci USA* 1997, **94**:13293-13298.

43. Changeux J-P, Dehaene S: **Hierarchical neuronal modeling of cognitive functions: from synaptic transmission to the tower of London.** *CR Acad Sci Paris III* 1998, **321**:241-247.

44. Hauser MD, Carey S, Hauser LB: **Spontaneous number representation in semi-free-ranging rhesus monkeys.** *Proc R Soc Lond* 2000, in press.

45. Shallice T: **Specific impairments of planning.** *Philos Trans R Soc Lond B Biol Sci* 1982, **298**:199-209.

46. • Crick F, Koch C: **Consciousness and neuroscience.** *Cereb Cortex* 1998, **8**:97-107.
The authors supply a useful introduction to the problem of consciousness and possible experimental approaches, and define the concept of the neural correlate of consciousness (NCC).

47. •• Tononi G, Edelman ME: **Consciousness and the integration of information in the brain.** In *Consciousness: At the Frontiers of Neuroscience. Advances in Neurology,* vol 77. Edited by Jasper HH, Descarries L. Philadelphia: Lippincott-Raven; 1998:245-279.
The authors provide a self-contained review of their work on consciousness, including a clear description of the notions of neural complexity, clustering index and the 'dynamic core'.

48. Tononi G, Edelman GM: **Consciousness and complexity.** *Science* 1998, **282**:1846-1851.

49. •• Llinás R, Ribary U, Contreras D, Pedroarena C: **The neuronal basis for consciousness.** *Philos Trans R Soc Lond B Biol Sci* 1998, **353**:1841-1849.
This is a remarkable synthesis of experimental work done on the role of large-scale thalamic activity in functional states characterizing cognition and consciousness.

50. Llinás RR, Pare D: **Commentary: of dreaming and wakefulness.** *Neuroscience* 1991, **44**:521-535.

51. Tononi G, McIntosh AR, Russel DP, Edelman GM: **Functional clustering: identifying strongly interactive brain regions in neuroimaging data.** *Neuroimage* 1998, **7**:133-149.

52. Tononi G, Sporns O, Edelman GM: **A measure for brain complexity: relating functional segregation and integration in the nervous system.** *Proc Natl Acad Sci USA* 1994, **91**:5033-5037.

53. •• Dehaene S, Kerszberg M, Changeux J-P: **A neuronal model of a global workspace in effortful cognitive tasks.** *Proc Natl Acad Sci USA* 1998, **95**:14529-14534.
The authors describe a biologically realistic implementation of the global workspace and of the relevant processors in terms of a neural network model capable of passing the effortful Stroop test.

54. Baars BJ: *A Cognitive Theory of Consciousness.* Cambridge, UK: Cambridge University Press; 1989.

55. Iacoboni M, Woods RP, Brass M, Bekkering H, Mazziotta JC, Rizzolatti G: **Cortical mechanisms of human imitation.** *Science* 1999, **286**:2526-2528.

56. Shallice T: *From Neuropsychology to Mental Structure.* Cambridge, UK: Cambridge University Press; 1988.

57. Baddeley AD: *Working Memory.* Oxford, UK: Clarendon Press; 1986.

58. Bush G, Whalen PJ, Rosen BR, Jenike MA, McInerney SC, Rauch SL: **The counting Stroop: an interference task specialized for functional neuroimaging — validation study with functional MRI.** *Hum Brain Mapp* 1998, **6**:270-282.

59. McIntosh AR, Rajah MN, Lobaugh NJ: **Interactions of prefrontal cortex in relation to awareness in sensory learning.** *Science* 1999, **284**:1531-1533.