H.B.M. Uylings, C.G. Van Eden, J.P.C. De Bruin, M.G.P. Feenstra and C.M.A. Pennartz
Progress in Brain Research, Vol 126

CHAPTER 14

Reward-dependent learning in neuronal networks for planning and decision making

Stanislas Dehaene[1,*] and Jean-Pierre Changeux[2]

[1]INSERM U. 334, Service Hospitalier Frédéric Joliot, CEA/DSV, 4 Place du Général Leclerc, 91401 Orsay Cedex, France
[2]Neurobiologie Moléculaire, Institut Pasteur, 25 rue du Dr. Roux, 75015 Paris, France

Introduction

Prefrontal cortex is thought to participate in supervisory attentional functions of the human brain by selecting a cognitive strategy that seems most appropriate to the task at hand and monitoring its execution (Luria, 1966; Shallice, 1988; Fuster, 1989). Yet how is the appropriateness of a strategy evaluated by prefrontal neurons? The neuronal network models developed by Jean-Pierre Changeux and myself have aimed at providing testable hypotheses concerning the organization of this evaluation and decision process as well as its putative cerebral and molecular bases (Dehaene and Changeux, 1989, 1991, 1993, 1996, 1997; Dehaene et al., 1998).

Our two basic hypotheses are that tentative plans or strategies of behavior are generated through the variable activation of neuronal assemblies in prefrontal cortex, which thus acts as a 'generator of diversity'; and that reward signals act to select among these activations those that are best adapted to the current environment. The models that we have introduced thus implement a generalized variation/selection scheme, which was initially explored under the name of 'reinforcement learning' by computer scientists (e.g. Sutton and Barto, 1998) and has also been called 'neural Darwinism' by neurobiologists (Edelman, 1987, 1993; Changeux and Dehaene, 1989).

This short review is focused on the biological mechanisms and functional significance of reward systems in this variation/selection scheme. We first briefly describe the three main roles that have been attributed to reward processes in models of learning by neural networks: the control of synaptic modification, the anticipation of further rewards, and the control of decision processes. We then focus on the latter process: how can neural networks implement decision making under the control of reward systems? A specific biological implementation is presented, and three simulations are described in which this mechanism was used. Finally, predictions for neurophysiological and brain imaging experiments are considered.

Three roles of reward in theoretical models of reinforcement learning

Most neural network simulations are framed in a supervised learning paradigm, in which an external teacher provides instructive signals which directly specify the patterns of neural output that must be learned by the network; more realistic however, from a biological standpoint, are simulations that rely on reinforcement learning. In this situation, the only feedback signal which is received by the simulated organism is an occasional reward which indicates the outcome of past actions, either right or

[*] Corresponding author: Tel.: +33 1 69 86 78 73; Fax: +33 1 69 86 78 16; e-mail: dehaene@shfj.cea.fr

wrong. The organism actively generates behavioral strategies and must use reward signals to optimize the adequacy of these strategies to the situation at hand. A thorough review of reinforcement learning algorithms can be found in (Sutton and Barto, 1998); here, we concentrate on three basic aspects of reward processing that were found useful in reinforcement learning models.

Use of reward signals in the control of synaptic modification

A first use of reward signals is in the control of changes of synaptic efficacy that underlie learning. In neural network models based on Hebbian learning or back-propagation, only information local to the synapses, such as recent pre- and post-synaptic activity, is used to alter synaptic efficacy. In reinforcement learning, however, an additional global signal coding for recent rewards is used to control the amplitude, and often the direction of synaptic change in order to adapt subsequent behavior to optimize the amount of reward received. For instance, our simulations have been based on a simple rule which, like the Hebb rule, is sensible to correlation of pre- and post-synaptic activity, but where the direction of the synaptic modification is determined by the sign of the reward signal received:

$$\Delta w = \varepsilon\, S_{\text{pre}}(2S_{\text{post}} - 1)R$$

where w is the synaptic weight, S_{pre} and S_{post} are the recent presynaptic and postsynaptic activities (between 0 and 1), and R is the reward (between -1 and $+1$). When the reward is positive, this equation implies that the classical Hebb rule is followed, which tends to stabilize recent activations. When the reward is negative, however, an anti-Hebbian rule is used which diminishes the probability of reproducing similar behavior in the future (Dehaene and Changeux, 1989, 1991, 1993, 1997; Dehaene et al., 1998). More complex rules, such as the linear reward-penalty algorithm, have been described (Sutton and Barto, 1998, Chapter 2).

Two properties that characterize the use of reward signals for synaptic modification may be relevant for real biological reward systems: First, information about rewards must be available to *all* synaptic sites at which reward-dependent plasticity is needed; this suggests that the reward must be transmitted by widely distributed neuromodulatory projections. This provides a functional interpretation for the well-known widespread distribution of neuromodulatory noradrenergic, serotoninergic, cholinergic, and dopaminergic projections to the cortex, although many of these systems may of course be implicated in non-reward-related global modulation functions such as arousal. Second, the above-described reward mechanism leads to behavioral adaptation on a slow time scale; due to the necessity of small and widespread cumulative synaptic modifications, learning typically takes hundreds to thousands of trials. Thus, this mechanism is compatible with the time scale of operant conditioning procedures in animals, but less so with the fast behavioral adaptation seen in humans and several other animal species. Further below we propose another mechanism for such fast reward-dependent adaptation.

Anticipation of reward

A second aspect of reward processing which has been found useful in neural network models is the anticipation of reward, also called 'value prediction' (Friston et al., 1994; Sutton and Barto, 1998), 'reward expectation' (Schultz et al., 1997) or 'auto-evaluation' (Dehaene and Changeux, 1991). In many environments, rewards are often infrequent; for instance, for a predator, the reward of eating the prey comes after a long non-rewarded chase. Likewise, for the backgammon player, the ultimate reward, winning the game, only comes after a long series of trials whose value remains uncertain until the very end. To circumvent this problem, theorists have shown that it is useful to incorporate an internal mechanism of reward prediction which anticipates on future external rewards. The output of this reward expectation system, rather than the actual external reward itself, is then used to direct adaptative behavioral changes. Possessing such an auto-evaluation loop is advantageous because it speeds up learning and partially solves the credit-assignment problem since each action can be immediately associated with an increase or decrease in the probability of subsequent rewards (Sutton and Barto, 1998). Most importantly, it gives the organism access to an internal mode of 'mental simulation' in which various courses of action can be evaluated without taking the risk of trying them out on the external world (Dehaene and Changeux, 1991, 1997).

Schultz and his collaborators (Schultz et al., 1993, 1997) have suggested that a circuit involving dopaminergic neurons in the ventral tegmental area and substantia nigra implements the expectation of rewards. Dopaminergic neurons normally fire in response to various appetitive stimuli such as food, but in the course of learning they can also become responsive to stimuli such as a light or a tone, that are not themselves primary rewards, but that reliably signal subsequent reward delivery. Montague and colleagues (Montague et al., 1996; Schultz et al., 1997) have suggested that the characteristics of this reward expectation property can be captured by a theoretical model of reinforcement learning, the temporal-difference (TD) algorithm, which was originally developed for computer-science purposes by Barto and Sutton (see Sutton and Barto, 1998). Currently, this specific theory, which interprets dopamine signals as indicating deviation from anticipated reward, remains controversial (see e.g. Pennartz, 1995; Redgrave et al., 1999; Spanagel and Weiss, 1999). Yet more generally, there is no doubt that the nervous system incorporates auto-evaluation mechanisms, which are at least reflected in, if not causally related to, the firing properties of dopaminergic neurons as well as cortical prefrontal and parietal neurons (e.g. Watanabe, 1996, Platt and Glimcher, 1999).

Selection of an appropriate decision

A third use of reward signals is in the direct control of neural activity. It is often necessary for an organism to react immediately to the occurrence of a positive or negative reward; for instance, a bee which evaluates the potential reward value of several flowers must rapidly decide in favour of one or the other (Montague et al., 1995). The synaptic modification mechanisms discussed earlier in this section are too slow to support such fast, reward-based decision; hence, neural networks models have also incorporated additional hypotheses about how rewards lead to explicit changes in on-line behavior. In many cases, no explicit biological mechanism has been proposed for this important function of reward systems. For instance, in the work of Montague and colleagues (Montague et al., 199, 1996; Schultz et al., 1997), an unspecified 'action selection' mechanism is postulated to lead the simulated organism to select the action which is associated with the greatest expected reward.

Our models include an implementation and simulation of an elementary mechanism of decision selection guided by an auto-evaluated reward (Dehaene and Changeux, 1989, 1991, 1993, 1997; Dehaene et al., 1998) (see Fig. 1). In these models, prefrontal connectivity is modeled at a coarse level by postulating that various clusters of prefrontal neurons, with a high level of spontaneous activity, encode a repertoire of rules whose activation modulates a lower-level sensori-motor network.

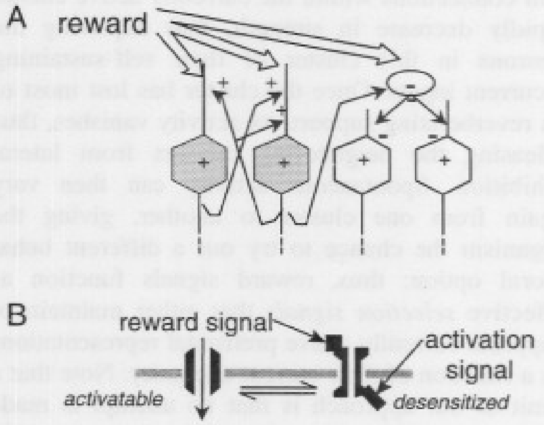

Fig. 1. Theoretical mechanism for the cellular and molecular implementation of a reward-dependent change in neuronal activity associated with decision making. Clusters of neurons in prefrontal cortex maintain a long-lasting activity through self-sustaining recurrent connections (top). The efficacy of those connections is assumed to be submitted to reward-dependent desensitization, thus allowing for a rapid change of activity following errors. One plausible molecular mechanism assumes a fast transition of postsynaptic receptor molecules to a desensitized state when a conjunction of a reward signal and a recent activation signal is present (bottom) (redrawn from Dehaene and Changeux, 1991).

Clusters are postulated to have a strong recurrent connectivity which implies that, although mathematically their activity can vary continuously anywhere between 0 and 1, they have two *stable* modes of activity: one in which the cluster is inactive (activity close to zero), and the other in which activity remains at a high level for a prolonged period (activity close to 1). Once activated, clusters can therefore remain in a state of self-sustained activation for a long duration through their local reverberating circuitry.

Action selection is implemented by a destabilization mechanism. Negative reinforcement, when impacting on an excitatory synapse between two currently active neurons, is assumed to cause a fast synaptic desensitization with a time scale of a few tens of milliseconds; later, the synapse spontaneously recovers its original strength with a slower time scale of a few seconds (for mathematical implementations of corresponding updating rules, see e.g. Dehaene and Changeux, 1989, 1991, 1997). The net result of this mechanism is that whenever negative reinforcement is received, recurrent connections within the currently active cluster rapidly decrease in strength, thus depriving the neurons in this cluster of their self-sustaining recurrent inputs. Once the cluster has lost most of its reverberating support, its activity vanishes, thus releasing the neighboring clusters from lateral inhibition. Spontaneous activity can then vary again from one cluster to another, giving the organism the chance to try out a different behavioral option; thus, reward signals function as effective *selection signals* that either maintain or suppress currently active prefrontal representations as a function of their current adequacy. Note that a limit of our approach is that no attempt is made here to solve the temporal credit-assignment problem; for simplicity, we merely assume that, in order to be selected, the prefrontal representation must still be active at the time when reward is received.

At the molecular level, the reward signal is postulated to be a neurotransmitter such as dopamine, acetylcholine or a coexisting messenger exerting a global modulatory action either via volume transmission or via targeted synaptic triads. Although most models of synaptic modification are based on the coincidence-detection properties of the NMDA glutamate receptor, our own tentative molecular mechanism for how a reward signal is used in decision selection is based on the known allosteric properties of a large body of non-NMDA receptor molecules, the archetype of which is the nicotinic acetylcholine receptor (Changeux, 1981). The latter can exist under four different states accessible via discrete conformational transitions: a resting, activatable state (with ion channel closed), an active state (ion channel open), and two desensitized states, respectively with fast (~100 ms) and slow (seconds) dynamics, in which the ion channel is closed.

Our proposed mechanism (Fig. 1) assumes that fast synaptic depression can be achieved through a desensitization reaction, in which postsynaptic receptor molecules switch to a desensitized state. We assume that the desensitisation reaction is enhanced by the co-occurrence of two signals converging on the same postsynaptic receptor molecules. The first one, endogenous to the postsynaptic cell, signals the recent activation of the synapse; this role may be assumed, for instance, by the high local intracellular concentration of calcium or a high extracellular concentration of neurotransmitter or of co-existing messengers. The second signal, diffused to all synapses throughout the relevant network, indicates a recent negative reward. Such gating of synaptic modifications by reward may be achieved for instance by diffuse neuromodulatory projections of catecholamine neurons from the mesencephalon to the prefrontal cortex. The simultaneous reception of these two converging signals would trigger a conformational change of receptor molecules into a state where the ion channel is closed, and thus the synapse depressed; recovery by the reverse reaction would occur on the 0.1–1 s. time scale.

This scheme is made more plausible by the observation that dopaminergic inputs to prefrontal cortex participate in 'synaptic triads' (Williams and Goldman-Rakic, 1993) : many of them are precisely targeted to dendritic spines on which a glutamatergic synapse from another prefrontal neuron is already present, thus putting them in an ideal position to modulate the efficacy of cortico-cortical excitatory connections between prefrontal neurons as required by the model.

Applications to frontal cortex tasks

Simulations of this variation-selection scheme have successfully accounted for the main features of several tasks that depend on prefrontal cortex integrity in humans, such as the Wisconsin card sorting test (Dehaene and Changeux, 1991), the Tower of London test (Dehaene and Changeux, 1997) and the Stroop test (Dehaene et al., 1998). We consider these in turn.

A simple example: the Wisconsin Card Sorting test

A classical test of prefrontal cortex in humans, which directly involves reward processing, is the Wisconsin card sorting test. In this task, subjects are required to infer a rule according to which they have to sort cards; feedback from the experimenter takes the form of a simple positive or negative reward (correct or incorrect). A simple neural network simulation that passes this test has been proposed (Dehaene and Changeux, 1991; see also Levine and Prueitt, 1989). The global architecture of our network (Fig. 2) comprises two distinct levels of organization : a low level (level 1) that governs the orientation of the organism toward an object with a defined feature and would correspond to a visuo-motor loop including visual areas and the premotor cortex, and a high level (level 2) that controls the behavioral task according to a memory rule and would be homologous to the prefrontal cortex or closely related areas.

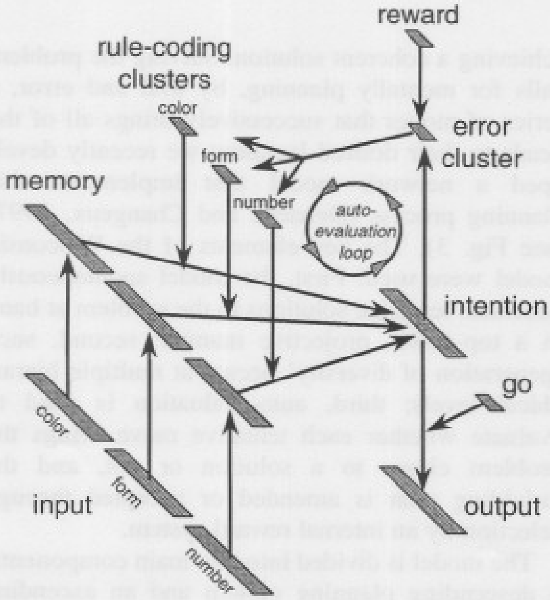

Fig. 2. Architecture of a neural network passing the Wisconsin card sorting test of prefrontal function (see text for details: adapted from Dehaene and Changeux, 1991).

A key feature of the model is that level 2 contains a particular category of cluster of neurons, referred to as rule-coding clusters, each of which code for a single dimension (position, colour, shape. . .) of the environment. Their connectivity is hierarchically organized in such a way that a rule-coding cluster globally regulates the efficacy of bundles of connections involved in the processing of particular features of the environment. For instance, the rule-coding cluster coding for 'sorting by color' selectively gates all connections associated with the processing of color information.

During the acquisition step, the layer of rule coding neurons is assumed to play the role of a generator of diversity; according to the above-described mechanism, the rule-coding clusters activate spontaneously, but because of lateral inhibition, only one cluster is active at a time. As long as negative reward is received, indicating that the correct rule has not been found, the activity of rule-coding cluster keeps changing at random with time in such a way that the organism is able to successively test different sorting rules upon its environment. In other words, a search by trial and error takes place, until a positive reward is received; then, the particular cluster active at this precise moment is selected. While this model is admittedly very simple - the range of available rules being quite limited and hardwired – it is able to pass the test, to successfully reproduce the behavior of normal subjects, and to fail in a manner similar to prefrontal patients if the frontal or reward units of the model are lesioned or removed.

In the course of the modelling of the Wisconsin card sorting task, we found it useful to introduce an auto-evaluation loop which, as described above, can short-circuit the reward input from the exterior. It allows for an internal evaluation of covert motor intentions without actualizing them as behaviors but by testing them by comparison with memorized former experiences. This element of architecture,

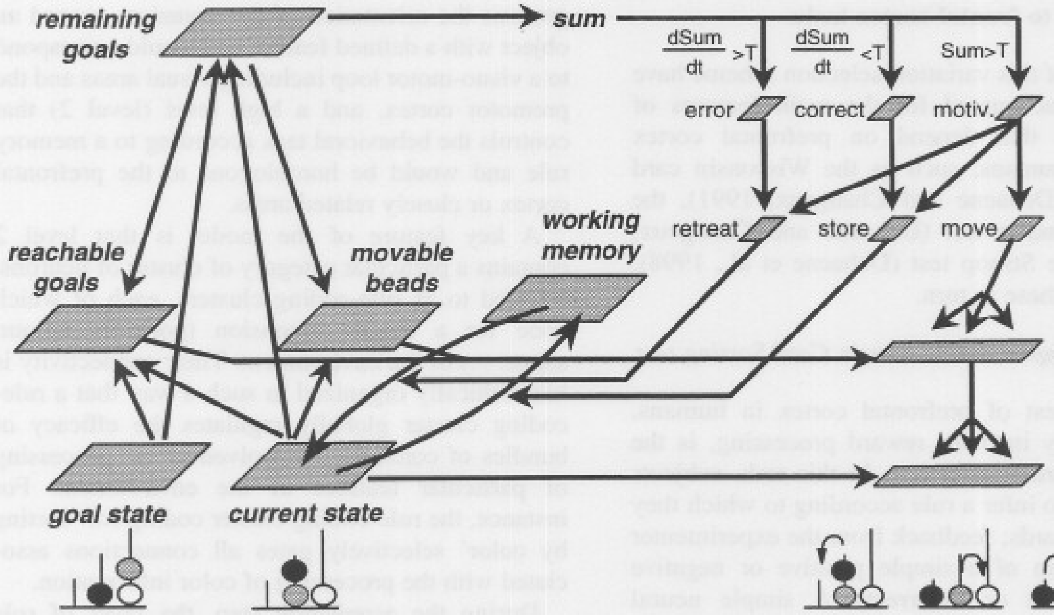

Fig. 3. Architecture of a neural network passing the Tower of London test of prefrontal function. The left column shows a hierarchy of neural networks (the 'ascending evaluation system') that compute an estimate of distance between the current state and the goal state. This system culminates in the delivery of reward and motivation signal, based on an internal evaluation of whether the distance to the goal has been recently increasing or decreasing. Those signals are used by a second hierarchy of neural networks (the 'descending planning system') that generates a tentative plan for solving the problem, with increasingly more refined details of the motor program being specified at lower levels (adapted from Dehaene and Changeux, 1997).

analogous in spirit to the 'adaptive critic' of Barton and Sutton's reinforcement learning models (Sutton and Barto, 1998), gives access to enhanced rates of learning via an elementary process of internal or covert mental simulation. Still, the 'mental experiments' authorized by this auto-evaluation loop are rather simple; for a more complex behavioral paradigm where the above neural architectural principles can be applied, we now turn to another classical test of prefrontal function, the Tower of London.

A more complex example: the Tower of London test

The Tower of London test (Shallice, 1982) is derived from the classical Tower of Hanoi test; it consists of moving three coloured beads, mounted on vertical rods of unequal length, from an initial position to a pre-specified goal. Patients with prefrontal cortex lesions experience difficulties in achieving a coherent solution. Solving the problem calls for mentally planning, by trial and error, a series of moves that successively brings all of the beads to their desired location; we recently developed a network model that implements this planning process (Dehaene and Changeux, 1997) (see Fig. 3). The key elements of the Wisconsin model were used. First, the model spontaneously generates tentative solutions to the problem at hand in a top-down, projective manner; second, such 'generation of diversity' occurs at multiple hierarchical levels; third, auto-evaluation is used to evaluate whether each tentative move brings the problem closer to a solution or not, and the unfolding plan is amended or accepted through selection by an internal reward system.

The model is divided into two main components: a descending planning system and an ascending evaluation system; in the descending planning system, the current plan unfolds internally at each of three hierarchical levels: plans, operations and gestures. Activation of plan units causes a series of activations at the lower operation level, with a fringe of variability; each activation of an operation unit in turn causes the sequential activation of two units at the lower gesture level, one for pointing to a bead and another to point to its new location. Hence, the descending planning system generates a variable, 'embedded' hierarchical sequence of internal moves. In essence, this system can be considered as a hierarchical version of the 'generator of diversity' used in the Wisconsin card sorting simulation.

The sequence of moves, however, is not entirely random, but is limited by constraints provided by the ascending evaluation system, a hierarchical system of areas that culminates, as in the Wisconsin card sorting model, in an auto-evaluated anticipation of reward. Based on the given of an initial state and a goal state, this system computes which beads are movable, which subgoals (misplaced beads) remain to be solved, and which subgoals can be directly fulfilled. When a bead can be placed directly at its final location, the corresponding operation is activated and executed immediately, without calling for plan unit activation. Only if no such move is available are plan units needed to activate the operation units and thus to generate a tentative move.

As in the Wisconsin card sorting model, a key element of the network is the internal reward system; in the Tower of London test, no external feedback is received at all about the correctness of tentative moves. In our model, reward units are now exclusively activated by an internal auto-evaluation loop; the total activation of remaining goal units is used to compute an internal estimate of distance to the goal: when this total activation decreases, it means that the last move brought the problem closer to a solution. Hence, in our network, the temporal derivative of total remaining goal unit activity is used to activate reward units. In the current simulation, we found it important to introduce separate units for positive or negative rewards (respectively activated by decreasing and by increasing total goal unit activation) and for motivation (activated whenever at least one remaining goal unit is active). All three units map onto plan units, but with slightly different connectivity patterns. While the motivation unit indiscriminately activates all plan units, thus 'turning on the generator of diversity', the positive and negative reward units are preferentially connected respectively to plan units that either validate the previous move and store it in working memory, or reject it and return to a previous memorized state. The net result is that tentative moves, generated semi-randomly by the descending planning system, are either maintained or rejected depending on whether the ascending evaluation system judges that the distance to the goal has decreased or increased.

When presented with Tower of London problems, the model generates solutions in a manner remarkably similar to normal subjects; in particular, it shows a gradient of difficulty similar to humans. Simple problems that call only for one, two or three direct moves are solved without trial-and-error. For more difficult problems, the network generates a complex internal sequence of trial-and-error that often rapidly converges to a valid solution (in less than 1000 update cycles or approximately 20 attempted moves, perhaps equivalent to about 30 seconds of reflection in humans). Measurement of the network's error rates and solution times indicate a close match to data from normal human subjects. When plan or reward units are deteriorated in the simulation, however, the resolution of complex problems becomes selectively impaired, as observed in actual experiments with prefrontal patients; the lesioned networks generate random trajectories that wander aimlessly in problem space. Their planning deficit can be attributed to an inability to guide the selection of motor operations by an internal evaluation of their relevance to reaching the goal, a characterization which also applies to human frontal patients.

The model makes several novel behavioral, neuropsychological and physiological predictions for experiments; most important in the present context is the role of internal reward systems in guiding the reasoning process. Diffuse catecholaminergic projection systems are predicted to be active and to play an important role in problem solving. Lesions of dopaminergic neurones may be simulated in the model by removing the reward units, while alterations of dopamine action on its receptors and/or related signal transduction mechanisms may be mimicked by altering the parameters determining the impact of reward units on plan units. In both cases, a severe planning deficit similar to that caused by plan unit lesions is observed, in good agreement with the deficits of Parkinsonian patients in the Tower of London test (Morris et al., 1988; Owen et al., 1992, 1995).

A general scheme for effortful tasks: the workspace model

We have recently extended these ideas and proposed a generalized model for the interaction between prefrontal cortex, reward systems, and the execution of effortful tasks (Dehaene et al., 1998). This model, which we call the workspace model, can be viewed as a generalization of the above models. It distinguishes two main computational spaces within the human brain (see Fig. 4): a unique global workspace composed of distributed and heavily interconnected neurons with long-range axons, and a set of specialized and modular perceptual, motor, memory, evaluative and attentional processors. We hypothesize that routine tasks that can be executed automatically, and without paying attention, are handled by specialized circuitry in the modular processors. Workspace neurons, however, are necessary for non-automatized effortful tasks for which the specialized processors do not suffice. They selectively mobilize or suppress, through descending connections, the contribution of specific processor neurons so as to implement, on the fly, operations that are not possible in the default configuration of processors. Hence, the processors and workspace levels correspond roughly to the two hierarchical levels of neural processing that were introduced in the Wisconsin model, as described above.

In order to perform the required interconnection of multiple modular processors distributed throughout the brain, workspace neurons must be characterized by their ability to receive from and send back to homologous neurons in other cortical areas horizontal projections through long-range

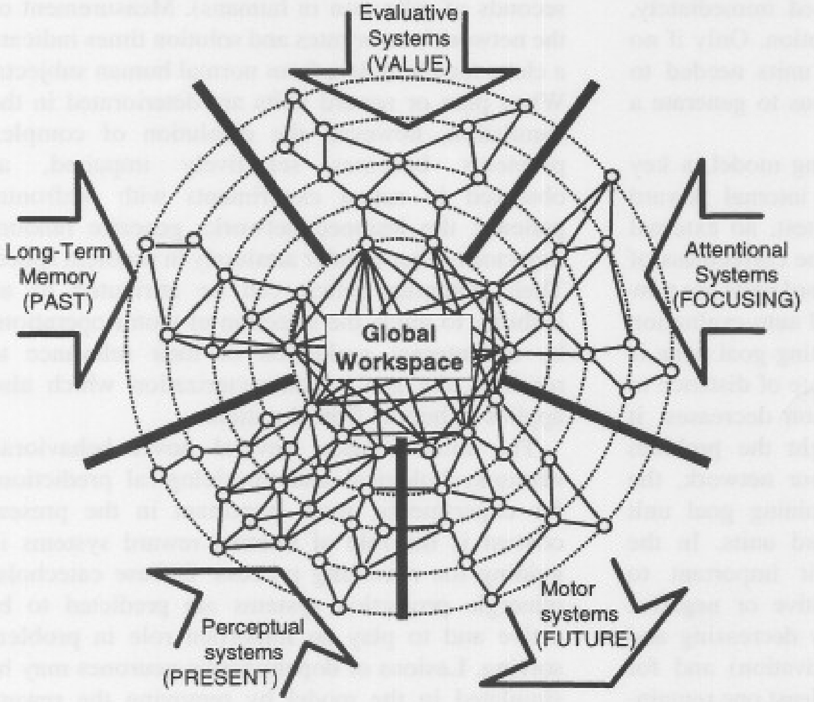

Fig. 4. Schematic representation of the five main types of neural processors connected to the global workspace (adapted from Dehaene et al., 1998).

excitatory axons (which may impinge on either excitatory or inhibitory neurons). Our view is that this population of neurons does not belong to a distinct set of 'cardinal' brain areas, but rather, is distributed among brain areas in variable proportions. It is known that long-range cortico-cortical tangential connections, including callosal connections, mostly originate from the pyramidal cells of layers 2 and 3, which give or receive the so-called 'association' efferents and afferents. We therefore propose that the extent to which a given brain area contributes to the global workspace would be simply related to the fraction of its pyramidal neurons contributing to layers 2 and 3, which is particularly elevated in von Economo's type 2 (dorsolateral prefrontal) and type 3 (inferior parietal) cortical structures.

We also postulate that workspace neurons are the specific targets of reward and vigilance signals that both modulate workspace activity; we use the same basic mechanism as above. In the course of task performance, workspace neurons become spontaneously co-activated, forming discrete though variable spatio-temporal patterns; those patterns are subject to modulation by a vigilance signal and to selection by a reward signal. The vigilance signal, analogous to the motivation unit of the Tower of London model, is postulated to have a gating effect on workspace unit activity. Thus, higher vigilance tends to activate workspace units, thus leading to greater spontaneous activation (see equations in (Dehaene et al., 1998)). The reward signal is identical to that used in the above simulations and serves a selection function: among the spontaneously activated workspace patterns, those that are associated with negative rewards are selectively eliminated. Note that the two systems are coupled, since the reward system is postulated to activate the vigilance system.

We have applied these principles to another well-known test of frontal function, the Stroop test (Dehaene et al., 1998) (see Fig. 5). This test comprises both an easy task (naming a color word) and a difficult task (naming the color of the ink in which a word is printed, when the word itself is the name of an incompatible color; e.g. saying 'blue' when seeing the word 'green' printed in blue ink). Like other previous simulations (Cohen et al.,

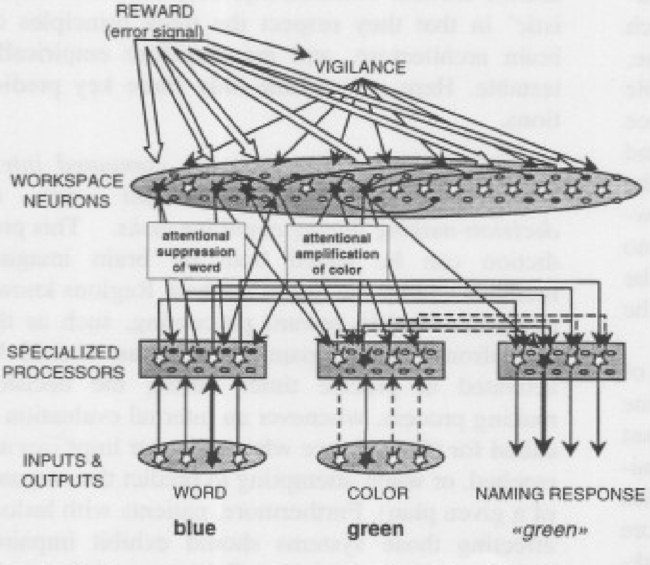

Fig. 5. Architecture of a neural network simulating effortful processing during the Stroop task (adapted from Dehaene et al., 1998). Three processors are simulated: perception of ink color, identification of written words, and spoken word production. All processors are connected bi-directionally to a large number of workspace neurons, themselves under the control of reward and vigilance signals. The appropriate state of workspace neuron activity, which is discovered by reward-dependent learning, allows the system to selectively amplify ink color information, which is then passed on to the word production system for naming.

1990; Kimberg and Farah, 1993), our model postulates that word naming is more automatized than color naming and is therefore the default processing strategy when words are presented. We therefore set up the processor connectivity with two input systems (one for color perception, the other for word recognition), one output system (for word production), and stronger connectivity between word perception and word production than between color perception and word production. All three processors were semi-randomly connected bidirectionally with a large set of workspace units.

The computer simulation showed that this model easily passed the easy word naming task, without needing to activate any workspace units. When the model was switched to the more difficult Stroop task, workspace activation initially increased during a search phase in which acquisition of the task was accompanied by an intense and highly variable activation of workspace units. This search phase ended when a workspace activation pattern was found which led to successful performance of the task. This pattern was characterized by a specific connectivity with processor units: the active workspace units tended to amplify the color perception and word production processor units, while deactivating the word perception units. The search phase was followed by an effortful execution phase, during which the workspace remained in a stable state of high activity; progressively, vigilance decreased as the task became routinized and transferred, through synaptic modifications, to the processor units and their interconnections. Following routinization, workspace activation was no longer needed for correct performance; but the workspace units reactivated sharply each time the network made an error.

The five observed stages – effortless execution of routine tasks; initial search during non-routine tasks; effortful execution with concomitant distant amplification or deactivation; progressive routinization; and error activation – are generic properties of the workspace model. They are therefore expected to characterize the activation of workspace neurons in various tasks other than the Stroop. Brain-imaging experiments indicate that dorsolateral prefrontal cortex (dlPFC) and anterior cingulate (AC) possess these properties. Both are active in effortful cognitive tasks, including the Stroop test, with a graded level of activation as a function of task difficulty (Pardo et al., 1990; Cohen et al., 1997; Paus et al., 1998). With automatization, activation decreases in dlPFC and AC, but it immediately recovers if a novel, nonroutine situation occurs (Raichle et al., 1994). AC activates in tight synchrony with subjects' errors (Dehaene et al., 1994; Niki and Watanabe, 1979; Carter et al., 1998). In the Wisconsin card sorting test, dlPFC activates when subjects have to search for a new sorting rule (Konishi et al., 1998). Finally, concomitant to dlPFC and AC activation, a selective attentional amplification is seen in relevant posterior areas during focused-attention tasks (Corbetta et al., 1991; Posner and Dehaene, 1994).

Summary and key predictions

Neural network models are only useful if they lead to empirical predictions. It is therefore natural that we end this chapter with a summary of the most important predictions made by our modeling approach. Although our proposed models do not incorporate highly detailed information about the known cortical connectivity, they are 'neuro-realistic' in that they respect the main principles of brain architecture, and are therefore empirically testable. Here, we outline only three key predictions.

1. Anticipations of rewards are computed internally and play an important role in decision-making during cognitive tasks. This prediction can be tested both by brain imaging methods and by the lesion method. Regions known to be involved in reward processing, such as the orbitofrontal and dopaminergic areas, should be activated at precise times during the decision making process, whenever an internal evaluation is called for (for instance when errors or impasses are reached, or when attempting to predict the outcome of a given plan). Furthermore, patients with lesions affecting those systems should exhibit impaired decision making (see e.g. Eslinger and Damasio, 1985; Damasio, 1994).

2. Reward inputs have a fast modulatory influence on prefrontal cortex activity. To be efficient in selecting an appropriate behavioral program, rewards should quickly affect prefrontal firing. This modulatory effect should have a fast onset (on a time scale of a few tens of milliseconds) and a duration that may vary from very short (a few hundreds of milliseconds or less) to long (seconds or minutes) depending on the time needed to discover an alternative program. This prediction, which should be tested by electrophysiological methods, also implies that a specific pattern of connectivity exists between reward systems and prefrontal neurons. As noted above, the observation that dopaminergic synapses to prefrontal pyramidal cells are often targeted to dendritic spines that already receive glutamatergic synapses from other neighboring neurons, thus forming synaptic triads (Williams and Goldman-Rakic, 1993), may provide the appropriate connectivity for this fast modulatory function.

3. Dorsolateral prefrontal cortex and anterior cingulate enter into an active search mode when a novel task is introduced or when errors are detected. Support for this hypothesis has recently been obtained with the observation of activation of these structures during error processing (Niki and Watanabe, 1979; Dehaene et al., 1994; Carter et al., 1998) and, most strikingly, during the search phase of the Wisconsin card sorting test (Konishi *et al.*, 1998).

Summary

Neuronal network models have been proposed for the organization of evaluation and decision processes in prefrontal circuitry and their putative neuronal and molecular bases. The models all include an implementation and simulation of an elementary reward mechanism. Their central hypothesis is that tentative rules of behavior, which are coded by clusters of active neurons in prefrontal cortex, are selected or rejected based on an evaluation by this reward signal, which may be conveyed, for instance, by the mesencephalic dopaminergic neurons with which the prefrontal cortex is densely interconnected. At the molecular level, the reward signal is postulated to be a neurotransmitter such as dopamine, which exerts a global modulatory action on prefrontal synaptic efficacies, either via volume transmission or via targeted synaptic triads. Negative reinforcement has the effect of destabilizing the currently active rule-coding clusters; subsequently, spontaneous activity varies again from one cluster to another, giving the organism the chance to discover and learn a new rule. Thus, reward signals function as effective selection signals that either maintain or suppress currently active prefrontal representations as a function of their current adequacy.

Simulations of this variation-selection have successfully accounted for the main features of several major tasks that depend on prefrontal cortex integrity, such as the delayed-response test, the Wisconsin card sorting test, the Tower of London test and the Stroop test. For the more complex tasks, we have found it necessary to supplement the external reward input with a second mechanism that supplies an internal reward; it consists of an auto-evaluation loop which short-circuits the reward input from the exterior. This allows for an internal evaluation of covert motor intentions without actualizing them as behaviors, by simply testing them covertly by comparison with memorized former experiences. This element of architecture gives access to enhanced rates of learning via an elementary process of internal or covert mental simulation.

We have recently applied these ideas to a new model, developed with M. Kerszberg, which hypothesizes that prefrontal cortex and its reward-related connections contribute crucially to conscious effortful tasks. This model distinguishes two main computational spaces within the human brain : a unique global workspace composed of distributed and heavily interconnected neurons with long-range axons, and a set of specialized and modular perceptual, motor, memory, evaluative and attentional processors. We postulate that workspace neurons are mobilized in effortful tasks for which the specialized processors do not suffice; they selectively mobilize or suppress, through descending connections, the contribution of specific processor neurons. In the course of task performance, workspace neurons become spontaneously co-activated, forming discrete though variable spatio-temporal patterns subject to modulation by vigilance signals and to selection by reward signals.

A computer simulation of the Stroop task shows workspace activation to increase during acquisition of a novel task, effortful execution, and after errors. This model makes predictions concerning the spatio-temporal activation patterns during brain imaging of cognitive tasks, particularly concerning the conditions of activation of dorsolateral prefrontal cortex and anterior cingulate, their relation to reward mechanisms, and their specific reaction during error processing.

References

Carter, C.S., Braver, T.S., Barch, D., Botvinick, M.M., Noll, D. and Cohen, J.D. (1998) Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, 280: 747–749.

Changeux, J.P. (1981) The acetylcholine receptor: an allosteric membrane protein. *Harvey Lectures*, 75: 85–254.

Changeux, J.P. and Dehaene, S. (1989) Neuronal models of cognitive functions. *Cognition*, 33: 63–109.

Cohen, J.D., Dunbar, K. and McClelland, J. (1990) On the control of automatic processes: a parallel distributed processing model of the Stroop effect. *Psychol. Rev.,*, 97: 332–361.

Cohen, J.D., Perlstein, W.M., Braver, T.S., Nystrom, L.E., Noll, D.C., Jonides, J. and Smith, E.E. (1997) Temporal dynamics of brain activation during a working memory task. *Nature*, 386: 604–608.

Corbetta, M., Miezin, F.M., Dobmeyer, S., Smulman, G.L. and Petersen, S.E. (1991) Selective and divided attention during visual discriminations of shape color and speed : functional anatomy by positron emission tomography. *J. Neurosci.*, 11: 2383–2402.

Damasio, A.R. (1994) *Descartes' error: emotion, reason, and the human brain*. NY: G.P. Putnam, New York.

Dehaene, S. and Changeux, J.P. (1989) A simple model of prefrontal cortex function in delayed-response tasks. *J. Cogn. Neurosci.*, 1: 244–261.

Dehaene, S. and Changeux, J.P. (1991) The Wisconsin Card Sorting Test: theoretical analysis and modelling in a neuronal network. *Cereb. Cort.*, 1: 62–79.

Dehaene, S. and Changeux, J.P. (1993) Development of elementary numerical abilities: a neuronal model. *J. Cogn. Neurosci.*, 5: 390–407.

Dehaene, S. and Changeux, J.P. (1996) Neuronal models of prefrontal cortical functions. *Ann. NY Acad. Sci.*, 769: 305–319.

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

Dehaene, S., Kerszberg, M. and Changeux, J.P. (1998) A neuronal model of a global workspace in effortful cognitive tasks. *Proc. Natl Acad. Sci. USA*, 95: 14529–14534.

Dehaene, S., Posner, M.I. and Tucker, D.M. (1994) Localization of a neural system for error detection and compensation. *Psycholog. Sci.*, 5: 303–305.

Edelman, G.M. (1987) *Neural Darwinism*. New York: Basic Books.

Edelman, G.M. (1993) Neural Darwinism: selection and reentrant signaling in higher brain function. *Neuron*, 10: 115–125.

Eslinger, P.J. and Damasio, A.R. (1985) Severe disturbance of higher cognition after bilateral frontal lobe ablation: Patient EVR. *Neurology*, 35: 1731–1741.

Friston, K.J., Tononi, G., Reeke, G.N., Sporns, O. and Edelman, G.M. (1994) Value-dependent selection in the brain: simulation in a synthetic neural model. *Neuroscience*, 59: 229–243.

Fuster, J.M. (1989) *The prefrontal cortex*. Raven: New York.

Kimberg, D.Y. and Farah, M.J. (1993) A unified account of cognitive impairments following frontal lobe damage: the role of working memory in complex organized behavior. *J. Exp. Psychol.: Gen.*, 122: 411–428.

Konishi, S., Nakajima, K., Uchida, I., Kameyama, M., Nakahara, K., Sekihara, K. and Miyashita, Y. (1998) Transient activation of inferior prefrontal cortex during cognitive set shifting. *Nature Neurosci.*, 1: 80–84.

Levine, D.S. and Prueitt, P.S. (1989) Modelling some effects of frontal lobe damage – novelty and perseveration. *Neur. Net.*, 2: 103–116.

Luria, A.R. (1966) *The higher cortical functions in man*. New York: Basic Books.

Montague, P.R., Dayan, P., Person, C. and Sejnowski, T. (1995) Bee foraging in uncertain environments using predictive hebbian learning. *Nature*, 377: 725–728.

Montague, P.R., Dayan, P. and Sejnowski, T.J. (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.*, 16: 1936–1947.

Morris, R.G., Downes, J.J., Evenden, J.L., Sahakian, B.J., Heald, A. and Robbins, T.W. (1988) Planning and spatial working memory in Parkinson's disease. *J. Neurol., Neurosurg. Psychiatry*, 51: 757–766.

Niki, H. and Watanabe, M. (1979) Prefrontal and cingulate unit activity during timing behavior in the monkey. *Brain Res.*, 171: 213–224.

Owen, A.M., James, M., Leigh, P.N., Summers, B.A., Quinn, N.P., Marsden, C.D. and Robbins, T.W. (1992) Fronto-striatal cognitive deficits at different stages of Parkinson's disease. *Brain*, 115: 1727–1751.

Owen, A.M., Sahakian, B.J., Hodges, J.R., Summers, B.A., Polkey, C.E. and Robbins, T. W. (1995) Dopamine-dependent fronto-striatal planning deficits in early Parkinson's disease. *Neuropsychology*, 9: 126–140.

Pardo, J.V., Pardo, P.J., Janer, K.W. and Raichle, M.E. (1990) The anterior cingulate cortex mediates processing selection in the Stroop attentional conflict paradigm. *Proc. Natl Acad. Sci. USA*, 87: 256–259.

Paus, T., Koski, L., Caramanos, Z. and Westbury, C. (1998) Regional differences in the effects of task difficulty and motor output on blood flow response in the human anterior cingulate cortex: a review of 107 PET activation studies. *NeuroReport*, 9: R37-R47.

Pennartz, C.M. (1996) The ascending neuromodulatory systems in learning by reinforcement: comparing computational conjectures with experimental findings. *Brain Res. Rev.*, 21: 219–245.

Platt, M.L. and Glimcher, P.W. (1999) Neural correlates of decision variables in parietal cortex. *Nature*, 400: 233–238.

Posner, M.I. and Dehaene, S. (1994) Attentional networks. *Trends Neurosci.*, 17: 75–79.

Raichle, M.E., Fiez, J.A., Videen, T.O., MacLeod, A.K., Pardo, J.V., Fox, P.T. and Petersen, S.E. (1994) Practice-related changes in human brain functional anatomy during non-motor learning. *Cereb. Cort.*, 4: 8–26.

Redgrave, P., Prescott, T.J. and Gurney, K. (1999) Is the short-latency dopamine response too short to signal reward error? *Trends Neurosci.*, 2: 146–151.

Schultz, W., Apicella, P. and Ljungberg, T. (1993) Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *J. Neurosci.*, 13: 900–913.

Schultz, W., Dayan, P. and Montague, P.R. (1997) A neural substrate of prediction and reward. *Science*, 275:1593–1599.

Shallice, T. (1982) Specific impairments of planning. *Philosoph. Trans. R. Soc. (Lond.), Ser. B*, 298: 199–209.

Shallice, T. (1988) *From neuropsychology to mental structure.* Cambridge University Press.

Spanagel, R. and Weiss, F. (1999) The dopamine hypothesis of reward: past and current status. *Trends Neurosci.*, 22: 521–527.

Sutton, R.S. and Barto, A.G. (1998) *Reinforcement learning: an introduction*. MIT Press: Cambridge, Mass.

Watanabe, M. (1996) Reward expectancy in primate prefrontal neurons. *Nature*, 382: 629–632.

Williams, S.M. and Goldman-Rakic, P.S. (1993) Characterization of the dopaminergic innervation of the primate frontal cortex using a dopamine-specific antibody. *Cereb. Cort.*, 3: 199–222.