Neuronal Models of Cognitive Functions Associated with the Prefrontal Cortex

J.-P. Pierre Changeux and *S. Dehaene*

Summary

Understanding the neural bases of cognition has become a scientifically tractable problem, and neurally plausible models are proposed to establish a causal link between biological structure and cognitive function. To this end, levels of organization have to be defined within the functional architecture of neuronal systems. Transitions from any one of these interacting levels to the next are viewed in an evolutionary perspective. They are assumed to involve:

1. the production of multiple transient variations and
2. the selection of some of them by higher levels via the interaction with the outside world.

The time scale of these "evolutions" is expected to differ from one level to the other. In the course of development and in the adult, this internal evolution is epigenetic and does not require alteration of the structure of the genome. A selective stabilization (and elimination) of synaptic connections by spontaneous and/or evoked activity in developing neuronal networks is postulated to contribute to the shaping of the adult connectivity within an envelope of genetically encoded forms. At a higher level, models of mental representations, as states of activity of defined populations of neurons, are suggested and their storage is viewed as a process of selection among variable and transient "pre-representations". Models are presented which can perform the delayed response task or the Wisconsin card sorting test and cognitive functions such as short-term memory, reasoning and handling of temporal sequences. Implementations of these mechanisms at the cellular and molecular levels are proposed. Finally, speculations are offered about plausible neuronal models and selectionist implementations of intentions.

Introduction

In biology, as in physics, the theoretical approach precedes experiment. Knowledge progresses by "conjecture and refutation" (Popper 1963), by the construction of *models* followed by submitting them to the experimental test. Any model is a "representation" of a natural object or process described in a coherent, non-contradictory and minimal form, if possible mathematical. To be

Christen et al. (Eds.)
Neurophilosophy
and Alzheimer's Disease

useful, the way in which it is formulated must allow comparison with outside reality. However, it cannot be expected that it will offer an exhaustive description of the latter. It may eventually be adequate, but will always remain limited.

Claude Bernard introduced a major distinction into life sciences by comparing anatomy (stable morphological organizations or "structures") with physiology (the dynamic processes by which an organism acts on the outside world or on itself). From there, the purpose of life sciences has become more precisely the determination of causal relationships between structure and function. This determination of relationships acquires a new dimension in the case of higher brain functions, where what is revealed of the mental state is still very often deliverately dissociated from subjacent neural organizations. Our purpose is completely opposed to any approach of this type since, in contrast, it concerns the creation of "a bridge" between neural sciences and mental sciences by developing neuronal models of cognitive functions.

Under these conditions, models of this type must be plausible at the neuro-biological level and not merely be "artificial", which would exclude from the start any comparison with neuronal reality. To be adequate, it is also necessary that the strucure-function relationship should provide correspondence between theoretical and experimentally observable variables in a *pertinent* manner and particularly *at the level of organization* involved (Changeux and Dehaene 1989).

As the physician P. Anderson wrote in 1972 "the ability to reduce everything to simple fundamental laws does not imply an ability to start from these laws and reconstruct the universe." The organization of living beings must be considered within the context of the evolution of species, and this evolution shows a consistent increase in complexity during palaeontological history. A first theoretical and essentially qualitative approach is to break down these complex organizations into hierarchical levels of organization whereby their appearance during evolution coincides with the appearance of new functions. For F. Jacob (1970), "living beings are therefore constructed by means of a series of packages. They are fitted together according to a hierarchy of discontinuous units or "integrons." Nerve cells are composed of molecules, but are assembled together into networks by dendrites and axons. The extension of this paradigm to clusters of neurons leads to differentiation of levels of organization within the brain itself. This must not be confused with the "levels" which Marr (1982) distinguishes in "Vision":

1. the "hardware" or neural machine;
2. the representation and algorithm and,
3. the computational theory.

In this case it concerns *levels of understanding* according to a scheme which perpetuates the cleavage between structure and function. This distinction may be useful for defining experimental pathways of approach or even for describing the system and its functional properties. However, it by no means takes any account of the levels *of organization* or of integration properly speaking and which may each highlight, at least in part, a description according to the terms of Marr.

Several models of the cleavage of functional brain organization into distinct "levels of integration" may therefore be proposed. Initially, the following levels were separated:

1. the level of elementary circuits and simple reflexes or fixed schemes of action (Kant's "sensitivity"?);
2. that of "groups of neurons" and of "symbolic representations" (Kant's "intendment"?);
3. that of complex assemblies of neuronal groups that we may refer to as "reason" (Kant) a "knowledge" (Newell) level (Newell 1982; Dehaene and Changeux 1989). However, we must expect even finer hierarchical cleavages.

The transition from one level of organization to the next is considered within the general conceptual context, which in this laboratory (Changeux et al. 1973; Changeux 1983) has always been that of an evolutionary epistemology (Popper 1966; Campbell 1974) inspired from Darwin's ideas (Darwin 1859; Poincare 1913) based upon:

1. a "blind" "generator of diversity" which introduces "variations" into the functional organization at the level being considered;
2. a mechanism of conservation and/or of propagation of the selected variation.

Within the context of the Darwinian scheme of evolution of species, variation occurs at the genome level (mutations, chromosome rearrangements, etc.), and the conservation of variations is performed by conservative replication of DNA and by propagation via sexual reproduction.

The application of this paradigm to the "internal" levels of organization of the brain does not postulate covalent variation of the genome but, in contrast, "epigenetic" variations of connectivity during development (time scale: years, minutes) or states of activity of neuronal clusters at levels of symbolic representation or architectures of reason (time scale: 0.1 second, minute). The proposed models (Dehaene and Changeux 1989, 1991) relate to the general problem of transition from the "symbolic" level to the "reason" level (Changeux and Dehaene 1989), dealt with in the particular example of the prefrontal cortex.

Functional Organization of the Prefrontal Cortex

The prefrontal cortex is the region of the neocortex in which the surface area has increased relatively most during the course of mammalian evolution; it increases from 3.5 % in the cat to 17 % in the chimpanzee and, finally, 29 % in humans (see Fuster 1989). Its histological organization is not uniform. In the 19th century, Brodmann subdivided the prefrontal cortex into several distinct territories, essentially on an anatomical basis. However, the latter mainly shows the six layers characteristic for the entire associative cortex, with a well-differentiated granular layer IV. The connectivity of the hundreds of millions of

neurons which compose it (900 millions according to some estimates) is one of extreme richness; this is most often established in a reciprocal or "re-entrant" manner according to Edelman's (1978) term between the intrinsic neurons which compose it, but also with the neurons of many regions of the encephalon. The frontal cortex exchanges connections with regions of the cortex which are hierarchically inferior to it, parieto-temporal associative areas, then primary sensory areas and motor areas. Of all the cortex, it is the domain which is most densely connected with the limbic system, which, as is known, participates in emotional responses. It is also linked to the thalamus as well as to the basal ganglia, which are involved in the control of movement. Finally, several nuclei of the reticular formation (containing neurotransmitters such as dopamine, noradrenaline, etc.) send highly divergent axons towards the frontal cortex where they control activity in a "global" manner. Examination of the connectivity of the frontal cortex allows at least three levels of hierarchical organizations to be defined, "nested" (Campbell 1974) one within the other. Very schematically, the prefrontal cortex is surrounded by an inferior level which corresponds to the areas of association, and by a more global level represented by the reticulo-frontal loops.

As Harlow noted in 1868, lesions of the frontal lobe in humans are accompanied by emotional disorders (hyperemotivity, character instability; Nauta 1971) as well as profound "cognitive" disorders which are expressed both by excessive obstinacy (perseverance in error) and then abnormal tendency to distraction, with a general decrease in critical activity. For Diamond (1988) the frontal cortex ensures "the correlation of information with space or time" and "inhibits dominant action tendencies." It constructs and updates "representations of the environment" (Goldman-Rakic 1987) which allow the subject to "plan and elaborate anticipations" (Teuber 1964, 1972) of actions on the surrounding world. For Shallice (1982) it constitutes a system of "supervisory attentive system", hierarchically superior to the "routine" "contention scheduling" system. It ensures the construction of plans in non-routine situations and *selects* schemes appropriate to those situations, all the while recording and taking into account errors likely to intervene in the realization of the plan. The prefrontal cortex produces "mental syntheses" (Bianchi) and is the site of "intentional behaviour" (Pavlov). It is also involved in making decisions and forming plans in relation to social behavior (Damasio et al. 1990).

It may be suggested that it corresponds to the "knowledge level" which theoreticians of artificial intelligence (Newell 1982) locate *above* the "symbolic" level and which may be the homologue of the "reason" level. Under these conditions, the prefrontal cortex would participate in the neural *architectures of reason* (Changeux 1988).

Functional Analysis of the Prefrontal Cortex by Various Tasks with Delayed Responses

In 1914, Hunter developed a behavioral test called the "delayed response task" which has since been very widely used with laboratory animals and even with

children for the experimental analysis of prefrontal functions (Diamond 1988; Piaget 1954; Fuster 1984; for review see Dehaene and Changeux 1989). The experimental design is as follows. A stimulus or cue object is initially presented to the subject at a precise point in a scene, then a screen falls and covers the scene from the subject's sight for a variable duration: two objects are then presented at two different sites and the subject must choose one of the two. The rule defining the correct choice varies with the type of task involved. In the delayed response (DR) task in its strict sense, and in task $A\bar{B}$ (A not-B; Piaget 1954), the rule is to choose the object which is located in the *position* occupied by the cue before the interval. In the DR task, the position of the cue is changed at random from one test to another, whereas in task $A\bar{B}$ its position is changed only after the child has been successful in accomplishing the task. In the so-called "delayed matching-to sample" task (DMS), the subject must choose an object identical to the cue irrespective of its position. Finally, a third task, called delayed alternation (DA), may be considered as part of this group of tasks. After having successfully performed a task at a given position, the subject must choose the alternative position in the following response. In all of these protocols, the subject learns the task during a training phase in which he receives the reward for each successful test (fruit juice in the monkey, playing with a toy in the case of children, etc.).

In all cases these are sensory-motor tasks which involve short-term memory of the subject and require selective attention. During the task, the subject makes a *decision* by comparing a test object with a memorized representation of the cue object. Finally, during training the subject performs an *induction* process in time and space by discovering the abstract rule (pertinent choice of feature) which governs the reinforcement.

In young children, systematic success in the $A\bar{B}$ test develops around the age of 7.5 months and performances improve up to 12 months. The young macaque monkey masters these two tests between 1.5 and 4 months, and ablation of the frontal cortex at birth interferes with mastery of the test. In the absence of the delay, an immature subject or injured adult passes the test but, if the delay exceeds 1–2 seconds, performance deteriorates and essentially becomes random (for review see Diamond 1988).

The delayed response task reveals early "cognitive" functions which are nevertheless of a high order and linked to the integrity of the prefrontal cortex.

Electrophysiological Recordings During the Execution of a Delayed Response Task

To our knowledge, only a few rare electrophysiological data exist on the *acquisition* of mastery of the delayed response task (Kubota and Komatsu 1985). The principal data available are recordings of individual neurons in the monkey (macaque) *during the performance* of the DR test after training (for review see Fuster 1989; Watanabe 1986a). Neurons of a first type come into action when the cue is presented (Fuster 1973; Niki 1974; 1975). Their activity represents

either an invariant early response appropriate to the *task*, which relates to the focussing of attention on the cue, or a response to the *test* itself, which distinguishes one cue from another, or finally, in some cases, both. Neurons of a second type, most often excitatory, change their activity in relation to the execution of the task; in general, their activity varies with the direction of movement of the hand which is required in the execution of the task (Kubota and Niki 1971; Watanabe 1986b). Their most remarkable feature is that their activity may anticipate the motor response by several seconds. Finally, the neurons of a third type are permanently active during the delay period, sometimes for a minute or more. Cells of this type are also encountered in the medio-dorsal thalamus and also in the temporal cortex, but to a much more limited degree. Neurons of this type are only observed in animals which have undergone training (it therefore does not concern a delayed sensory discharge). In addition, their activity is related to the state of *alertness* of the animal. Finally, there is a correlation between the activity of the cells during the delay period and the success of performance (distraction of the animal by an auditory stimulus during the delay period interferes both with the delay period activity and with success in the test). These neurons therefore establish a "temporal contingency" between presentation of the cue and motor performance. Their activity is not uniform. The discharge of a fraction of these cells tends to diminish during the delay period; these are *short-term memory cells* which participate in retention of the "representation" of the stimulus. The frequency of discharge of another fraction of neurons increases with time; they correspond to *motor anticipations* which prepare for movement. Coordination of the activity of these two categories of neurons ensures integration between short-term memory and preparation for motor action. Some cells which become temporarily active during the delay periods may participate in this coordination (Batner et al. 1981). At the end of the test, some prefrontal cells become active when the animal receives the reward and drinks the few drops of fruit juice received for a correct response. In contrast, other cells become active when the reward is expected but *is not* received (Rosenkilde et al. 1981).

These different types of cells are found throughout the prefrontal cortex; some are present in greater abundance in certain areas. The "reward" cells are present in the orbital regions, richly connected to the limbic system. The cells which are activated (or inhibited) during the delay period are mainly found in the region of the main sulcus of the dorso-lateral prefrontal cortex (Fuster 1989).

A Model of the Formal Neuron Network Accomplishing the Delayed Response Test

The objective of this modelling (Dehaene and Changeux 1989) is to construct a minimal and biologically plausible neuronal network which successfully passes the delayed-response task. Such a network must allow identification of the critical elements which are *necessary* for success in the task, and prediction of new properties subject to experimental validation.

The Formal Organism and Its Environment

The formal neuronal network is contained in a "formal organism" which interacts with a very restricted environment. This environment is *a priori* limited initially to the objects serving as a cue, then to some pertinent features of the latter which are likely to be taken into account by the formal organism: position (dimension 1), with two possibilities, right or left; color (dimension 2) with three possible hues; and finally no more than two objects may be presented to the formal organism at any given moment.

Each task is composed of successive tests and each test comprises four stages: presentation of an object, interval in the absence of the cue object, presentation of two objects, choice of object with reward or punishment, and an interval between two tests.

In tests of type 1 (analogues of DR or $A\bar{B}$), the correct object is the one having a position (dimension 1) coinciding with that of the cue; in those of type 2 (analogues of DMS), the correct choice is that of color of the cue (dimension 2).

The reward (or punishment) signal is applied from outside (by a master who decides his mark) or, under more natural conditions, as a result of the sensory qualities (taste, nutritional value, etc.) which are intrinsic to the object and recognized by the organism as favorable (or unfavorable) to survival (as a result of its past evolution). The reinforcement parameter covers an interval $[-1, +1]$ where 0 is neutral, $+1$ is maximum reward and -1 is maximum punishment.

Elementary Components of the Network

The network is composed of *"formal neurons"* of the McCulloch and Pitts type (1943) (see Amit 1989 for discussion), linked together by synaptic contacts of either the excitation or the inhibitory type. Each neuron is able to exist in two states, active (discharge) or inactive (rest). However, the states of activity of individual neurons (or synapses) are not explicitely modelled.

The basic unit of the network is a *"cluster of synergic neurons"*, the state of activity of which is assumed to code for an elementary "neural representation." This latter is analogous to Mountcastle's elementary module or "column" (Mountcastle 1978) or to Edelman's "group" of neurons (1978). It is defined and formalized here (Dehaene et al. 1987) as one hundred (or several hundred) neurons densely interconnected by excitatory synapses and, because of this, likely to exist in two self-sustained states of activity with either a high or a low frequency of discharge.

The neuron clusters are linked together by *"axon bundles"* of two types. The static bundles, not modulated by the activity of the network, propagate either lateral inhibition between clusters or the output of calculations performed by groups of clusters (or assemblies). The efficacy of the *modulated bundles*, for example between A and B, is regulated (to a *maximum* value) by the activity of a third neuronal cluster, for example C, called a modulator. The maximum efficacy value reached is itself variable and regulated by training (see below).

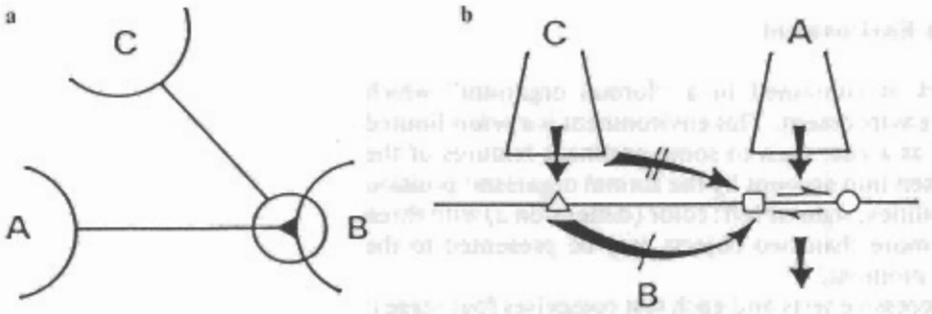

Fig. 1. Synaptic triad. Signals from synapse *C-B* modulate the efficacy of the neighboring *A-B* synapse. *b* is an enlarged view of *a* (taken from Dehaene et al. 1987)

Regulation of synaptic efficacy between A and B is undertaken in a "heterosynaptic" manner by C according to the *"synaptic triad"* scheme (Fig. 1); Dehaene et al. 1987) where the signal produced by synaptic terminal C acting on neuron B regulates, by an extra- or intracellular signal, the allosteric transitions (Heidmann and Changeux 1982; Changeux and Heidmann 1987) of the postsynaptic receptor of synapse A → B. All the synaptic triads between neurons belonging to clusters of neurons A, B or C constitute a "modulated bundle."

Architecture of the Network

A major feature of the architecture of the network is its differentiation into two hierarchical levels of organization (Fig. 2). *Level 1* (or execution level) includes two layers of neuron clusters, respectively *"input"* and *"output."* Each of the characteristic features of a given object is analyzed and coded by a particular cluster of input neurons. The output clusters are connected in an isomorphic manner with the input clusters, and the activity of the output clusters governs the orientation of the organism towards a defined object possessing a particular feature. *Level 2* (or regulation level) includes a layer of "memory clusters" and a layer of "rule-coding clusters," and controls the processing of an object according to a defined rule. The memory clusters, which are self-excitatory and mutually inhibitory, project isomorphically and modulate the input-output connections. Each cluster of rule-coding neurons codes not for a particular feature of the object but for *one* dimension which groups together several features of the object (in the very restricted case of the proposed model, there are only two). The clusters of rule-coding neurons project onto bundles which link input clusters with memory clusters and regulate their efficacy. By analogy with the primate neocortex, level 1 would correspond to a visuo-motor loop which includes secondary visual areas and the motor or pre-motor cortex, and level 2 would be identified with the prefrontal cortex.

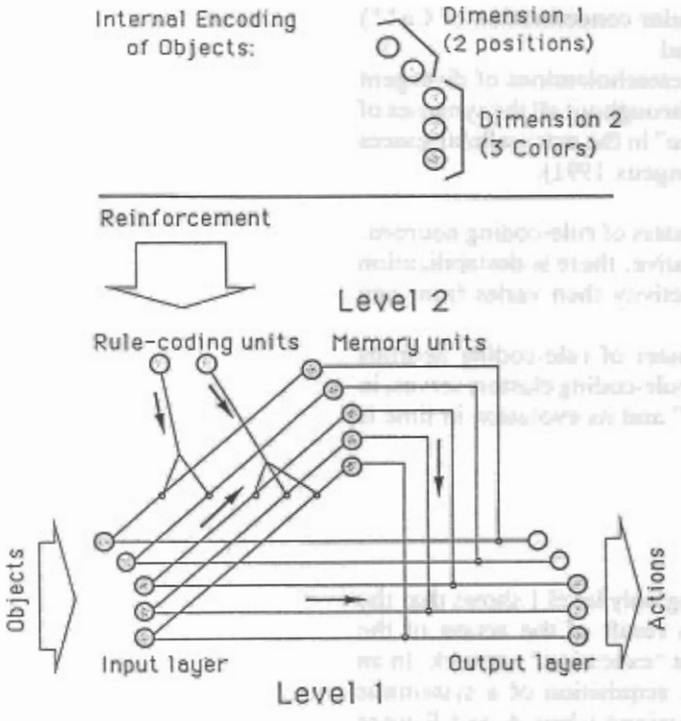

Fig. 2. Model of the role of the frontal cortex in learning and execution of delayed response tasks (taken from Dehaene and Changeux 1989)

Learning a Behavioral Rule

The organism learns a defined behavioral rule by interpreting a reinforcement signal which governs both modifications of synaptic efficacy and random variations in spontaneous activity of clusters of rule-coding neurons.

The reinforcement triggered "by return" as a result of the action of the organism on the environment during the learning process is "internalized" in the form of a parameter R which represents satisfaction (from 0 to +1) or dissatisfaction (from 0 to −1) of the organism. A first effect of R is to modulate the *maximal efficacy* of a synaptic triad according to Hebb's law. When R is positive and the postsynaptic neuron B (Fig. 1) is active at the same time, maximal efficacy increases; it decreases when the postsynaptic neuron is inactive. When R is negative, the rule is reversed.

Application of this rule is based on the allosteric properties of the postsynaptic receptor of synapse A → B. It is known that the nicotinic receptor for acetylcholine can exist in at least two desensitized states for which the ionic channel is closed (Changeux and Heidmann 1987). State I, of rapid access from the resting state, is involved in the functioning of the synaptic triad. The fraction of receptors in state D, of slower access, determines the maximum amplitude of variation in synaptic efficacy and is stabilized by the co-occurrence of two signals:

1. a *postsynaptic* signal (for example the intracellular concentration of Ca^{++}) which indicates recent activation of the cell, and
2. a diffuse *extracellular* signal (for example, the catecholamines of divergent reticulo-frontal pathways) which is propagated throughout all the synapses of the network for instance by "volume transmission" in the extracellular spaces (Fuxe and Agnati 1991; see Dehaene and Changeux 1991).

A second effect of R is to modify the activity of clusters of rule-coding neurons. When the organism is dissatisfied, R becomes negative, there is destabilization of all the rule-coding clusters, and spontaneous activity then varies from one cluster to another.

Learning takes place by selecting a defined cluster of rule-coding neurons according to its actual state of activity. The layer of rule-coding clusters serves, in some way, as a Darwinian "generator of diversity" and its evolution in time is under the control of the reinforcement signal.

Functional Properties of the Model

Simulation of the behavior of a network comprising only level 1 shows that the organism which possesses it is able to learn as a result of the action of the reinforcement loop on the triads of the input-output "execution" network. In an $A\bar{B}$ test, it ceases to orient at random. There is acquisition of a systematic orientation towards position A for which it was trained when A and B were presented simultaneously. However, like all infants or monkeys before the development of frontal connections, there will be systematic error when the position of the cue is changed from A to B. The organism which possesses only level 1 thus fails in the DR and DMS tasks. In contrast, it succeeds in all these tasks when it possesses levels 1 and 2.

The rule-coding neurons play a decisive role in the behavior of the organism. Their activity commands the memorization of a particular feature of the cue by modulating the efficacy of the connections between input clusters and memory clusters. If the rule-coding neurons which code for color are active, only the particular color of the cue will be memorized, but not its position. The neurons of the memory group themselves will govern the orientation of the organism towards the object possessing the memorized feature. In other words, the organism selects the object which possesses the characteristic feature of the cue *to the extent* that the rule-coding neurons which code for the particular *category* (position, color) to which this feature belongs are active.

The activity of the rule-coding neurons definitively "channels" the rule of behavior of the organism towards the choice. Learning therefore consists of a *search* among the various states of activity of rule-coding clusters to find the particular state which leads to the satisfaction of the organism. During learning, by successive "anticipations" based on the spontaneous variable and "blind" activity of rule-coding neuron clusters, the organism tests various features of the environment and selects the particular category of features of the object for which the "reward" is systematically positive.

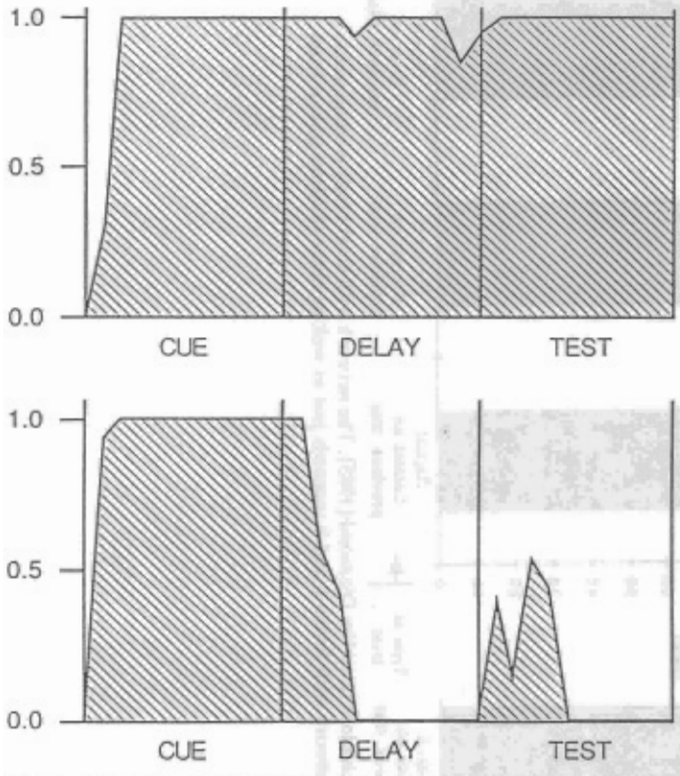

Fig. 3. Simulation of the activity of a memory neurons cluster during the delay. *Top*, the group remains active during the delay; performance in the next test is correct. *Bottom*, the group is inactivated due to internal noise; the organism now fails the test (taken from Dehaene and Changeux 1989)

The model allows simulation of the activity of defined neuron clusters during or after learning. In particular, neurons of the memory clusters display an activity which resembles that of neurons which are active during the delay period (see previous chapter) and the activity of which anticipates the behavior of the monkey during the choice when it is successful (but also when it fails; Fig. 3).

Simulation of the behavior of the formal organism shows that, with level 1 only, its behavior is analogous to the performances of infants aged from 7.5 to 9 months, of monkeys of 1.5 to 2.5 months or of monkeys with prefrontal lesions 17 (Fig. 4).

With level 2, the performances of the formal organism become practically identical to those of a child aged 12 months or of a rhesus monkey aged 4 months with respect to the learning of task $A\bar{B}$ or DMS. In addition, the organism is capable of passing from one task to another without difficulty.

Despite these successes, the formal organism modelled in this way presents three groups of limitations:

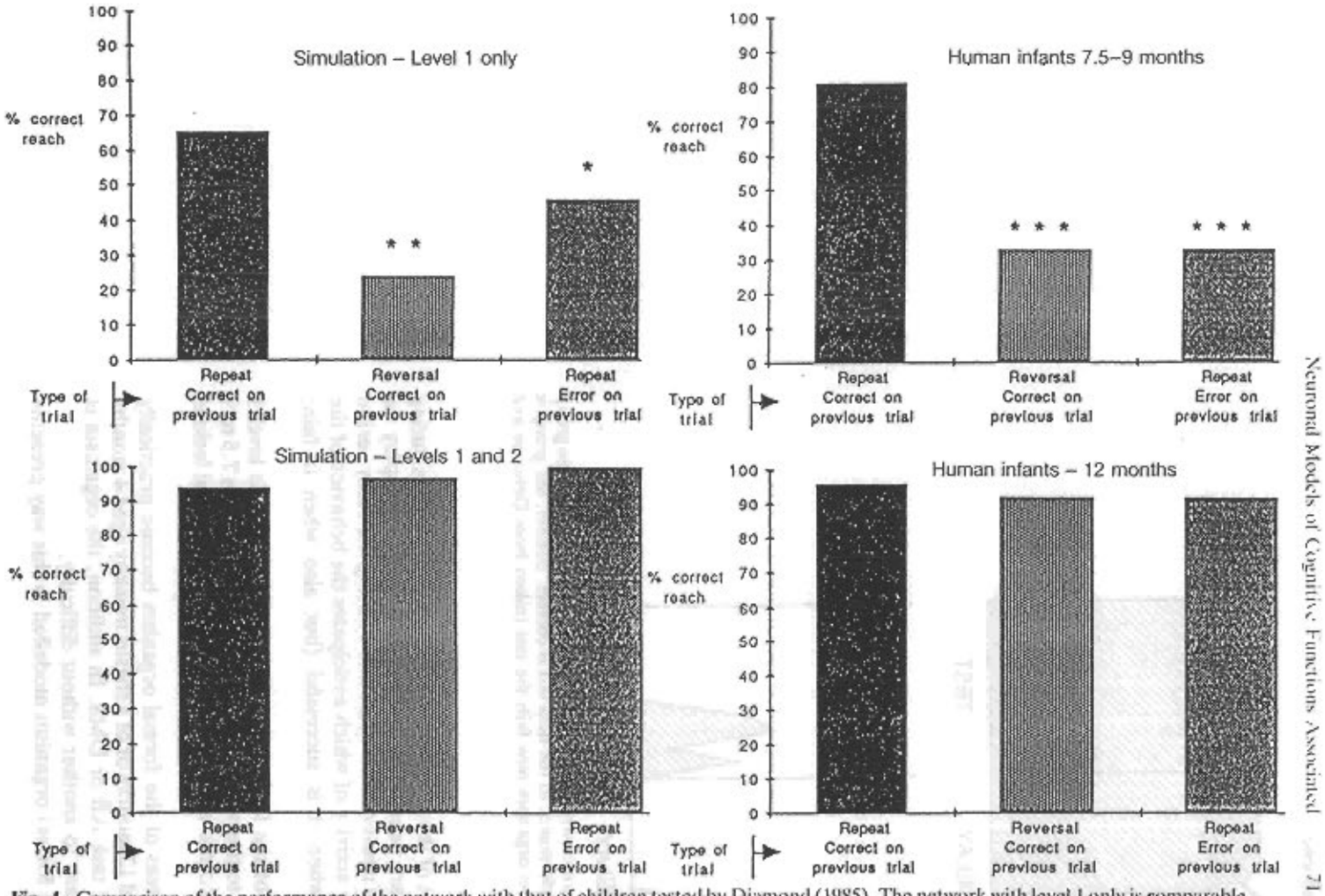

Fig. 4. Comparison of the performance of the network with that of children tested by Diamond (1985). The network with level 1 only is comparable to children before the development of frontal connections. The network with levels 1 and 2 succeeds just as well as older children (taken from Dehaene and Changeux 1989)

1. the sensory-motor tasks are highly reduced in the number of sensory categories or features and types of motor behavior;
2. the architecture is extremely simple: the number of formal neurons is six-seven orders of magnitude lower than that of the neurons of the prefrontal cortex in humans; and
3. the range of available rules is very small in size.

With the purpose of extending this research to more complex functions and to richer networks, a modeling of the Wisconsin card sorting test was investigated (Dehaene and Changeux 1991).

The Wisconsin Card Sorting Test

This test which is used to detect prefrontal cortex lesions classically consists of discovering the principle according to which a deck of cards must be sorted (Grant and Berg 1948; Milner 1963). The cards bear geometric figures of different shapes (triangle, star, cross or circle), color (green, red, blue or yellow) and number (1, 2, 3 or 4 figures). Four *reference* cards are permanently placed in front of the subject. The subject has another deck of cards called the *response* cards. He is asked to match each response card successively with one of the four reference cards. After each response, he is told whether this is correct or not. The subject tries to achieve the maximum of correct responses. The rule will, for example, be sorting according to color. Once the subject is systematically successful in this, the rule is changed, for example from color to shape. The subject must understand that the rule has changed and discover the new rule (Fig. 5).

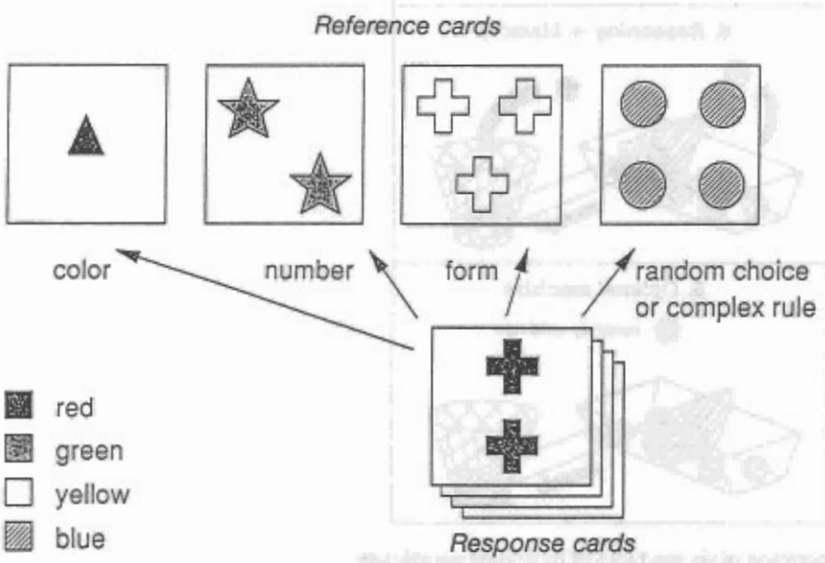

Fig. 5. Cues used in the Wisconsin card sorting test (taken from Dehaene and Changeux 1991)

Normal subjects do not always succeed in passing the test, particularly in the case of elderly subjects. However, subjects with a prefrontal lesion are systematically less successful than normal subjects. Frontal subjects make errors of a particular type called "perseveration" when they persist in using a rule which was initially correct even after they have been told that it is no longer in use. They exhibit difficulties in passing from one rule to another. The lesions which lead to the most marked deficit are located in the median frontal cortex.

Funtional analysis of the abilities of a formal cognitive system to pass the test (Dehaene and Changeux 1991) leads to the distinction of 6 "formal machines" according to the manner in which they select a new rule (Fig. 6).
The first three machines are "blind" in the sense that, each new rule is drawn at random from a repertoire of available rules without resort to reasoning. The simplest possible machine (random) draws a new rule entirely at random to replace the previous rule. The second, more complex [random + context] machine avoids again drawing a rule which has just been rejected. The third [random + memory] machine keeps an "episodic memory" of the previously rejected rules and draws only from among the remaining possible rules.

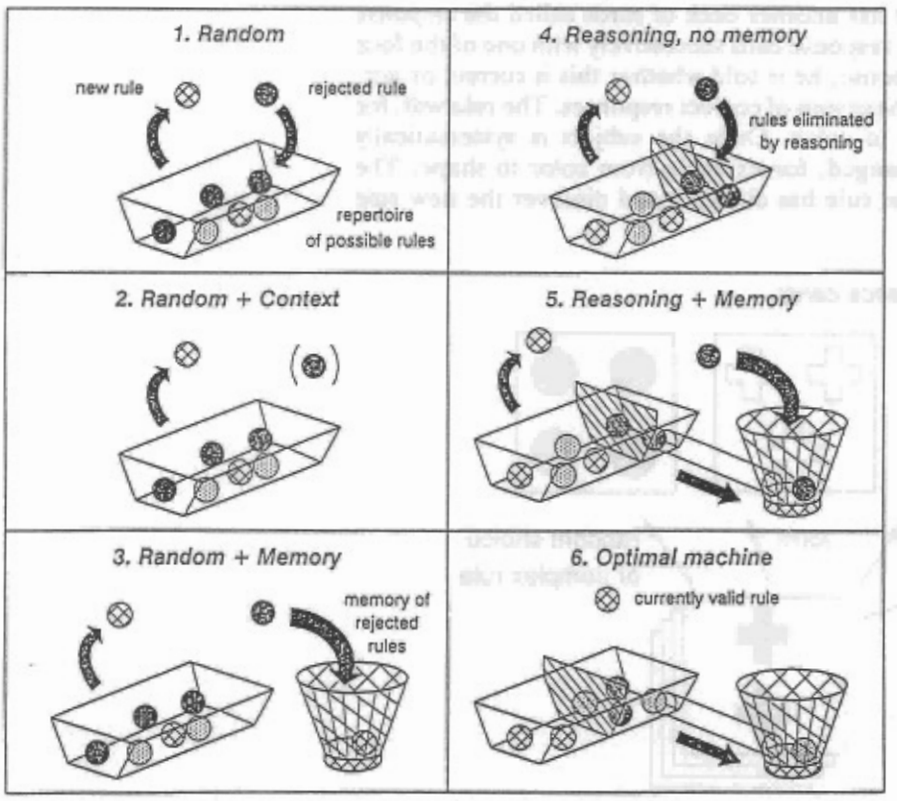

Fig. 6. Diagrammatic representation of the operation of six machines of increasing complexity learning the rules of the Wisconsin test by selection (taken from Dehaene and Changeux 1991)

The other three machines possess the additional faculty of rejecting rules by a type of tacit *"reasoning"* without having tested them overtly by trial and error. This is a very simple form of reasoning in the sense that, in the event of negative reinforcement, the machine eliminates all rules (in addition to that actually tested) which would lead to the same failure. The fourth machine [reasoning, no memory] utilizes reasoning in an extemporaneous manner without applying memory. The fifth machine [reasoning + memory] keeps in the memory a sketch of the previously rejected rules. Finally, the sixth, called "optimal," applies reason to the failures and memorizes the rules which have failed, but also applies reason to positive tests. Rules which have not given the same positive response are rejected and memorized as incorrect.

Comparison of the properties of these machines with the results of the Wisconsin card sorting test shows that this test does not allow all these properties to be tested. Of the three fundamental cognitive abilities of these machines – namely

1. ability to change the rule when punished,
2. memory of rules already tested and
3. *a priori* rejection of rules by reasoning – only the first is tested in a critical manner by the test.

A Formal Neuronal Architecture Able to Pass the Wisconsin Card Sorting Test

The model presented here (Dehaene and Changeux 1991) covers the major outlines of the architecture of the proposed formal organism in the case of tasks with delayed response, naturally with several major additions and modifications (Fig. 7).

The clusters of input neurons are more numerous since there are more categories and cue features involved in the test. Clusters of memory neurons are also present and receive projections from imput clusters with conservation of topography. There is competition between input clusters, with reciprocal inhibition so that only one feature is memorized for each dimension.

The memory neuron clusters project onto a new layer compared with the previous model. This layer is composed of neuron clusters which code for *"intentions"* of motor response that are distinct from the motor command itself. The model includes four clusters of intention neurons. Each code for the choice of a particular reference card, and activation of one of them excludes activation of the others. The intention is converted into an output command when an external "go" signal is received. Finally, like the previous model, this model includes clusters of rule-coding neurons which play a critical role in the performance of the test. Indeed, they modulate the connections between memory clusters and intention clusters according to a defined category (color, shape, number, etc.) and their activity varies with time during learning since each cluster which is active at a given moment inhibits the others. The organism uses them to test "hypotheses" of rules of behavior and selects a particular rule by interpreting a reinforcement signal. In fact, the network has been modelled in

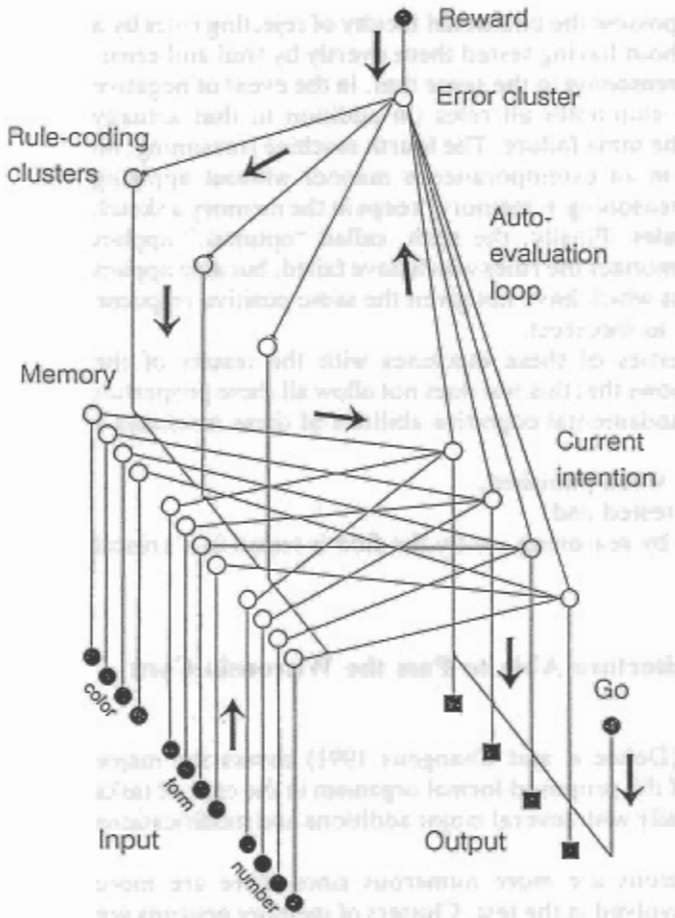

Fig. 7. Model of the role of the frontal cortex in the Wisconsin card sorting test. *Filled circles* represent the input clusters, *filled squares*, the output clusters, and *empty circles*, the internal clusters. Each *line* represents a bundle of synapses; auto-excitatory connections and lateral inhibitory connections within each assembly are not shown. On the input side, cards are coded along the dimensions of color, form, and number, and their features are stored in the short term by the self-excitatory connections of memory clusters. Memory clusters in turn activate the clusters defining the current intention for output. *Rule-coding clusters* modulate this transmission between *memory* clusters and *current intention* clusters, thus effectively deciding on the sorting rule. When the *go* unit is activated, activity coding for the current intention is in turn transmitted to output units. The subsequent entry of positive or negative reward (*top*) selects among the possible states of activity of the rule-coding layer, until the appropriate rule-coding cluster is activated. In the absence of an external reward, an auto-evaluation loop enables the system to reject rules autonomously, by evaluating the current intention with respect to memorized situations (taken from Dehaene and Changeux 1991).

such a way that each incorrect rule leads to punishment which, as in the previous model, destabilizes all the rule-coding neuron clusters so that they fluctuate in time and serve as the "generator of diversity."

Another novelty of the model is the differentiation of a *"cluster of error neurons"* which project and modulate the connections with rule-coding neuron clusters. The activity of error neurons is itself governed by reward signals so that a negative reward leads to short-term depression of excitatory connections in clusters of active rule-coding neurons. A molecular embodiment of this effect is allosteric regulation of postsynaptic receptor desensitization of the type described previously. This depression is spontaneously reversible and the speed of recovery is a crucial parameter which determines the memory range of the generator of diversity. If this speed is fast, the cluster of rule-coding neurons which has just been eliminated immediately enters again into the generator of diversity; it is a [random] machine. If the recovery speed is slow, a [random + context] machine is obtained which retains only the rule that has just been eliminated. Finally, when this speed is very slow, recovery extends over several consecutive tests and the network memorizes all the rules which have failed. It then behaves like a [random + memory] machine. The most original feature of the new model is the *"auto-evaluation loop"* which short-circuits the reward input from the exterior. This allows endogenous activation by intention clusters of error clusters, the efficacy of which is changed according to a classical Hebb's scheme. When a negative reward is received, the error neurons are activated and the connection linking intention clusters which are active at that moment with error clusters is reinforced. This intention is labelled as incorrect. Due to the persistence of activity in the error neurons, a new rule is tested within the rule-coding layer. This new rule is applied to the memorized features of the preceding cue, which produces a new distribution of intention cluster activity. If this distribution is identical to the previous one, the rule is rejected because the activity of the error cluster is maintained by potentiation of the intention/error connection, which prevents stabilization of the new rule. The "internal evaluation" of rules sequence is pursued until a correct rule is found.

Simulation of networks possessing auto-evaluation and memory shows a single trial percentage success rate much higher than that of the [random + memory] machine (98.4% versus 39.8%). Similarly, a network with an auto-evaluation loop but no memory is more successful than the [random + context] machine (Fig. 6).

Lesioning of the error cluster leads to slowing of learning and an increase in perseverations similar to those observed in frontal patients (see above). The inertia of the generator of diversity becomes very large. As in the case of the simple network, lesioning of rule-coding clusters interferes with the acquisition of a "systematic" rule of behavior. Lesioning of the auto-evaluation loop has no major qualitative effect on the behavior of the organism except for a loss of ability to reason, which significantly slows the learning process. However, it might offer a formal explanation of the "sociopathic" behavior resulting from ventromedian lesions of the frontal cortex (Damasio et al. 1990). Damasio et al. (1990) consider that this deficit is due to the inability to activate somatic states linked to the punishment or reward which the subject has experienced in association with specific social situations and which must be reactivated in connection with the anticipated result of a possible response. Injury to the intention/error connection might, according to our scheme, be the origin of this

type of syndrome, evidently within a context both verbally and socially richer than that which served for modelling.

Conclusion

The two proposed formal models of the neuron network take into account the characteristic functional abilities of the prefrontal cortex: success in various delayed response tasks and in the Wisconsin card sorting test. They are based on principles of molecular, cellular and histological architecture that are plausible at the neurobiological level. These models are extremely simple and might even appear simplistic to cerebral cortex specialists. Nevertheless, they provide several original and specific predictions able to delineate novel experimental tests. One bears on the existence of "rule-coding neurons," the activity state of which varies randomly during the learning period until a rule of behavior is selected. Another concerns the mechanism, or mechanisms, of reinforcement by "error neurons." On the one hand their activity is regulated by that of neurons coding for motor intentions, and on the other hand they exert a regulatory action on rule-coding neuron clusters.

At a more general level, the induction of rule by trial and error followed by selection integrates perfectly with evolutionary epistemology (Changeux and Dehaene 1989; Changeux et al. 1973; Changeux 1983; Popper 1966; Campbell 1974; Darwin 1859; Poincare 1913; Dehaene and Changeux 1991; Edelman 1978) and illustrates the concept of "mental Darwinism" (Changeux and Dehaene 1989; see also Campbell 1974). In this context, clusters of rule-coding neurons would constitute the "generator of diversity." Memorization by selection may be considered as a homologue of "amplification" since the organism will re-use the memorized trace repeatedly in its subsequent behaviors.

The models also illustrate the precise contribution of hierarchical levels of network architecture to defined behaviors and particularly:

1. the ability to generalize a rule acquired for a particular cue to an entire class of cues, or *systematicity* (Dehaene and Changeux 1989; Fodor and Pylyshyn 1988);
2. the ability to "memorize" rules which have already been tested on the outside world; and
3. the ability to evaluate new rules in a tacit manner by internal auto-evaluation which may be taken as a very simple form of "reasoning" (Dehaene and Changeux 1991).

Finally, these models and their simulation show how some elementary components of the network (e.g. allosteric receptors, synaptic triads) can introduce constraints into higher cognitive functions ("bottom-up" regulation). They also illustrate how a global process of interaction with the outside world, such as reward or reinforcement, can govern regulation at a more elementary level, such as the regulation of conformational transitions of allosteric receptors ("top-down" regulation). Last of all, they offer a specific illustration of the

interdependence between levels of organization which confer structural coherence and functional integration on the system.

References

Amit DJ (1989) Modeling brain functions. Cambridge University Press, Cambridge

Batuev AS, Orlov AA, Pirogov AA (1981) Short-term spatio-temporal memory and cortical unit reactions in the monkey. Acta Physiol Hung. 58: 207–216

Campbell DT (1974) Evolutionary epistomology.In: Schilpp PA (ed). The philosophy of Karl Popper. La Salle, Open Court

Changeux JP (1983) L'Homme neuronal. Fayard, Paris: English edition: Neuronal man, Pantheon New York

Changeux JP (1988) Molécule et mémoire. Bedou, Gordes

Changeux JP, Courrege P, Danchin A (1973) A theory of the epigenesis of neural networks by selective stabilization of synapses. Proc Natl Acad Sci USA 70: 2974–2978

Changeux JP, Dehaene S (1989) Neuronal models of cognitive functions. Cognition 33: 63–109

Changeux JP, Heidmann T (1987) Allosteric receptors and molecular models of learning. In: Edelman G, Gall WE, Cowan WM (eds) Synaptic Function. John Wiley, New York, pp 549–601

Damasio AR, Tranel D, Damasio H (1990) Individuals with sociopathic behavior caused by frontal damage fail to respond autonomically to social stimuli. Behav Brain Res 41: 81–94

Darwin C (1859) On the origin of species. London: Murray

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

Dehaene S, Changeux JP (1991) The Wisconsin card sorting test: theoretical analysis and simulation of a reasoning task in a model neuronal network. Cerebral Cortex, in press

Dehaene S, Changeux JP, Nadal JP (1987) Neural networks that learn temporal sequences by selection. Proc Natl Acad Sci USA 84: 2727–2731

Diamond A (1985) The development of the ability to use recall to guide action, as indicated by infant's performance on AB. Child Develop 56: 868–883

Diamond A (1988) Differences between adult and infant cognition: is the crucial variable presence or absence of language? In: Weiskrantz L (ed.) Thought without language. Clarendon Press, Oxford

Edelman GM (1978) Group selection and phasic reentrant signaling: a theory of higher brain function. In: Edelman GM, Mountcastle VB (eds) The mindful brain: cortical organization and the group-selective theory of high brain function. MIT Press, Cambridge, pp. 51–100

Fodor JA, Pylyshyn ZW (1988) Connectionism and cognitive architecture: A critical analysis. Cognition 28: 3–71

Fuster JM (1973) Unit activity in prefrontal cortex during delayed-response performance: Neuronal correlates of transient memory. J Neurophysiol 36: 61–78

Fuster JM (1984) Electrophysiology of the prefrontal cortex. Trends Neurosci 1: 408–414

Fuster JM (1989) The prefrontal cortex (2nd edition). Raven Press, New York

Fuxe K, Agnati L (1991) Volume transmission in the brain: new aspects for electrical and chemical communication. Raven, New York

Goldman-Rakic P (1987) Circuitry of the primate prefrontal cortex and the regulation of behavior by representational knowledge. In: Mountcastle V, Plum KF (eds) The nervous system: Higher functions of the brain, Vol. 5. Handbook of Physiology. American Physiological Society, Washington, DC:

Grant DA, Berg EA (1948) A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. J Exp Psych 38: 404–411

Heidmann T, Changeux JP (1982) Un modèle moléculaire de régulation d'efficacité d'un, synapse chimique au niveau postsynaptique. C R Acad Sci Paris, série 3, 295: 665–670

Jacob F (1970) La logique du vivant. Gallimard. Paris
Kubota K, Niki H (1971) Prefrontal cortical unit activity and delayed alternation performance in monkeys. J Neurophysiol 34: 337–347
Kubota K, Komatsu H (1985) Neuron activities of monkey prefrontal cortex during the learning of visual discrimination tasks with go-no go performances. Neurosci. Res. 3: 106–129
Marr D (1982) Vision. Freeman, San Fransisco
McCulloch WS, Pitts W (1943) A logical calculus of the ideas imminent in nervous activity. Bull Math Biophys 5: 115–137
Milner B (1963) Effects of brain lesions on card sorting. Arch Neurol 9: 90–100
Mountcastle VB (1978) An organizing principle for cerebral function: the unit module and the distributed system. In: Edelman GM, Mountcastle VB (eds) The mindful brain. The MIT Press, Cambridge
Nauta WJH (1971) The problem of the frontal lobe: a reinterpretation. J Psych Res 8: 167–187
Newell A (1982) The knowledge level. Artificial Intelligence 18: 87–127
Niki H (1974) Differential activity of prefrontal units during right and left delayed response trials. Brain Res 70: 346–349
Niki H (1975) Differential activity of prefrontal units during right and left delayed response trials. In: Kawai M, Ehara A and Kawamura S (eds) Symposia of the Fifth Congress of International Prematological Society, Kondos, Japan Science Press, Tokyo, pp 475–486
Piaget J (1954) The construction of reality in the child. Basic Books, New York
Poincare H (1913) Science et Méthode. Flammarion, Paris
Popper KR (1963) Conjectures and refutations. Basic Books, New York, 1963
Popper KR (1966) Of clouds and clocks: an approach to the problem of rationality and the freedom of man. Washington University Press. St. Louis, Missouri
Rosenkilde E, Bauer CE, Fuster JM (1981) Single cell activity in ventral prefrontal cortex of behaving monkeys. Brain Res 209: 375–394
Shallice T (1982) Specific impairments of planning. Phil Trans Royal Soc London B, 298: 199–209
Teuber HL (1964) The riddle of frontal lobe function in man. In: Warren JM, Akert K (eds) The frontal granular cortex and behavior. McGraw-Hill, New York, pp. 410–477
Teuber HL (1972) Unity and diversity of frontal lobe functions. Acta Neurobiol Exp (Warsz.) 32: 615–656
Watanabe M (1986a) Prefrontal unit activity during delayed conditional Go/No-Go discrimination in the monkey. II. Relation to Go and No-Go responses. Brain Res 382: 15–27
Watanabe M (1986b) Prefrontal unit activity during delayed conditional discriminations in the monkey. Brain Res. 225: 51–65