18

Formal Neuronal Models for Cognitive Functions Associated with the Prefrontal Cortex

J.-P. CHANGEUX[1] and S. DEHAENE[2]

[1]UA CNRS D1284 Neurobiologie Moléculaire,
Département des Biotechnologies, Institut Pasteur,
75724 Paris CEDEX 15, France
[2]Laboratoire de Sciences Cognitives et Psycholinguistique,
75270 Paris CEDEX 06, France

ABSTRACT

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 ffom any one of these interacting levels to the next are viewed in an evolutionary perspective. They are assumed to involve: (a) the production of multiple transient variations and (b) 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. Aselective 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 viewed as a process of selection among variable and transient "pre-representations." Models are presented that 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.

INTRODUCTION

In the introduction of his classical "Textbook of Psychology, Briefer Course" of 1908, William James stated that his aim was to deal with psychology as a "natural science."

Exploring Brain Functions: Models in Neuroscience
Edited by T.A. Poggio and D.A. Glaser © 1993 John Wiley & Sons Ltd.

In doing so, James introduced, into psychology and brain sciences, Claude Bernard's basic distinction between anatomy (stable morphological organizations or "structures") and physiology (the dynamic processes by which an organism acts on the outside world or on itself). The goal of psychology, as a "natural science," then became the establishment of a causal relationship between structure and function, the creation of "bridges" between neural and mental sciences. In past decades, a fruitful approach toward such a goal has been the proposal of experimentally testable neuronal models of defined cognitive functions (see review by Changeux and Dehaene 1989).

First, models of this type must not merely be "artificial" but "realistic" and plausible at the neurobiological level. Furthermore, to be adequate, it is also necessary that the structure–function relationship yields a *pertinent* correspondence between theoretical and experimentally observable variables. In our opinion, the choice of the *level of organization* at which such correspondence should be established plays a crucial role (Changeux and Dehaene 1989; Changeux and Connes 1989).

Our view (see Changeux and Dehaene 1989) is that within the brain, several such levels of organization might be distinguished. These levels, however, should not be confused with those that Marr (1982) distinguished in "vision," i.e., (a) the "hardware" or neural machine, (b) the representation and algorithm, and (c) the computational theory. Marr's levels are *levels of understanding* which perpetuate the cleavage between structure and function and actually take into account only a single level of functional organization.

Several models of hierarchical cleavages have been proposed to attempt to establish causal relationships between structure and function within the brain. For instance, one classically distinguishes (a) the level of elementary circuits and simple reflexes or fixed schemes of action; (b) "groups of neurons" and "symbolic representations" (Kant's "intendment" or "understanding"); and (c) complex assemblies of neuronal groups (Newell 1982; Dehaene and Changeux 1989) that we may refer to as "reason" (Kant) or "knowledge" (Newell) level. However, additional, finer hierarchical cleavages might be defined.

The transition from one level of organization to the next is considered within the general conceptual context, which has always been that of our laboratory (Changeux et al. 1973; Changeux and Danchin 1976; Changeux 1983; see also Edelman 1978, 1987) of an evolutionary epistemology (Darwin 1859; Poincaré 1913; Popper 1966; Campbell 1974). It is based upon:

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

The application of this paradigm to brain *internal* levels of organization does not postulate covalent variation of the genome but, in contrast, *epigenetic* variations of connectivity during development (time scale: years, minutes) and/or of states of activity of neuronal clusters at levels of either symbolic representation or "architectures of reason" (time scale: 0.1 second, minute). The proposed models of prefrontal cortex function (Dehaene and Changeux 1989; Dehaene and Changeux 1991) illustrate a plausible application of such an evolutionary scheme to the relationship between the "symbolic" and "reason" levels.

FUNCTIONAL ORGANIZATION OF THE PREFRONTAL CORTEX

The prefrontal cortex is the region of the neocortex in which the surface area has relatively increased the most during the course of mammalian evolution; from 3.5% in the cat to 17% in the chimpanzee and, finally, 29% in humans (see Fuster 1989). Its lesions in humans are accompanied by emotional and "cognitive" disorders, which are expressed both by error perseverations and abnormal tendency to distraction, with a general decrease in critical judgement. For Diamond (1988), the frontal cortex relates "information over space or time" and inhibits "predominant action tendencies." It constructs and updates "representations of the environment" (Teuber 1964, 1972; Goldman-Rakic 1987) and is involved in planning the interaction of the organism with the environment. For Shallice (1982), it constitutes a "supervisory attentive system," hierarchically higher than the "routine" or "contention scheduling" system. It ensures behavioral guidance in nonroutine situations and *selects* schemes appropriate to these situations. The prefrontal cortex produces "mental syntheses" (Bianchi) and is the site of "intentional behavior" (Pavlov). It is also required for adequate social conduct and related decision making and planning (Damasio and Damasio 1990). As the following models illustrate, it may be suggested that it is located *above* the "symbolic" or "intendment" level and contributes to the "neural architectures of reason" (see Changeux 1988).

FUNCTIONAL ANALYSIS OF THE PREFRONTAL CORTEX BY VARIOUS DELAYED-RESPONSE TASKS

Delayed-response (DR) tasks have been used with intact or lesioned laboratory animals and even with human adults and babies for the experimental analysis of prefrontal functions (Piaget 1954; Fuster 1984; Diamond 1988; review by 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 the scene, then a screen falls and covers the scene from the subject's sight for a variable duration. Then, two objects are presented simultaneously at two separate locations and the subject must choose one of them. The rule defining the correct choice varies with the type of task involved. In its strict sense, in the DR task and in the task $A\overline{B}$ (A, not B; Piaget 1954), the rule is to choose the object that stands at the *position* occupied by the cue object before the delay. In the DR task, the position of the cue is changed at random from one test to another, whereas in the $A\overline{B}$ task, its position is changed only after a criterion of success at that location has been reached. 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 alternate position in the following response. In all instances, the subject learns the task during a training phase in which he receives a reward for each successful test (fruit juice in the monkey, playing with a toy in the case of children, etc.).

All tasks are sensorimotor, tax short-term memory, and require selective attention. During the task, the subject makes a *decision* by comparing the test object with the stored representation of the cue. Finally, during training the subject performs an *induction* over time and space by discovering the abstract rule (pertinent choice of feature) that governs the reinforcement.

Human infants systematically succeed in the $A\overline{B}$ test around the age of 7.5 months and performance improves up to 12 months. The young macaque monkey masters these two tests between 1.5 and 4 months; ablation of the frontal cortex causes failure of the test. In the absence of the delay, an immature subject or lesioned adult passes the test; however, if the delay exceeds 1-2 seconds, performance deteriorates and essentially becomes random (review by Diamond 1988).

The DR task thus reveals early "high-level cognitive" functions which are linked to the integrity of the prefrontal cortex. To our knowledge, only a few, rare electrophysiological data exist on the *acquisition* of mastery of the DR task (Kubota and Komatsu 1985). The main data available are single-unit recordings in the monkey (macaque) *during the performance* of the DR test *after training* (review Watanabe 1986b; Fuster 1989). 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 focusing of attention on the cue, or a response to the *test* itself. Neurons of a second type, most often excitatory, change their activity in relation to the execution of the task (Kubota and Niki 1971; Watanabe 1986a). 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. Their activity is related to the state of *alertness* of the animal, and a correlation exists 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.

MODEL OF A FORMAL NEURAL NETWORK THAT COMPLETES DELAYED RESPONSE TASKS

The aim of this model (Dehaene and Changeux 1989) was to construct a minimum and biologically plausible neuronal network which would successfully pass the DR tasks. The model may lead to the identification of structural elements that are *necessary* to success in the tasks, and to the prediction of new properties subject to experimental validation.

The Formal Organism and its Environment

The formal neuronal network is contained in a "formal organism" that interacts with its 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, i.e., 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 trials, and each trial comprises four stages: presentation of the cue, delay, presentation of two objects, choice of object with reward or punishment, and an interval between two successive tests.

In trials of type 1 (analogues of DR or A$\overline{\text{B}}$), 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 the color of the cue (dimension 2).

The reward (or punishment) signal is applied from outside (either by a master who decides his score) 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 or experience). The reinforcement parameter covers an interval (–1, +1) where 0 is neutral, +l 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); (for discussion, see Amit 1989), linked together by synaptic contacts of either excitatory or 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 explicitly modeled.

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

The 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 modulator. The maximum efficacy value reached is itself variable and regulated by training (see below). Regulation of synaptic efficacy between A and B is undertaken in a heterosynaptic manner by C according to the "*synaptic triad*" scheme (Fig. 18.1; 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 compose a "modulated bundle."

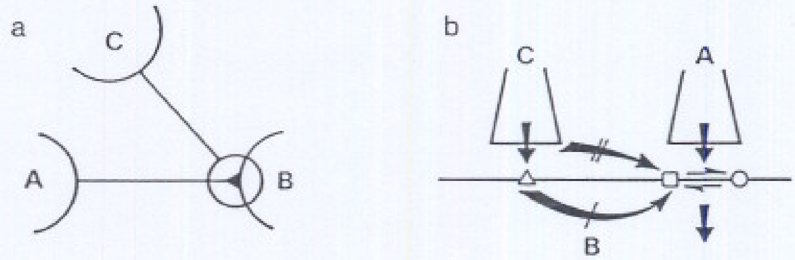

Figure 18.1 Synaptic triad. Signals from synapse C-B modulate the efficacy of the neighboring A-B synapse (from Dehaene et al. 1987).

Architecture of the Network

A major feature of the architecture of the network is the *distinction of two hierarchical levels of organization* (Fig. 18.2). Level 1 (the execution level) includes two layers of clusters: input and output. Each characteristic feature of a given object is decomposed 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 (the 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 topographically onto the output clusters and modulate the input–output connections. Each cluster of rule-coding neurons codes not for a particular feature of the object but

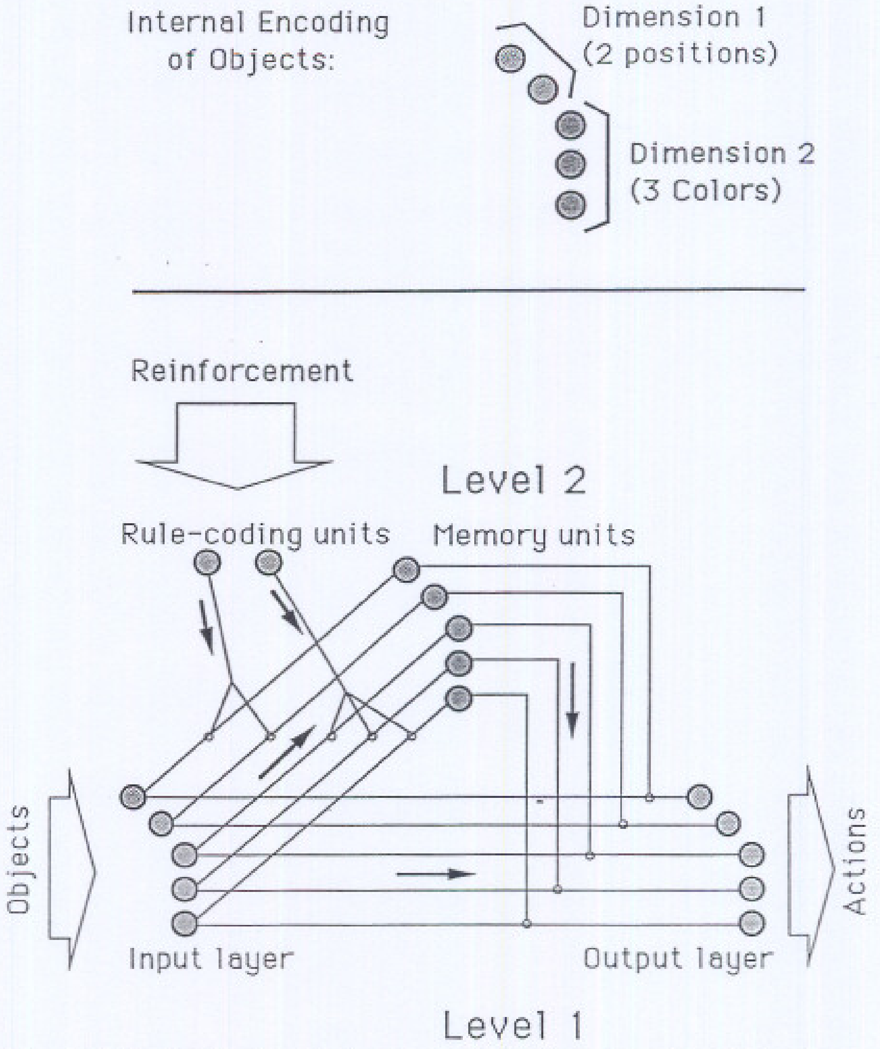

Figure 18.2 Model of the role of the frontal cortex in learning and execution of delayed response tasks (from Dehaene and Changeux 1989).

for *one dimension*, which groups together several features of the object. 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; level 2 would be identified with the prefrontal cortex.

Learning a Behavioral Rule

The organism learns a defined behavioral rule by interpreting a reinforcement signal that 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. 18.1) 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 may exist in at least two desensitized states for which the ionic channel is closed (review Changeux 1990). State I, of rapid access from the resting state, would be involved in the functioning of the synaptic triad. The fraction of receptors in state D, of slower access, would determine the maximum amplitude of variation in synaptic efficacy and would be stabilized by the time-coincidence of two signals: (a) a *postsynaptic* signal (e.g., the intracellular concentration of Ca^{2+}), which indicates recent activation of the cell, and (b) a diffuse *extracellular* signal (e.g., the catecholamines of divergent reticulo-frontal pathways), which is transmitted 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 rule-coding clusters, and spontaneous activity then varies from one cluster to another.

Learning takes place by *selection* of particular cluster of rule-coding neurons according to its actual state of activity. The layer of rule-coding clusters thus serves, in the framework of evolutionary epistemology, as a "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 such an organism is able to learn systematic orientation towards position A, for which it has been trained, when A and B were presented simultaneously. However, like infant humans or monkeys before maturation of pre-frontal functions, it fails in the DR and DMS tasks. By contrast, the formal organism which possesses levels 1 and 2 succeeds in all these tasks.

Rule-coding neurons play a critical role in the behavior since 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 that code for color are active, only the particular color of the cue will be memorized, 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 *dimension* (position, color) to which this feature belongs, are active.

The activity of the rule-coding neurons thus "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 (pre-representations) to find the particular one 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 *dimension* of the cue for which the "reward" is systematically positive.

The model allows simulation of the electrophysiological activity of defined neuron clusters during or after learning. In particular, neurons of the memory clusters display an activity which resembles that of neurons active during the delay period, the activity of which anticipates the behavior of the monkey during the choice when it is successful (but also when it fails) (Fig. 18.3).

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 (Diamond 1988). 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 $A\overline{B}$ or DMS. In addition, the organism is capable of passing from one task to another without difficulty (Fig. 18.4)

Despite these successes, the formal organism modeled in this way displays three groups of limitations: (a) in the sensorimotor tasks, the number of sensory dimensions or features and types of motor behavior postulated is very small; (b) the architecture is extremely simple: the number of formal neurons is six to seven orders of magnitude lower than that of neurons present in the prefrontal cortex of humans; (c) the range of available rules is very small in size.

THE WISCONSIN CARD SORTING TEST

This test used to detect prefrontal cortex lesions is more elaborate and complex than the delayed response tasks: it 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

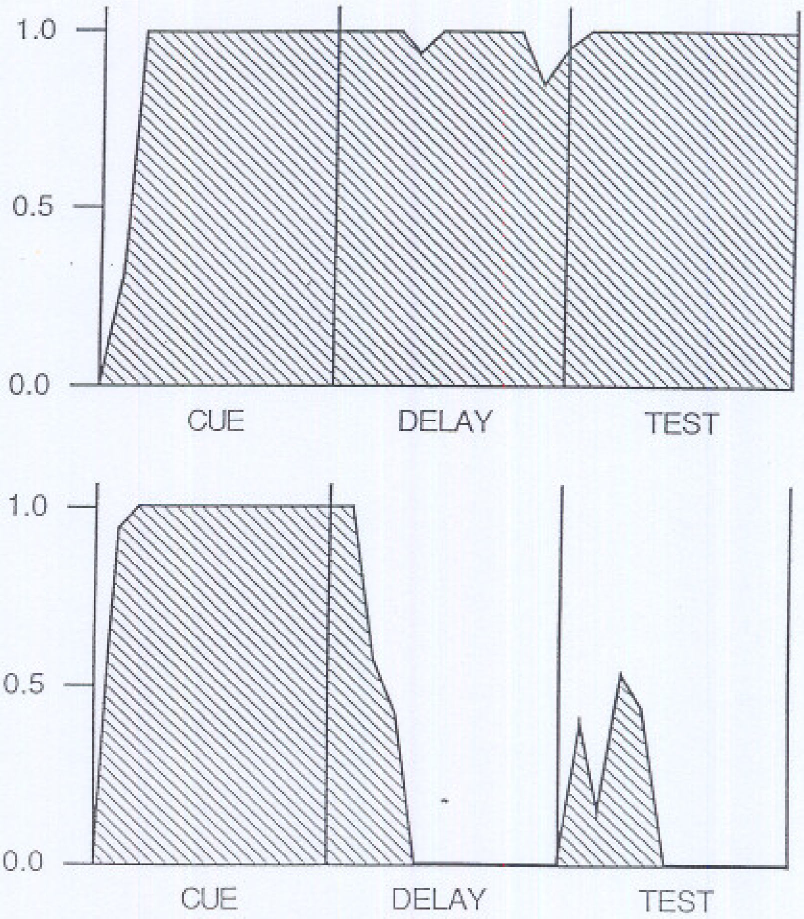

Figure 18.3 Simulation of the activity of a memory neurons cluster during the delay. Top: the group remains active during the delay; performance on this trial is correct. Bottom: the group is inactivated due to internal noise; the organism now fails on this trial (from Dehaene and Changeux 1989).

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 the response was "right" or "wrong." The subject tries to achieve the maximum of correct responses. The rule might be sorting according to, say, color. When performance is successful, the sorting rule is changed, for example, from color to shape. The subject must notice the change and discover the new rule (Fig. 18.5).

Many normal subjects fail to complete the test, particularly in the case of elderly subjects. However, subjects with a prefrontal lesion are systematically less successful

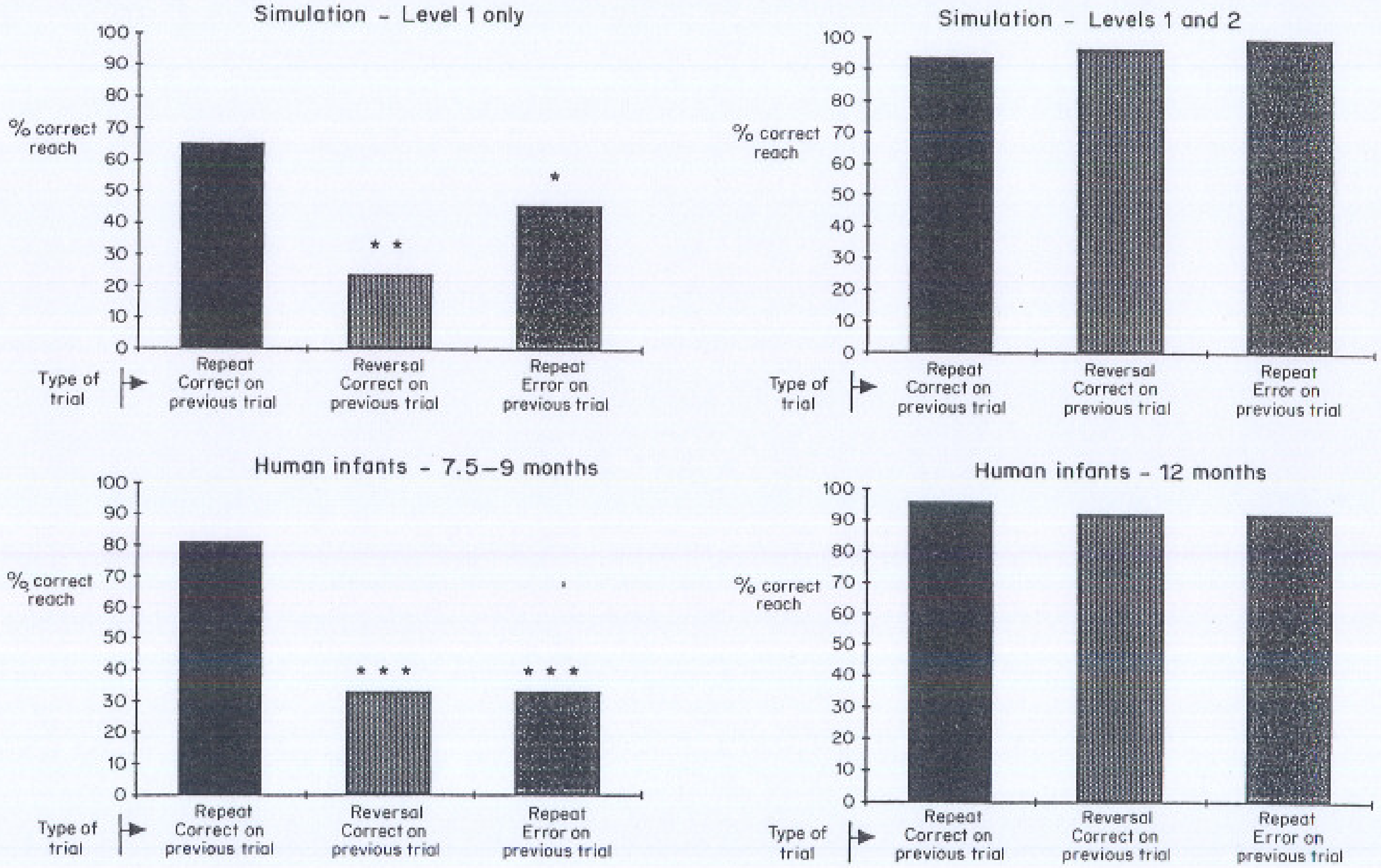

Figure 18.4 Comparison of the performance of the network with that of children tested by Diamond (1985). The network with level 1 only is comparable to children before the development of frontal connections. The network with levels 1 and 2 succeeds just as well as older children (from Dehaene and Changeux 1989).

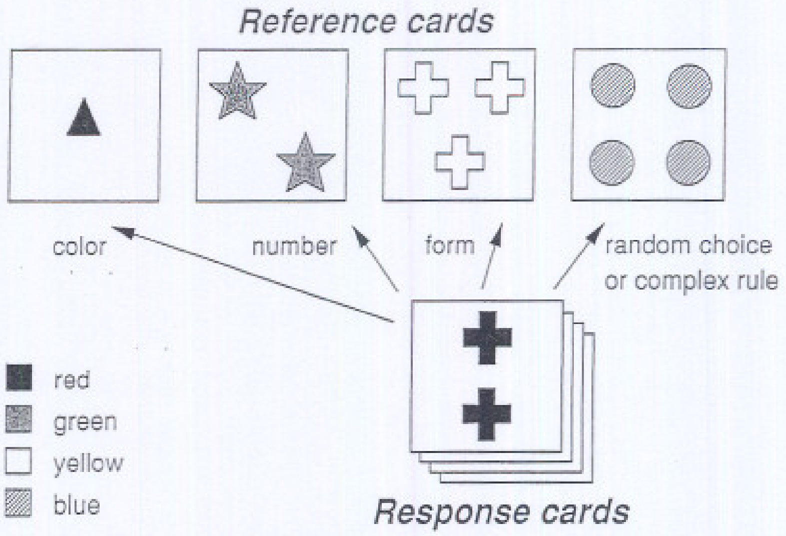

Figure 18.5 Cards used in the Wisconsin card sorting test (from Dehaene and Changeux 1991).

than normal subjects. Frontal subjects make errors of a particular type, called "perseveration," since they persist in using a rule that was previously correct, even after negative feedback was provided. They fail to shift from one sorting rule to another. Functional analysis of the abilities of a formal cognitive system to pass the test (Dehaene and Changeux 1991) leads to the distinction of six "formal machines," according to the manner in which they select a new rule (Fig. 18.6).

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 with replacement) draws a new rule entirely at random to replace the previous rule. The second, more complex (random + context) avoids drawing again a rule that has just been rejected. The third (random + memory) keeps an "episodic memory" of the previously rejected rules and draws only from among the remaining possible rules.

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," in addition to reasoning on negative trials and memorizing rejected rules, also reasons on positive trials. All the rules that would not have led to the same response are rejected as incorrect and memorized as such.

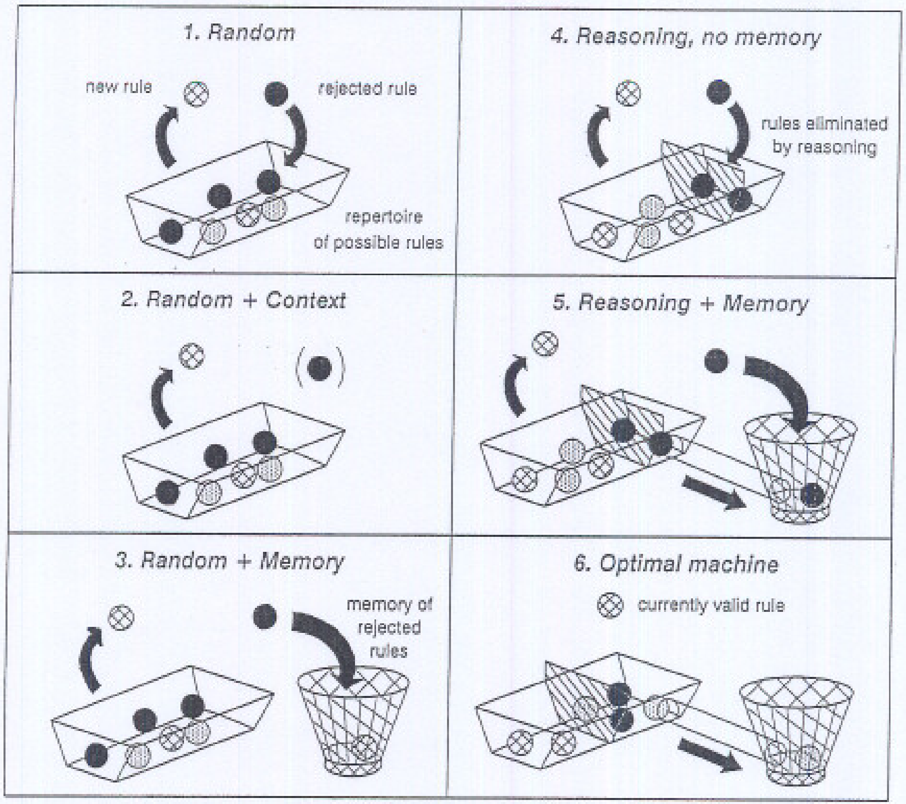

Figure 18.6 Diagrammatic representation of the operation of six machines of increasing complexity learning the rules of the Wisconsin test by selection (from Dehaene and Changeux 1991).

Comparison of the properties of these machines with the results of the Wisconsin card sorting test shows that the latter does not allow all these properties to be tested. Of the three fundamental cognitive abilities of these machines, namely (a) the ability to change the rule when punished, (b) memory of rules already tested, and (c) a priori rejection of rules by reasoning, only the first is tested in a critical manner.

A FORMAL NEURONAL ARCHITECTURE ABLE TO PASS THE WISCONSIN CARD SORTING TEST

This model (Dehaene and Changeux 1991) covers the major outlines of the architecture of the formal organism which passes the DR tasks (Dehaene and Changeux 1989; see also above section on A MODEL OF FORMAL NEURAL NETWORK

Passing the DR Test), but with several major additions and modifications (Fig. 18.7). The clusters of input neurons are more numerous since there are more dimensions and features of the cue involved in the test. Clusters of memory neurons are also present and receive projections from input clusters with conservation of topography. There is competition between input clusters, with reciprocal inhibition so that only one feature is memorized for each dimension.

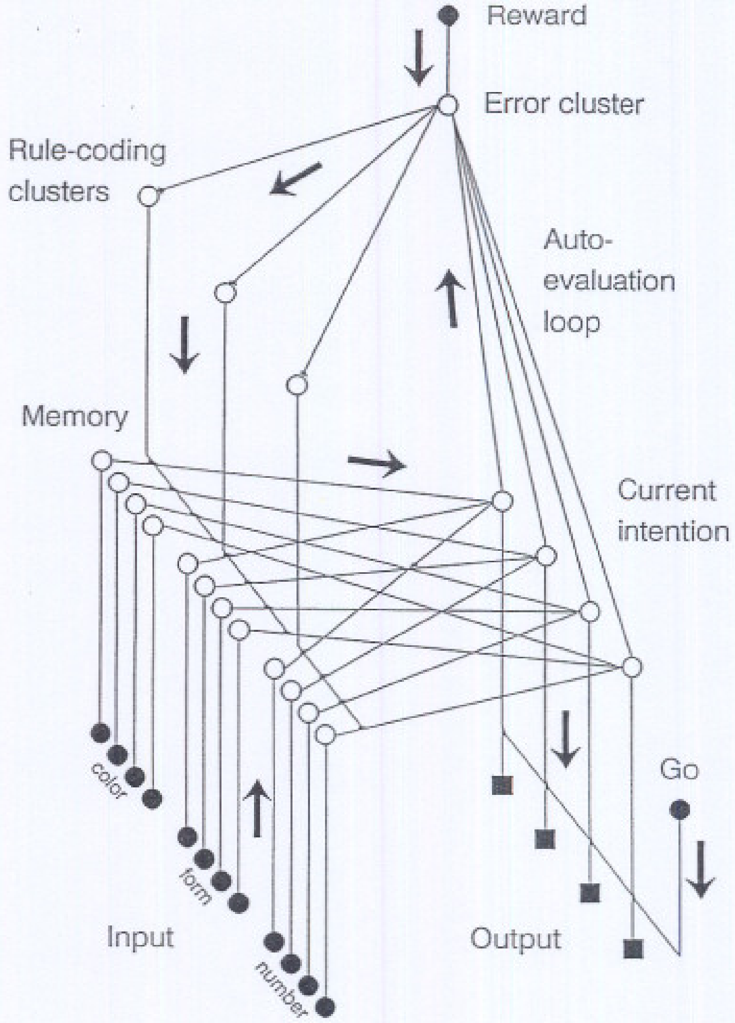

Figure 18.7 Model of the role of the frontal cortex in the Wisconsin card sorting test (from Dehaene and Changeux 1991).

The memory neuron clusters project onto a layer that was absent from the previous model. This new layer is composed of clusters which code for "*intentions*" of motor response that are distinct from the motor command itself. The model includes four 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 still play the key role in the performance of the test. Indeed, they gate the connections between memory and intention clusters according to a defined dimension (color, shape, number, etc.), and their activity varies with time during learning since each cluster that 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 the reward signal. In fact, the network has been modeled in 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 a "generator of diversity."

Another novelty of the model is the distinction of a "*cluster of error neurons*," which projects and modulates 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 the allosteric regulation of postsynaptic receptor desensitization of the type described previously. This depression is spontaneously reversible and the speed of recovery is, according to the model, a crucial parameter that 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 that 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 model is perhaps the "*auto-evaluation loop*," which short-circuits the reward input from the outside world. 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 labeled 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 to 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 percentage success rate for single trial 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. 18.6).

Lesion of the error cluster leads to slowing of learning and to an increase in perseverations similar to those observed in frontal patients (see above section on FUNCTIONAL ORGANIZATION OF THE PREFRONTAL CORTEX). The inertia of the generator of diversity becomes very large. As in the case of the simple network, lesion of rule-coding clusters interferes with the acquisition of a "systematic" rule of behavior. Lesions 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. It might, however, offer a formal explanation of the "sociopathic" behavior resulting from ventromedian lesions of the frontal cortex (Damasio et al. 1990). Damasio (1990) suggests that the deficit is due to the failures in the activation of somatic states linked to the punishment or reward which the subject has experienced, in association with specific social situations, and which must be reactivated as markers for the outcome of a response option. 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 modeling.

CONCLUSION

The two formal neuronal networks proposed to account for characteristic functional abilities of the prefrontal cortex give success in various DR 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. Despite their extreme simplicity, they provide several original and specific predictions susceptible to delineate novel experimental tests. One bears on the existence of "rule-coding neurons," the activity states of which vary randomly during the learning period until a rule of behavior is selected. Another concerns the mechanism, or mechanisms, of reinforcement by "error neurons."

At a more general level, the induction of rule by trial-and-error fits with the framework of evolutionary epistemology (Darwin 1859; Poincaré 1913; Popper 1966; Changeux et al. 1973; Campbell 1974; Edelman 1978, 1987; Changeux 1983; Changeux and Dehaene 1989; Dehaene and Changeux 1991). In this context, clusters of rule-coding neurons would constitute the "generator of diversity." Memorization by selection might then be viewed as homolog of the "amplification" mechanism postulated to follow selection, since the organism will reutilize the memorized trace repeatedly in its subsequent behaviors.

The models also illustrate the positive contribution of the distinction of hierarchical levels of network architecture to establish a pertinent relationship between structure features of the network and the following function: (a) the ability to generalize a rule acquired for a particular cue to a class of cues, or *systematicity* (Dehaene and Changeux 1989; Fodor and Pylyshyn 1988); (b) the ability to "memorize" rules which have already been tested on the outside world; and (c) the ability to evaluate new rules in a tacit manner by internal auto-evaluation, which may be taken as a simple and schematic 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 through "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 conformational transitions of allosteric receptors ("top-down" regulation). Last of all, they offer a specific illustration of the interdependence between levels of organization that confers structural coherence and functional integration on the system as a whole.

REFERENCES

Amit, D.J. 1989. Modeling Brain Functions. Cambridge: Cambridge Univ. Press.

Campbell, D.T. 1974. Evolutionary Epistemology. In: The Philosophy of Karl Popper, ed. P.A. Schilpp. La Salle: Open Court.

Changeux, J.-P. 1983. L'Homme Neuronal. Paris: Ed. Fayard. (Engl. ed.: Neuronal Man. New York: Pantheon.)

Changeux, J.-P. 1988. Molécule et mémoire. Gordes: Bedou.

Changeux, J.-P. 1990. Functional architecture and dynamics of the nicotinic acetylcholine receptor: An allosteric ligand-gated ion channel. In: Fidia Research Foundation Neuroscience Award Lectures, ed. J.-P. Changeux, R.R. Llinàs, D. Purves, and F.E. Bloom, vol. 4, pp. 21–168. New York: Raven.

Changeux, J.-P., and A. Connes. 1989. Matière à Pensée. Paris: Ed. Odile Jacob.

Changeux, J.-P., P. Courrège, and A. Danchin. 1973. A theory of the epigenesis of neural networks by selective stabilization of synapses. *Proc. Natl. Acad. Sci.* **70**:2954–2978.

Changeux, J.-P., and A. Danchin. 1976. The selective stabilization of developing synapses: A plausible mechanism for the specification of neuronal networks. *Nature* **264**:705–712.

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

Changeux, J.-P., and T. Heidmann. 1987. Allosteric receptors and molecular models of learning. In: Synaptic Function, ed. G. Edelman, W.E. Gall, and W.M. Cowan, pp. 549–601. New York: Wiley.

Damasio, A.R., D. Tranel, and H. Damasio. 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., and J.-P. Changeux. 1989. A simple model of prefrontal cortex function in delayed-response tasks. *J. Cog. Neurosci.* **1**:244–261.

Dehaene, S., and J.-P. Changeux. 1991. The Wisconsin card sorting test: Theoretical analysis and simulation of a reasoning task in a model neuronal network. *Cerebral Cortex* **1**:62–69.

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

Diamond, A. 1985. The development of the ability to use recall to guide action, as indicated by infant's performance on A$\bar{B}$. *Child Devel.* **56**:868–883.

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

Edelman, G.M. 1978. Group selection and phasic reentrant signaling: A theory of higher brain function. In: The Mindful Brain: Cortical Organization and the Group-selective Theory of High Brain Function, ed. G.M. Edelman and V.B. Mountcastle, pp. 51–100. Cambridge, MA: MIT Press.

Edelman, G. 1987. Neural Darwinism. New York: Basic.

Fodor, J.A., and Z.W. Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition* **28**:3–71.

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

Fuster, J.M. 1984. Electrophysiology of the prefrontal cortex. *TINS* **1**:408–414.

Fuster, J.M. 1989. The Prefrontal Cortex (2nd ed.). New York: Raven.

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

Goldman-Rakic, P. 1987. Circuitry of the primate prefrontal cortex and the regulation of behavior by representational knowledge. In: The Nervous System: Higher Functions of the Brain, ed. V. Mountcastle and K.F. Plum, pp. 373–417, vol. 5, Handbook of Physiology. Washington, DC: Am. Physiol. Soc.

Grant, D.A., and E.A. Berg. 1948. A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *J. Exp. Psychol.* **38**:404–411.

Heidmann, T., and J.-P. Changeux. 1982. Un modèle moléculaire de régulation d'efficacité d'une synapse chimique au niveau postsynaptique. *C.R. Acad. Sci. Paris* **925(3)**:665–670.

Kubota, K., and H. Niki. 1971. Prefrontal cortical unit activity and delayed alternation performance in monkeys. *J. Neurophysiol.* **34**:337–347.

Kubota, K., and 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. New York: W.H. Freeman.

McCulloch, W.S., and W. Pitts. 1943. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophysics* **5**:115–137.

Milner, B. 1963. Effects of brain lesions on card sorting. *Arch. Neurology* **9**: 90–100.

Mountcastle, V.B. 1978. An organizing principle for cerebral function: The unit module and the distributed system. In: The Mindful Brain, ed. G.M. Edelman and V.B. Mountcastle. Cambridge, MA: MIT Press.

Newell, A. 1982. The knowledge level. *Art. Intell.* **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: Symp. Fifth Congress Internatl. Prematological Soc., Kondos, ed. M. Kawai, A. Ehara, and S. Kawamura, pp. 475–486. Tokyo: Japan Science Press.

Piaget, J. 1954. The Construction of Reality in the Child. New York: Basic.

Poincaré, H. 1913. Science et Méthode. Paris: Flammarion.

Popper, K. 1966. Of Clouds and Clocks: An Approach to the Problem of Rationality and the Freedom of Man. St. Louis: Washington Univ. Press.

Shallice, T. 1982. Specific impairments of planning. *Phil. Trans. Roy. Soc. Lond. B* **298**:199–209.

Teuber, H.L. 1964. The riddle of frontal lobe function in man. In: The Frontal Granular Cortex and Behavior, ed. J.M. Warren and K. Akert. New York: McGraw-Hill.

Teuber, H.L. 1972. Unity and diversity of frontal lobe functions. *Acta Neurobiol. Exper. (Warsz.)* **32**:615–656.

Watanabe, M. 1986a. Prefrontal unit activity during delayed conditional discrimination in the monkey. *Brain Res.* **225**:51–65.

Watanabe, M. 1986b. 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.