20

Group Report: Architectures of Intelligent Systems

D. KIRSH, Rapporteur

J.S. ALTMAN, J.-P. CHANGEUX,
A.R. DAMASIO, R. DURBIN, A.K. ENGEL,
W.D. HILLIS, D. PREMACK, R. RIVEST,
P.E. ROLAND, P.S. ROSENBLOOM,
G.S. STENT, P. STOERIG

INTRODUCTION

Theories of intelligence can be of use to neuroscientists if they:

1. provide illuminating suggestions about the functional architecture of neural systems;
2. suggest specific models of processing that neural circuits might implement.

The objective of our session was to stand back and consider the prospects for this interdisciplinary exchange.

One of the facts that emerged early in our discussions was that given our current level of knowledge, it is hard to tie theories and models of intelligence to actual neural machinery. Skeptics see this bridging problem as nonaccidental. We know, for instance, that human intelligence depends on a range of knowledge representation capacities, reasoning methods and additional computational mechanisms that have both principled and unprincipled components. Theorists in artificial intelligence (AI), cognitive science, linguistics and psychology are beginning to tell us something about the principled components—the core competences of human intelligence. But there is widespread disagreement about how much of cognition is principled. Most of human intelligence has evolved through extension and repair of simpler cognitive systems. This suggests that biological designs may not be as cleanly principled as engineers would like. If intelligence is really the product of prolonged tinkering, high-level accounts of cognitive designs may be misleading. Each intelligent system might have

Exploring Brain Functions: Models in Neuroscience
Edited by T.A. Poggio and D.A. Glaser

its own idiosyncratic design, depending on its evolutionary history, and so theories of intelligence could be of use to neuroscience in one way alone: as *specifications* of behavioral capacities intelligent systems display. They could offer few concrete design constraints neuroscientists might find useful.

In our discussions we did not take this skeptical stance for two reasons. First, it is hard to imagine how any biologically designed intelligence, no matter how idiosyncratic, can fail to be organized around certain high-level principles of functional organization characteristic of rational systems. We may be wrong in some of the details of our account of these principles; however, it is unlikely that we could be wrong about the key competences underpinning rational intelligence. This is particularly so if we are motivated not only by abstract considerations of what is necessary for intelligence, but by studies of human subjects with brain damage and studies of comparative intelligence.

The second reason we rejected the skeptic's position is that more complex models of dynamical systems allow us to relax, somewhat, our idea of what a principled design is. PDP systems, cellular automata and other highly connected systems tend to implement functions in complex ways. These approaches have yet to yield theories about the global architecture of intelligence, but they are highly suggestive of some of the forms and properties of neural representations and circuitry.

Our report is organized into six sections, each addressing a particular issue. What are the key competences underpinning intelligence? How are we to understand the various levels of functional organization characteristic of brains? What are some of the forms and properties of neural representations? What makes conceptual knowledge special? What is the role of expectation in intelligent systems? And finally, because no free ranging discussion of intelligence would be complete without a reconsideration of consciousness and its functional role, we reviewed a few of the facts and difficulties associated with consciousness.

COMPETENCES UNDERPINNING INTELLIGENCE

It would be standard in philosophical circles to begin an inquiry into intelligence with an analysis of what we ordinarily mean by intelligence. The outcome of such an analysis might be a list of basic competences, such as the ability to reason and problem solve in propositional, spatial and other analogue domains; the ability to learn by induction, by inference to the best explanation; the ability to be self-aware; and the ability to explain the conduct of others, to name just a few. The chief element missing from such an analysis is a structural framework that might explain how basic competences *interact* to produce intelligent behavior. Our first step, accordingly, was to motivate such a list by reviewing some of the abstract properties of intelligent systems with an eye to uncovering what would be needed to build a computational system with human-like properties.

Intelligence in the Abstract

Following the analysis offered by Rosenbloom and Newell, we began by assuming that intelligence is a measure of a system's *rationality*: the more rational a system, the more intelligent it is. An agent is perfectly rational if it is able to bring all of its knowledge to bear in the service of its goals; it always chooses the optimal action, given its knowledge, goals, and action capabilities.

Intelligence, on this account, is to be sharply distinguished both from knowledge and behavioral competence per se. In any given case, inadequate performance in a task may be due either to inadequate knowledge or inadequate intelligence. If a creature has the requisite knowledge to choose the rational act relative to its goals, we will say that its failure to do so is evidence that it is not perfectly intelligent; it lacks certain capacities necessary for making the most of its knowledge. If, by contrast, it lacks certain bits of essential knowledge, no amount of intelligence can substitute for this knowledge. We will say that its inadequate performance is caused by ignorance, not stupidity.

One feature of this approach is that intelligence can now be understood as an attribute of design rather than just a set of behavioral competences. If a system is sufficiently well designed so that it is well adapted to its environment of activity, it is by definition highly intelligent. At first this may seem counterintuitive for it seems to imply that cognitively primitive creatures, such as aplysias or nematodes, no less than cats and humans, can be assessed for intelligence. But two facts qualify this. First, because intelligence has to do with using knowledge effectively, only systems that can be characterized as having knowledge can be assessed for degree of intelligence. Second, because systems surviving in complex environments require more sophisticated goals, knowledge, and action capabilities than systems surviving in simpler environments, the information-processing capacities which more complex creatures require for choosing rational actions will themselves have to rise in complexity. It is often this requirement of competence in complex domains that we associate with intellligence.

What then can be said from this perspective about the design of intelligent systems adapted to complex environments? Since the environments of activity in which humans operate range from simple physical worlds where obstacle avoidance, locomotion, and object manipulation are the basic requirements of life to complex social worlds where body language, conversational skill, and knowing when to be silent are critical acts, we know that the type of system that could cope with all of these requirements will minimally have to have the competence (a) to perceive its situation, (b) to represent situations in working memory, (c) to select among its goals, (d) to combine information of that goal and situation with potentially relevant background knowledge stored in long-term memory, (e) to perform appropriate computations to choose a rational or nearly rational action, and (f) to execute that action in the world. Note that all these are domain-independent competences.

Moreover, since in complex environments it is unlikely that a system has enough knowledge to attain all of its goals immediately, we can expect at least two additional

features: a reflective component whose job is to schedule goals, to plan, to determine relevant implications of existing knowledge, to problem solve; and a learning component that will update existing knowledge in the light of the activities of the reflective component and perceived consequences of actions on the world (see Fig. 20.1).

This list of competences and components is, to be sure, abstract: it abstracts from particular domains, and leaves unspecified all details of representation and algorithm. Yet any functional decomposition of a system based on an analysis of the requirements of rationality will be similarly abstract. It is only when we add facts about the processing time and memory size a system has available, and about the particular task environments the system must succeed in that we can say much more about the

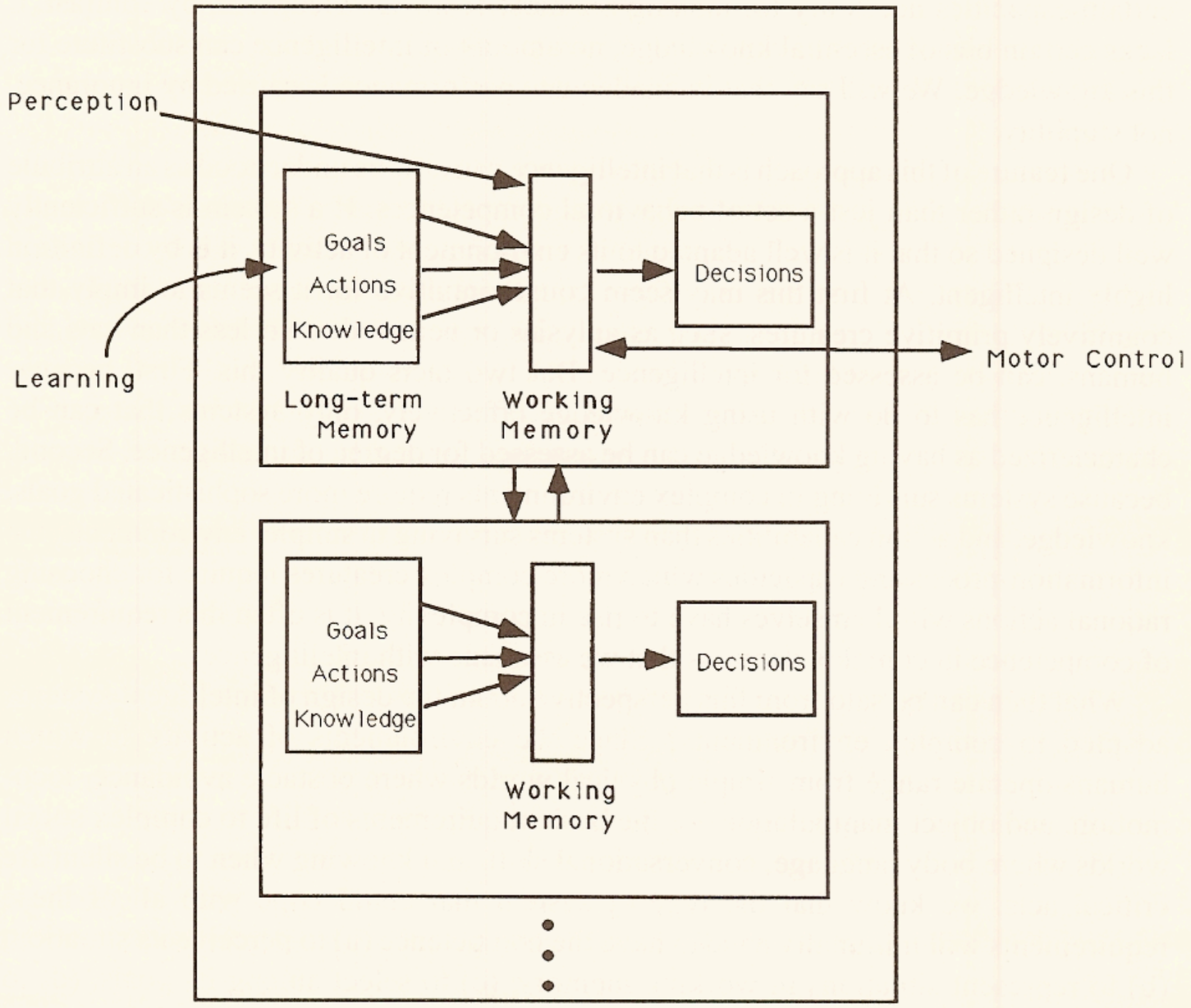

Figure 20.1 The key components of an intelligent system, on this view, are a set of actions the system can perform, goals it seeks to attain, a body of facts, laws, principles rules, etc. that it knows and can use, as well as a mechanism for selectively retrieving goal-relevant knowledge and actions, and a mechanism for rationally deciding a course of action given that retrieved material. Implicated are also a reflection component for overcoming impasses at the original level and a learning component to ensure continuous adaptation to a dynamic world.

functional structure of these general mechanisms. One such further decomposition has formed the basis for the Soar architecture (Newell 1991; Rosenbloom et al. 1991). In Soar, each of these components is given a specific implementation. The result is a computational system that has been shown to provide a good match to a range of human psychological data.

Having blocked out some of the abstract requirements for intelligent systems we next considered evidence from the study of lesions to human brains to gather further evidence about the key competences of intelligent agents.

Lessons from Lesions

According to Damasio, the study of brain damage gives broad confirmation for a functional decomposition akin to Rosenbloom's. But from a clinical perspective planning, attention, and learning (especially learning associated with punishment and reward) are of particular interest for human intelligence. See Fig. 20.2 for a diagram of some of the key dependency relations between functional systems that are motivated by clinical case studies.

Although lesions in varied sensory cortices can reduce overall intelligence because of the ensuing deficits in perception, memory, or language, it is damage in the frontal cortices, especially in the prefrontal sectors, that most consistently produce the defects in intelligence, in the sense most regularly used in the clinic. Curiously, some of those defects may not be readily measured by psychological test scores; many frontal lobe patients maintain IQs in the normal range and may have largely intact basic neuropsychological performances (e.g., in perception, conventional memory, and language).

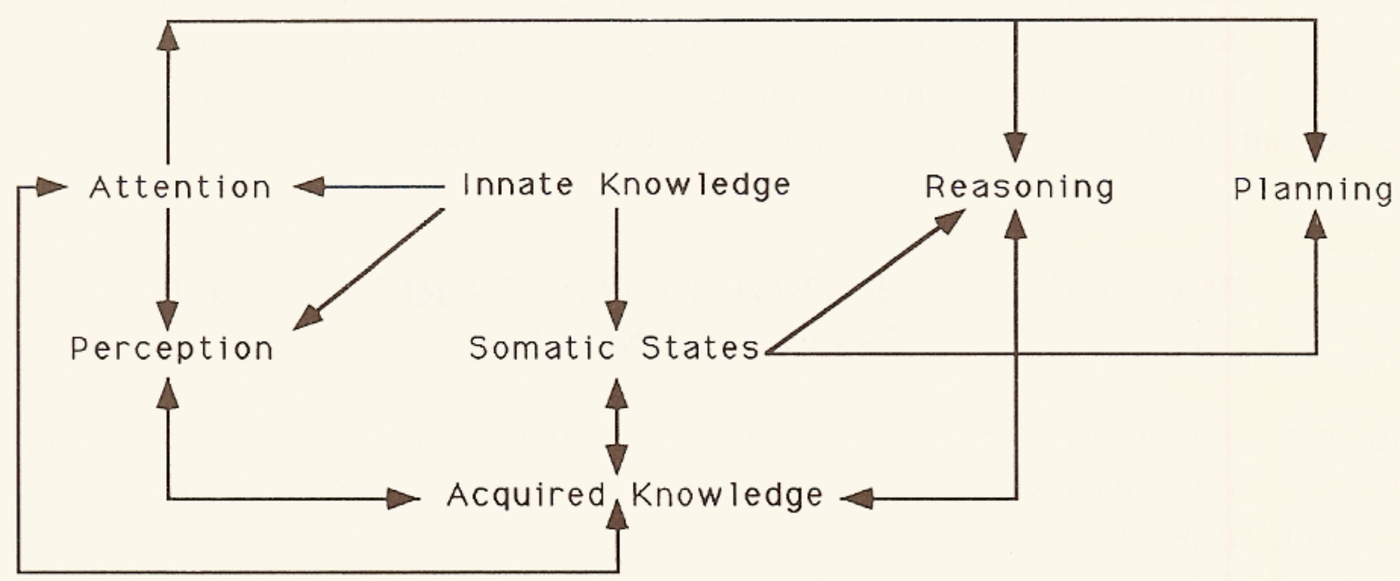

Figure 20.2 In Damasio's view, human intelligence can break down if lesions are made to any of several systems in cerebral cortices. The magnitude of the defect varies with the system affected. When lesions compromise systems which include prefrontal cortices, there tend to be major defects in reasoning and planning. Yet damage to other systems, by compromising attention, or active perception, or the retrieval of knowledge, can also diminish the scope and level of intelligent behavior.

However, that intelligence is defective is apparent in real-life situations. Patients fail to plan their activities properly, often in the immediate- and medium-term ranges, and may be entirely disrupted in their ability to plan for the long-term future. Furthermore, where their social behavior was intact before the claimed onset, there can develop marked defects in social conduct.

A class of frontal lobe patients that deserves special mention is made up of patients with ventromedial frontal lesions (involving largely the orbital prefrontal cortices); they develop profound disturbances of planning and social conduct in the face of otherwise intact intelligence. Patient EVR is an important example. Up to his mid-thirties, when he underwent a surgical resection of ventromedial frontal cortex to remove a meningeoma, he was a successful professional, husband, and parent, respected and loved by his friends and relatives. After the lesion, he was no longer able to decide intelligently how to run his daily life—in professional or personal terms—and he was no longer able to plan for the future. He could not maintain his job; he made disastrous financial decisions; he was not able to maintain previously stable and advantageous relationships; and he initiated personal relationships that led to personal tragedy. He has not been able to learn from any of his mistakes, in spite of the fact that they have been explained to him in detail and that he has clearly understood the implications of his actions. Yet EVR is not only psychometrically intelligent, he is personally pleasant.

In EVR and in four other comparable individuals, recent research has identified a defect in the ability to generate and experience somatic states relative to emotionally charged stimuli. In other words, EVR and similar patients, seem to be unable to "feel" much in situations in which previously they apparently had "normal" feelings, and in which presumably normal individuals also have normal "feelings." These findings have led to the hypothesis that a failure to reenact negative (or positive) somatic states, clearly associated to punishment and reward, may deprive these individuals of an important signal marking potentially dangerous (or advantageous) future outcomes as they are internally represented in decision-making scenarios.

WHAT DOMAIN OF KNOWLEDGE IS CRITICAL FOR INTELLIGENCE?

Our account of competences has so far been at the functional level; nothing concrete has been said about the kind of knowledge essential for human intelligence as we know it. On the basis of comparative studies with animals and studies in developmental psychology, Premack approached the question of what knowledge is essential for minimal performance in human environments. Based on his view, not only must an agent have certain skills and mechanisms before it can be intelligent in a human way, it also must have certain ontological categories and predispositions for conceptualizing. Chief among the critical elements of knowledge are the capacities to divide the world into categories, such as physical object, mind, biological kind, and number.

Possession of these concepts is significant for two reasons. First, it is probable that only systems with certain higher functional capacities are capable of possessing such categories, and second, it is probable that possession of such categories is a precondition for acquiring a host of additional capacities. We will return to this notion of higher functional capacity shortly.

It is easy to appreciate how the concept of a physical object is central to a range of cognitive capacities we rely on to make sense of our world. Without this concept, systems may have the ability to respond in similar ways to similar stimuli if the notion of "similar" is simple enough; however, a more sensitive response often requires an ability to identify, individuate, and reidentify entities.

Evidence suggests that this capacity to identify is innate or nearly innate. Four-month-old infants tested with habituation/dishabituation, for instance, show "knowledge" of what an object is; they rely on von Uexküll's criterion: an object is an item whose parts move in unison (Spelke 1985). Indeed, infants at this age rely exclusively on this criterion, making no use of the object quality information that adults and older children use. They also "understand" some basic facts about the interaction between objects, e.g., that two solid objects cannot occupy the same place at the same time. In this same period, however, they display no knowledge of either gravity or inertia.

The concept of mind and mental causation is equally important for understanding our world. Teleological, intentional, and purposive explanations are different in logical structure than simple law-like explanations. This fact is significant because of the obvious link between explanation and understanding. The more we can explain the more we can predict and respond adaptively.

There is some data based on three-year-old children that suggest that humans are endowed early on with the capacity to distinguish intentional from nonintentional actions and systems (Dasser et al. 1989). For example, children appear to distinguish objects that seem to be self-propelled, i.e., objects that start and stop their own motion from objects that move only when acted upon. The self-propelled objects they seem to interpret as intentional, whereas non self-propelled objects are not seen as intentional. Moreover, infants further distinguish the valence of intentional actions. Gentle rubbing is seen as positive, hitting as negative. When an infant perceives one object act upon another in a valenced manner, it attributes social intention to the one object and expects the action to be reciprocated, and for the reciprocation to preserve valence (Premack 1990).

Related to the capacity to distinguish intentional from nonintentional actions, but obviously more sophisticated, is the capacity to attribute mental states to oneself and others. Often called possessing a "theory of mind" (TOM), this capacity is acquired by all children without formal pedagogy (Premack and Woodruff 1978; Premack 1988a). Children attain an adult level of competence by the fourth year (Wimmer and Perner 1983). This is a vital acquisition, for arguably TOM can be seen as a precondition for pedagogy, the teaching of one organism by another, which is not seen in other species (Premack 1984; Premack 1991).

LEVELS OF FUNCTIONAL ORGANIZATION IN BRAIN AND COMPUTER

If possession of certain concepts is a precondition for some of our higher capacities, it is vital that we understand what the expression "higher functional capacities" means. The expression is naturally allied to the idea of higher levels of function, or functional organization. This idea, it seems, has several different senses. We distinguished three:

1. level of a function as its place in an epistemological hierarchy—processes nearer the sensory periphery are at lower levels, processes nearer the reasoning centers are at higher levels;
2. level of a function as its place in a control hierarchy—processes that can turn on or shut off lower-level processes are higher in a control sense; and
3. level of a function as its level in a decompositional hierarchy—processes analyzed at higher levels are implemented in processes at lower levels .

Level as Epistemological Hierarchy

For Changeux, who introduced the discussions, the aim of the life sciences is not only to specify function but, most of all, to relate a given function to an appropriate anatomical organization. In this search for anatomical implementation of functions, we need to map out an order of functioning. Changeux suggested that the logic of processing information from sensory perception to understanding ("Verstand" or "entendement"), up to reasoning ("Vernunft" or "raison") and planning might be related to gross neuroanatomical compartments of the brain. These would, for instance, include primary and secondary cortical areas, parietotemporal cortex with other association areas, and prefrontal cortex.

The functional level of a process viewed from this epistemological perspective is quite distinct from the level at which a process may be *analyzed*. Changeux emphasized that the Marr/Poggio computational/algorithmic/hardware distinction does not mark a difference in level of organization, but of analytic understanding. In principle, any process may be studied at different levels of analysis, the difference being the kind of questions we answer about the process. For instance, at the computational level we analyze a process in terms of the informational problem it is meant to solve. We can ask about the well-posedness of the problem and its absolute complexity. At the algorithmic level we can ask questions about the costs and benefits of particular representations and particular algorithms. And at the hardware level we can ask about the suitability of particular neural mechanisms to instantiate particular algorithms and representations. In most cases, each of these questions is worth asking but they are orthogonal to our concern with localization at a compartmental level.

Level as Control Hierarchy

Cognitive processes that are deeper or higher in an epistemological sense are often higher in a control sense too. Let us say that an event or process is higher up the control hierarchy if it can initiate, terminate, or interfere with events or processes at lower levels. Premack offered a thought-provoking experiment to show that human intelligence is organized hierarchically in a control sense.

Assume we have conditioned both a human and a dog to flex their legs to a conditioned stimuli associated with a shock to the foot. The next day we run an extinction session. Before our subjects arrive, however, there is a power failure in the village so that dog and human find a lab lit by candles. The difference in behavior of the two species is dramatic. Dogs produce a normal extinction curve; people do not. They do not flex their leg even to the first presentation of the conditioned stimuli.

The reason, of course, is that in people, knowledge that the shock is caused by electricity coupled with knowledge that there is a power failure overrides learning. Learning is not eliminated; when power is restored, people will generate a perfectly normal extinction curve. The organization of human intelligence, however, grants knowledge the power to override the performance called for by learning. Whether the same process operates in nonhumans is an open question. At present, we have little evidence of this type of knowledge operating in nonhumans. Clearly more must be said. We need a proper account of the difference between informational states acquired through nonconditioned learning (e.g., through reasoning) and informational states acquired through conditioning or association. Presumably, the two are encoded in different ways. Information based on conditioning may be represented by low-level representations (e.g., by sensory images) whereas information based on causal reasoning cannot be so represented; it requires more highly processed mental events, sensory images that are coded for type, token and beyond.

A second example of multilevel organization is the disparity we sometimes find between habituation/dishabituation and explicit choice. Infants show recognition for a distinction of which they can make no instrumental use. For example, 18-month-old infants, both humans and chimpanzees, show differential habituation/dishabituation to sameness and difference. Following experience with AA (two like objects) they respond less to BB than CD; likewise, following experience with EF (two different objects) they respond less to CD, than BB. But they cannot *match* sameness and difference, i.e., match AA to BB, or CD to EF. Disparities of this kind speak to the multilevel organization of intelligence, showing that processing of events which is successful on one level is unsuccessful at a higher level (Premack 1988b; Wellman 1991; Premack 1976).

Level as Decompositional Hierarchy

The third type of hierarchy we need to distinguish is, in certain respects, the most well understood. In computers we know that there are well-defined functional levels, such

as devices, circuits, register level descriptions, symbolic or program-level descriptions, agorithmic descriptions, etc. Each level is well defined in terms of components and operations, and each level can be used to implement the next. A level provides a clean abstraction in the sense that a function on one level can generally be described without reference to the levels below.

The question whether analogous decompositional levels exist within the brain is open for several reasons. First, computers have levels because they have been engineered to do so. Evolution may have adapted this engineering strategy, but then again it need not have. In other domains, nature seems disposed to decompositional hierarchies. Complex organizations are easier to create (Simon 1962) if they are constructed in Chinese box fashion, and decompositional levels confer a certain robustness on processors. Yet, as Hillis argued, although decompositional levels are likely within the brain, they are not nearly as distinct and well-defined as in the computer, and they stand to each other in more complex relations than simple hierarchies.

A second reason pointed out by Kirsh is that in computers, levels are assumed to be insulated against processes occuring at lower levels because the speed of processes occurring at different levels is sufficiently far apart that one can be confident that lower-level processes will settle fast enough to serve as building blocks for higher-level processes. Level-specific regularities in computers are reliable as long as there is no breakdown and a lower level shows through. Programs are excellent descriptions of system behavior unless registers fail to behave as assumed. Hence, in normal conditions, each level can be treated as modular and can be assumed to behave according to its own laws. However, in biophysical substrates, features of lower levels may show through regularly because of the importance of timing effects, connectivity, and local inhomogeneities in detailed aspects of the biophysics of neurons. This means that information can leak through levels, making it difficult to construct clean models of the information processes occuring at a given level. (See Koch et al, chapter 6, for a more complete account of this problem.)

A further, more specific question that provoked discussion was whether there exists a level of the brain that corresponds to the symbol manipulation level in computers. It is a central thesis of mainstream AI that there exists such a level—a level where compositional representations can be combined, uncombined, and manipulated. There was partial consensus on this point; however, those who disagreed rejected the idea completely.

This led to consideration of a different level at which models of brains and computers might connect: the level of idealized "neuron," the level of connection models, a subset of which are familiar connectionist models. In the ensuing disagreement, a useful consensus was reached in that whatever model—symbolic or connection —proves more useful, it will largely be irrelevant to the type of regularities to be found at higher computational levels.

This is not to say that insight into higher levels may not be obtained by studying lower levels. Virtually everyone present assumed that biological models of brain

performance set important constraints on the class of computational models empirically minded computationalists wish to propose. But, as Hillis pointed out, before computationalists can exploit these constraints, they must be told the different types of computational elements experimentalists believe there are; how many of each type there are; the kinds of patterns they communicate in; and the time scale of their communication. These questions are more important for the computationalist to know than details of function. For example, it is more useful for building computational models to know how many neurons there are and how fast they are than to understand their specific responses to stimuli.

Despite periodic moments of agreement in our discussion, true consensus was seldom reached.

A Cautionary Note (C. Koch, R. Douglas, P. Roland)

The computer is not a model of the brain, and explanations that assume that it is are actively dangerous and pernicious for the progress of neuroscience. A model is a simplified description of the system under study. As such, the model must reflect the cardinal properties of the architecture and functions of the system that it models. Because neurons are organized in ways that are fundamentally different to general purpose computing machines, similes and explanations of behavior that arise out of computer science are essentially inadequate. Concepts such as "programming," "input buffers," "pipe-lining," and the like have no correlates in biological neuroscience. Because the ontology of computer models is not committed to brain properties, predictions and proposed experiments will not be matched to their biological counterparts. This results in a misleading coupling between theory and experiment.

NEURAL REPRESENTATION

Whether one accepts the standard computer model, a connection model, or some even less idealized account of neural processing, it is accepted as dogma today that the central function of neural tissue is to process and represent information. Some of this information can be implicit in the structural organization of the cortex, but much of it must be explicitly represented in patterns of neural activity, which dynamically emerge as functional states in the cortical network. Intelligent behavior of an organism will certainly be constrained by its ability to create representations of objects and events in a flexible way. This appears as a prerequisite for the acquisition of memory, for the anticipation of future events in the outside world, and for the planning and execution of purposeful behavior. Because of this fundamental importance, it is worth clarifying what some of the means are for representing information in patterns of neural activity.

Models of Neural Representation

Engel pointed out that the formation of cortical representations can essentially be viewed as a problem of *binding*, which arises from the highly distributed nature of cortical information processing. The binding problem can be exemplified by considering the visual system. Due to the finite extension of their receptive fields, most neurons in visual cortical areas integrate information only from a limited part of the visual field. In addition, all these neurons respond only to a limited range of feature constellations. Thus, for instance, certain cells are sensitive to the movement direction of a figure, but not to its particular shape, orientation, or color. These features, in turn, are registered by other neurons. Representing a visual object therefore requires the binding of information between different parts of the visual field. In addition, binding has to occur across feature domains, that is, the shape of the object has to be linked with its color, movement trajectory, texture, and all other possible attributes. This need to integrate distributed information into representational entities is not confined to the visual modality, but constitutes a general problem which applies to all cortical systems.

A number of models have been proposed which approach the binding problem in quite different ways. These models are schematically reviewed in Table 20.1. A well-known classical proposal that has pervaded cortical neurobiology for decades was made explicit by Barlow. He introduced the notion of "cardinal cells" which extract (in a serial-hierarchical manner) information from lower-level neurons and thus acquire sufficient specificity to represent complex objects or events. Thus, in the visual cortex, for instance, individual dedicated cells were assumed to represent a whole object such as a grandmother's face. The difficulties with single-cell representations have now been widely recognized. Severe problems arise first from the enormous number of cells which would be needed to represent all possible feature constellations (combinatorial explosion); second, representations which rely on a very small number of cells are highly vulnerable; and third, since in this model binding is expressed by convergence of anatomical connections, it is not flexible enough to account for the fact that humans clearly can create novel representations within 100–200 ms, which occurs, for instance, during tachistoscopic presentation of novel visual scenes.

Thus, an alternative proposal made by Hebb seems to be more advantageous. Hebb's fundamental idea was that assemblies of cells rather than single units should be the correlate of representations. Such assemblies are formed by cooperative interactions between large numbers of neurons distributed in the cortex. In the Hebb model, an assembly is defined by the fact that the participating cells concurrently elevate their average firing rates. Thus, binding is expressed by response amplitudes. Clearly, assembly representations provide higher flexibility and reliability than single-cell representations. Defining assemblies by response amplitudes, however, still exhibits a major drawback, that has be addressed as the "coexistence problem." Facing a natural environment, the organism has to represent multiple objects or events at any one time. Coexistence of several representations, however, constitutes a major problem

Table 20.1 Comparison of different models for the formation of cortical representations (Engel). The rightmost column summarizes features which are common to the correlational models proposed by von der Malsburg, Singer and coworkers, and Damasio. For further details, see Singer et al. and Damasio and Damasio (chapters 13 and 17, respectively).

	Barlow	Hebb	Correlational Models
Neural correlate of representation	single cell	assembly	assembly
Code for binding	anatomical convergence	response amplitude	synchrony
Parsimony	low, combinatorial explosion	high, the same cells can participate in different assemblies	
Flexibility	low	rearrangement by "slow" plasticity	high, rapid changes of temporal relationships
Coexistence of representations	possible	impossible	possible, uncorrelated firing of cells belonging to different assemblies

for the Hebb model. In this case, several cortical assemblies raise their average firing rates, and it cannot be determined which of the active cells pertain to which of the representations.

This severe restriction led to the suggestion of a third class of models which may be called "correlational models" for assembly formation. Such models have been suggested by Abeles, von der Malsburg, and more recently also by Singer and coworkers and by Damasio (Table 20.1). The key assumption of these models is that temporal correlation on a fast time scale serves a code for binding. Cells representing a particular object fire synchronously, whereas cells belonging to different representations fire in an uncorrelated manner. This mechanism would solve the coexistence problem mentioned above.

Two particular architectures have been discussed. One model that was introduced by von der Malsburg and later modified by Singer et al. (see chapter 13) implies that cells with oscillatory firing patterns, i.e., cells which exhibit recurrent bursting, may be useful for the establishment of temporal correlation. This notion has now received experimental support (Singer et al., chapter 13). In particular, it has been verified in experiments on cat visual cortex that two assemblies of oscillating cells can coexist in the same cortical region which are distinguished by the absence of fixed phase-relationships between the assemblies (Engel et al. 1991b).

A related model has been suggested by Damasio (see chapter 17). In contrast to Singer and coworkers, Damasio assumes that specialized sets of cells located in

circumscribed "convergence zones" induce the synchronous firing of lower-level neurons. This implies that the binding is carried out at anatomical sites which are different from those containing the elements of the representation. As suggested by Engel, it may be more economical if those cells that are bound into representations accomplish the binding by cooperative interactions among themselves. Current data support this latter possibility. Recent experiments suggest that synchronization of cells in the visual cortex is mediated by interaction of these cells via tangential connections (Engel et al. 1991b). Certainly, the existence of "convergence zones" requires further experimental verification.

Altogether, recent developments in cortical neurobiology provide new ideas about how representations may be created in a highly parsimonious and flexible manner, which is certainly required for adaptive intelligent behavior. However, there are a number of serious and yet unresolved issues that have to be addressed in the future. First, how can hierarchically structured representations be formed in the cortex, i.e., how can a hierarchical relationship be expressed between a representation of the whole object and representations of its parts? Second, how can one generalize over individual representations to form more abstract concepts of objects and events? Third, how can representations sequentially be linked in time, which is probably required, for instance, to track a moving object and establish its identity from one observation to the next?

Argument for a Localist Representation (Durbin)

Despite the reasonableness of accounts for distributed representations, recent evidence of neural localization of function has strengthened the localist reply.

Although the idea of using a separate cell to represent each possible conjunction of concepts and percepts is clearly ridiculous, there is a strong case to be made for the localization of representation even at high levels. By this I mean that the concept is characterized by an activity pattern within a patch of cortex of perhaps a few hundred microns in diameter. There is no real danger of a combinational explosion at the level of, for example, "grandmother," since a standard human vocabulary including proper nouns and known names is under 100,000. Experimental evidence from the last 30 years has been progressively restricting the alternative view that representations are widely distributed. This comes both from the increasing awareness of specificity of human lesion deficits and from physiological data about the localization of specific meaningful properties in progressively higher areas. These data strongly suggest that there are representations localized to within a centimeter, and do not rule out smaller scales. The computational advantage of a localist representation is that it provides independent, nonoverlapping basic components for cognition to combine and manipulate without problems of interference.

If mental processing is carried out by activity in the cortex, then the organization of the substrate for that activity, the cells and connections of the cortex, provide a fundamental constraint. The vast majority of cortical connections are local (on the order of a millimeter or less), and it seems inevitable that much of the computational

processing is local. Even where a concept involves binding together neuronal activity over widely separated areas, it has been suggested that this is achieved by linking those areas in an activity pattern in a local patch of higher association cortex (e.g., the convergence zones) (discussed by Damasio, chapter 17). This emphasis on local circuitry goes hand in hand with the richness of spatial structure found on the cortical sheet. Physiologists have observed localization of function at all scales, from major zones (visual, temporal, frontal, etc.) to areas, to topographic maps in the sensory areas and finally down to ordered substructures on the submillimeter scale (stripes, "columns," patches).

The intimate relationship between this spatial localization and function is emphasized by experiments showing that functional activity is required for the organization of the spatial structure on scales below that of areas. There is a family of models that have been developed over the last 20 years (e.g., von der Malsburg 1973; Willshaw and von der Malsburg 1976; Durbin and Mitchison 1990) that are effective in modeling the experience-dependent development of cortical structure, including the results of experiments which perturb that structure. These models tie important aspects of the large scale layout of the cortex to a small number of basic mechanisms (competition for activity, Hebbian-type synaptic adaptation and local interactions).

An overlapping set of models (e.g., Linsker 1986; Miller et al. 1989) using the same mechanisms give indications of some expected properties of the wiring between neurons. In particular, they suggest that cells will tend to come to respond to expected correlation among their inputs i.e., to patterns repeatedly seen in their input that might provide a good basis for a description of the world. Of course this does not imply that cortical processing is simply correlation. Cognition requires suitable components to work with, and detected correlations may well provide such basic components (Barlow and Foldiak 1989).

Cautionary Note against Localizationism (A.K. Engel, P. König)

We want to express our disagreement with the view that recent findings provide any particular support for localized representations. There can be little doubt that perception of natural scenes and control of complex behavior involve enormously large numbers of cells in the brain, distributed through widely different regions and areas. Contrary to Durbin, it seems to us that a combinatorial problem certainly arises from the fact that one does not only want to represent individual items such as words, but also combinations of words into sentences and paragraphs, for which there are infinitely many possibilities. Apparently, this task cannot be accomplished by small numbers of neurons in a particular location. Recent data from single unit recordings do not seem to support the localist view because receptive field properties are never highly specific. The so-called "face-selective" cells do actually respond to a whole class of objects, and they are never specifically tuned to one particular face. Even if this happened to be the case, there would still be the combinatorial problem of, for instance, linking the activity of the face cell to that of cells representing other parts of

the body. Thus, interactions are certainly required between spatially separate populations of neurons. Finally, it seems erroneous to cite Damasio for a localist viewpoint. His data support the idea that the process of representation requires binding between quite different brain areas, and he explicitly argues for the distributed nature of cortical representations.

CONCEPTUAL KNOWLEDGE

Whatever one's opinion of local vs. distributed representation, one fact that must be explained, or explained away, is the compositional nature of much higher-level knowledge. The very idea that there is such a thing as higher-level knowledge was a recurrent theme in our discussion. Changeux emphasized the need for tracking the logical or epistemological path of information from the level of sensory input to the level of reasoning and planning, the implication being that at higher levels of functional organization, higher forms of knowledge were created. Premack, even more explicitly, pointed out the importance of higher-level knowledge in the control of voluntary action and the significance of the categories of object, mind and number for higher cognitive capacities. Damasio, too, reported on instances of the need for apparently higher-level knowledge, such as might be found in the planning centers, to help agents act with foresight. The common theme to these discussions is that not all information is represented in the same way, nor does it have the same logical structure.

Kirsh followed this theme by elaborating on some of the special features of concepts and conceptual knowledge. He argued that it is possible to distinguish at least three logically distinct capacities involved in the human ability to use concepts:

1. the ability to partition entities into equivalence classes; that is, to recognize, classify, and categorize;
2. the ability to know the meaning of a linguistic term, even if one does not yet attach the right lexeme to the concept; and
3. the ability to combine more or less arbitrarily elements of a scene or an abstraction into a thought.

Research tends to focus on one or the other of these abilities. Of particular interest, however, are the second and third abilities; the capacities to grasp linguistic meanings and to create structured thought. For whereas the ability to partition a set is an ability animals, even low mammals, display, we have yet to find convincing proof that animals can form arbitrary combinations of ideas, such as are displayed by humans in speech.

In current AI theory (Winston 1992), the output of linguistic processing and the output of visual processing are both structure descriptions: a linked list of symbols organized into certain framelike structures. Although there are constraints governing the filling of such frames, there is tremendous freedom too. Within the confines of

semantic and syntactic well formedness, we are free to construct arbitrary propositions. The same applies to visual imagery: we are free to put together arbitrary elements of scenes we have perceived or imagined, providing certain semantic and syntactic constraints are obeyed. This ability to *detach* elements of a scene and make them separate objects of thought confers on the agent the ability to construct thoughts about how the world might be, and what the world that is distant to it in space and time might be like.

One reason to view these thoughts and expectations to be higher forms of knowledge is that they presuppose the ability to *refer* to enduring objects. The notion of reference is itself the subject of a vast literature (Evans 1982). But minimumally, if a creature has the capacity to refer, to have thoughts *about* things, as opposed to just respond to things, it must be able to grasp the difference between being presented with identical twins, one after the other, and being presented with the same person twice — a distinction known as the difference between two tokens of the same type and similar tokens of different types. Creatures who possess the capacity to classify by appearances alone, will be unable to distinguish the two cases.

The formal defence of the type-token capacity flows from an analysis of predication. It is inconceivable that a person can genuinely understand the meaning of a sentence such as "John loves Mary" without also having the ability to understand the meaning of the sentence "Mary loves John." That is because an agent who grasps the concept *love*, as we mean it, knows that it takes two arguments (love x y). It is possible, of course, that the agent only knows the concept *loved-by-John* — a predicate taking one argument. But the fact that we distinguish these predicates is proof that we refuse to grant that an agent has the normal concept of love unless it grasps that it is possible to apply the predicate to different x's and y's . More generally, the ability to predicate *H* of *a* presupposes the ability to predicate *H* of *b, H* of *c*, and also the ability to use different predicates for the same individual, as in *Ha*, *Ga*, *Ja*, so as to guarantee the capacity to refer to individuals (ibid.).

A second and equally profound ability that human concept users possess is the social skill necessary to bind ourselves to the norms of the correct application of a concept, so that we learn to partition sets the way our neighbors do, even if there is no natural grouping based on structural features. Premack mentioned earlier that possession of certain concepts are necessary for pedagogy. One crucial element necessary for higher types of learning is to be able to understand what it is necessary for an answer to be *correct*, or *relevantly similar*. This is one factor making explanation-based learning and one-shot learning possible.

Finally, and perhaps most significant of all, the ability to use concepts means that we are able to recognize a public nonegocentric space where other agents have different spatial perspectives on objects (Kirsh 1991). That is, we have a theory of mind. Not all forms of cooperation require knowing what one's partners will perceive; however, there is a large class of cooperative actions that does, since to distribute goals so as to cooperate in a planned way, we must be able to aim for the same *objective* states of the world—a capacity not shared by animals, to our knowledge.

This difference has deep consequences for theories of behavior because it implies that some behavior does not require fully blown concepts while other behavior clearly does. Concept-free behaviors are the product of skills, which may be defined as complex organizations of perceptuo-motor control systems that enable a creature to manipulate objects in order to regulate certain properties. A cook, for example, has knife-using skills, frying skills, and so on which are highly tuned to local properties that are learned to be important for the task. These properties are monitored from the agent's egocentric perspective. Globally managing such skills, however (Kirsh 1990), requires some element of planning and hence concepts for understanding states of the public, nonegocentric world.

More Lessons from Lesions

Damasio presented a series of results illustrating how patients with lesions in occipito-temporal cortices may have selective losses of access to concepts of concrete entities. For instance, they may be unable to recognize the pictures of certain animals, but be perfectly capable of recognizing most human-made and manipulable items. Those same patients will have no difficulty in comprehending the actions or relationships instantiated by the entities whose concept they fail to access, e.g., a patient with bilateral damage to infero-temporal cortex may consistently fail to recognize pictures of raccoons, but shown pictures of raccoons eating, or fighting, or protecting their offspring will easily identify those actions in spite of not identifying the subject of the picture, i.e.,"it is eating," or "mothering."

Some patients with left anterior temporal lobe lesions become unable to access the names that denote certain entities, but are perfectly able to retrieve the concepts related to those entities and can access the concepts that describe those stimuli. Those same patients can easily access verbs or functions (closed class words such as prepositions or conjunctions), which denote actions or relationships of entities or among entities.

These results imply that the brain utilizes different neural systems to map the properties, actions, and relations of different entities and that, furthermore, it utilizes different systems for the representation of objects and for the representation of the names that denote them. The evidence also makes specific suggestions about how conceptual knowledge is inscribed in neural systems. For instance, as far as concrete entities are concerned, it suggests that the brain does not carry a permanent lexicographic definition of each concept but rather that it reconstructs representations of appropriate elements of typical exemplars of the entity and generates a narrative out of those representations.

The mechanisms for accessing nouns given concepts, or for the reverse operation, uses the framework of *synchronous multiregional retroactivation* (see Damasio and Damasio, chapter 17). The mechanism calls for a "third party" convergence zone, that mediates between convergence zones related to concept representation, on the one hand, and word reactivation, on the other. This binding device is bidirectional and does not embody knowledge about either concept or word but rather knowledge about the

probability of the two having been linked in experience. The framework predicts that such a "third party" mediation should occur at a hierarchy higher than that required for either concept or word processing, and the lesion data offers some support for this prediction: lesions that disrupt name access (or access of concept given the name) are located in left anterior temporal lobe, and more anteriorly than lesions that would disrupt either concept access or word generalization.

We have argued that with concepts comes the ability to plan in the full and proper sense. But much adaptive response does not require planning so defined. Appropriately tuned perceptually driven control systems — skills — can serve a creature remarkably well.

It is characteristic of skills that they evolve as a result of a dynamic relationship with the environment actively organizing perception to probe the environment for properties that are maximally informative relative to the tasks and goals the creature has. We therefore next inquired into expectation-based perception and behavior.

EXPECTATION-BASED PERCEPTION AND BEHAVIOR

Most animal, if not human, behavior is goal-directed, that is, it aims to fulfill the needs of the animal: safety, food, shelter, reproduction. The existence of a goal presupposes that the animal has an expectation that is matched when the goal is fulfilled. Altman contended that to analyze these expectations, it is first necessary to appreciate that a large part of an animal's sensory input arises from self-generated stimuli: as an animal moves through the world, its changing relationship with the world provides a continual stream of input to the sensory system; likewise the internal state (equivalent to the somatic states posited by Damasio) continually changes, both on the level of short-term proprioceptive information and changes in concentrations of chemicals, such as glucose and hormones. These on-going changes, generated by action, form internal and external loops through which the animal receives information about the results of its action.

Novel stimuli, that is, changes arising in the environment independent of the animal's actions, actually form a small part of the input, except in humans, who have specialized in collecting and dealing with novel stimuli. Both our anthropocentric view of behavior and the prevalence of experimental designs in which stimuli are presented to and responses demanded of immobilized subjects have tended to emphasize the input–output aspect and have led to the neglect of the far more important output–input component. Furthermore, once an animal starts responding to a novel stimulus, that stimulus loses its novelty and becomes part of the animal's world, thus part of the ouput-input loop.

Information flowing through the output–input loop is matched against the expectation, or internal representation, of the goal. This may be either hard-wired or learned: crickets with species-specific courtship songs have neurons tuned to specific features of that song (Schildberger 1987). Hebb made the point that even the organization of

stepping involves a hard-wired expectation, in the form of sensory information that is gated to be available only in the appropriate phase of the movement.

A descriptive model of the insect motor system that includes the output–input loops (external and internal), as well as internal loops between the local networks at various levels of the nervous system (Altman and Kien 1989) has the form of an attractor neural network. Dynamic forms of such networks turn out to be excellent for formal modelling of this system on an abstract level (Nützel, Kien, Bauer, Altman and Krey, in preparation). One problem that is being examined with this method is the determination of the conditions that cause an animal to switch from one behavior to another when there is no sudden discontinuity in the intensity of the input conditions; in other words, what determines whether an expectation has been fulfilled? This turns out to be a general class of problem that applies to many dynamic systems from chemical oscillators to weather forecasting.

Expectation-based Learning: A Computational Viewpoint

Expectations may provide a bridge between low-level stimulus-response behavior and higher-order mental activities, according to remarks made by Rivest.

In this view, expectations provide a first mechanism for representing the future—certainly a key requirement for survival in a complicated world. In its simplest form, an expectation may merely be an *anticipation* that some event is about to happen. (Seeing the lightning, one expects to hear thunder.) Given that the future is also affected by one's own actions, an expectation may also represent a *potentiality* present in the current situation that may or may not be realized, depending on what actions are taken. (One may expect to find coins in one's pocket if one looks or to see a door if one turns around.) Such expectations can be simple or complicated, but will always have the property that they correspond to *testable predictions* about the current state of the world; an expectation is the belief that a certain prediction is true in the current state of the world. We may conjecture that a key function of intelligence is to maintain an accurate set of expectations in a dynamic environment.

It is also useful to consider an expectation as itself a sensation; expecting to see the door behind oneself if one turns around is a vision of a (possible) future. Such expectations help one to perceive one's immediate world as if one had greatly extended sensory capabilities (in this case, eyes in the back of one's head).

However, expectations are not founded on immediate sensory perception but derive from knowledge gleaned about the structure of the world. One learns to expect thunder after lightning or to expect to find food in the refrigerator if one opens it. Such knowledge may relate sensations to expectations (thunder after lightning), actions to expectations (if I look up I will see blue sky), or even expectations and actions to expectations (if I expect to get a coke from the vending machine after I put two quarters in it, then if I put one quarter in, I expect to get a coke if I put another quarter in). The way in which expectations are affected by sensations, actions taken, and other expectations forms an elementary "world model."

Nontrivial concept formation can be based upon the recognition of equivalences between expectations. A notion of an object may evolve from the cluster of interrelated expectations learned about that object. In a theoretical study, Rivest and Schapire (1990) demonstrated the application of this principle to "Rubik's Cube"; a learning algorithm was able to infer the structure of Rubik's Cube by experimentation, even though only three squares of the front face of the cube were observable. The algorithm invented concepts representing the other 51 visible squares of the cube by detecting equivalences between expectations.

We surmise that expectations will play an increasingly important role in our understanding of intelligence and its neural underpinnings.

CONSCIOUSNESS

To understand the function of consciousness it may be helpful to define the part of the brain involved in its mediation and to then try to evaluate its possible contribution to intelligence.

Stoerig suggested that consciousness allows the reflection on some parts of inner (motivations, emotions, feelings, and desires) and outer (social and physical environment) reality which are accessed via our perceptual systems. These systems provide information about our internal and external states, which to a certain extent only are consciously represented. Consciousness thus depends on perception and can be seen as a composite of modality-specific perceptual consciousnesses: to see and know that you see, to hear and know that you hear, to feel and know that you feel, to want and know that you want. Perceiving and knowing that you perceive do not always go hand in hand, and they may be pathologically dissociated in patients.

One example in question is provided by patients with "blindsight." These patients have suffered lesions in their primary visual cortex (V1) which cause visual field defects in the topographically corresponding portion of their contralateral visual hemifield. The field defects are perimetrically assessed and clinically classified with respect to the extent, density, and position within the visual field. Defects that are classified as absolute are experienced as blind by the patients. When visual stimuli are presented within these defects, the patients report no visual sensation. They may still, however, be able to process the visual information, as can be shown by measuring skin conductance response or pupil reflexes in response to visual stimulation or by asking them to guess whether or not a stimulus has been presented, where it has been presented, or which one of different stimuli has been presented (see Weiskrantz 1991, for a recent review). Depending on the experimental conditions, patients may perform at high levels of accuracy and sensitivity in such tasks, getting 100% correct in a motion detection task or exhibiting a detection sensitivity which is reduced by no more than 0.5 log units as compared to sensitivity in the normal hemifield (Stoerig and Cowey 1991).

Despite this level of performance, however, the patients claim that they do not see anything, that they are only guessing; in some cases it may even be difficult to convince them to give it a try because it seems an eccentric request.

It is not difficult to understand how the patients can process visual information in their field defects since up to nine pathways from the retina remain in the absence of V1 (Cowey and Stoerig 1991). What is remarkable is the dissociation between the performance and the visual experience, or lack thereof. It seems as if the visual information they are able to process can no longer be represented in perceptual consciousness: they "see" but they do not know that they see. As this happens only after lesions of V1, while lesions in higher visual areas produce specific deficits such as cortical color blindness, motion blindness, or face blindness, it may shed some light on the neuronal basis of visual consciousness. It would be interesting to know whether the loss of visual perception is due to the privileged position of V1 that distributes a vast part of visual information to the extrastriate visual cortical areas, or whether V1 is really responsible for accessing perceptual consciousness in that it not only forwards the information to the specialized higher areas, but also receives their outputs and maps them on its precise map of the visual world that the other areas lack (Stoerig and Cowey 1992). If this were the case, vivid visual dreams or imagery should activate visual cortical areas as far down as V1 in people who are able to really "see" internally generated images. At present, it is not known whether complete bilateral destruction of V1 abolishes all "seeing" in that sense, because no such patients are known.

To better understand the basis of blindsight and visual consciousness, monkeys with ablations of V1 are studied, who, like the patients, exhibit residual visual functions in their field defects (e.g., Pasik and Pasik 1982). Yet it is not known whether the monkeys, like the patients, have blindsight, i.e., whether they are also convinced that they are only guessing when in fact they perform at high levels of statistical significance. To test this, Cowey and Stoerig (1991b) are presently teaching monkeys two different response paradigms, between which they have to choose to optimize their rewards. They should use the first response when they know, or are confident, that they see the stimulus, because they are rewarded for correct responses only. They should use the second response when the task is very difficult because they are rewarded for 75% of responses regardless of whether or not they were correct. This is the better strategy to use when you are only guessing because in a two alternative forced choice you score approximately 50% correct when you are in fact just guessing. Therefore, we want to then test them in their cortically blind fields once they have mastered these strategies to see which strategy they will adopt. This should tell us whether they experience themselves as blind, as only guessing. Although this is a specific test for visual consciousness—a test we have yet to prove works—it may eventually help us design tests to distinguish between the more general "knowing" and "knowing that you know" by testing both the performance which is based on knowing, and the confidence or meta-knowledge about this basis.

Consciousness: A Methodological Comment

Consciousness can be treated as a recursion on knowing—an individual uses information, knows that he uses information, knows that he knows he uses information. We can equate consciousness with the third state—knows that he knows—and equate it, albeit in lesser degree, to the second state as well.

Of course, every species uses information, but which know that they do? Do chimpanzees? To answer this question we used a two-step procedure (Premack 1986). First, four young animals were taught to use plastic "words" in a simple way. Each animal was shown a plastic word displayed on a writing board, then trained to select one of three objects that was "named" by the word. When the animals attained criterion, we introduced this critical step: the writing board was rotated 180°, concealing the word and depriving the animal of information. How did they react to this loss? One animal acted as though nothing had been changed, choosing among the objects as promptly as ever, performing, of course, at chance level. A second animal performed in the same way but with great reluctance, whining, and "complaining." It appeared to "know" something was wrong, but not what, or how to correct the problem. The third and fourth animals redeemed the species: from the first trial, they went directly to the rotated board and turned it around: then they made their choice. Shyness or a reluctance to handle the board could not explain the behavior of the two animals that did nothing. On alternate trials, when the board was rotated only 90° (so that the animals could see the word simply by adjusting their posture) they again did nothing. Two of the four animals gave clear evidence that they knew they had used information, for when the information was removed, they took steps to restore it. This simple paradigm, the restoration of missing information, can be used to determine whether an individual is conscious of his actions.

A second test entailed a more complicated setting. In the test (Premack 1988a), the chimpanzees were required to choose between two opaque containers, one of which held food. The chimpanzee's view of the baiting was blocked by a barrier but they could see two onlookers: one who had a clear view, the other's view was blocked like that of the animal. Before choosing between the containers, the animal was given the opportunity "to ask the advice" of one of the onlookers, and then pointed to one container or the other: the correct container if he was the "clear view" onlooker, the incorrect one if he was the "blocked" onlooker.

Three of the four chimpanzees chose the correct onlooker from essentially the first trial, giving evidence for the assumption that chimpanzees have a "theory of mind," viz., they understand the conditions on which seeing depends and attribute it appropriately. Does the chimpanzee know why it makes the choice it does? Does it know that it chose a particular onlooker because it attributed seeing to him? To answer this question we next gave the animal two kinds of trials: those on which it could not see where food was placed and those on which it could (the barrier was removed). Both onlookers were available as before, but now the animal had to pay for their advice (using tokens that it had been taught to exchange for candy). One should not, of course,

pay to be given information about where food is located if one already knows where it is: one should pay only if one does not know. An individual who does this can be said to know why he chooses a particular onlooker and to be conscious of his actions, or on the road to becoming so. However, none of the chimpanzees made this discrimination. They paid for the trainer's advice on all trials, whether needed or not. In a simple setting, chimpanzees gave evidence of consciousness, of knowing what they were doing, but they failed to give evidence of consciousness in a complex one.

Yet More Lesion Lessons

Damasio outlined several results in humans with lesions which exemplify performances without consciousness.

For instance, patients with face agnosia are able to generate accurate discriminatory skin conductance responses to faces which they clearly can not recognize consciously. Those same patients also generate scanpaths to familiar faces which are clearly different from those that they would generate for entirely novel faces.

Patients with amnesia provide other examples. For instance, patient Boswell develops preferences for persons who often reward him and avoids those who do not, in spite of the fact that he has no conscious knowledge of the person's identity. This is an example of habit learning. It does not require hippocampal circuitry and it can occur below the level of consciousness.

Damasio also outlined some requirements for self-consciousness as a complex function distinct from awareness and wakefulness. They are:

1. the ability to correlate current knowledge with previously acquired knowledge;
2. the ability to refer the correlated information to the self defined as follows:
   a) a repeatedly updated collection of autobiographical items from the individual's past and expected future, i.e., a memory of past and planned future;
   b) a repeatedly updated representation of the individual's somatic state; this collection may not be sufficient but the different components are seen as necessary.

It is difficult to imagine that the higher levels of human intelligence (e.g., pertaining to personal relationships and creativity) could operate in the absence of self-consciousness as defined above.

A Reductionist Account of Consciousness

Koch argued concisely for a radical reductionist thesis concerning the neuronal localization of awareness (Crick and Koch 1990).

"When will I know that a given neuron is involved in awareness?" Let us consider an awake monkey shown a rivalrous stimulus. In such an experiment, the left eye is

presented with an upward moving grating, while the right eye is stimulated by a downward moving grating. Under these conditions of binocular rivalry, the percept is ambiguous and alternates between seeing only *upward* or only *downward*. Early on in its visual system, say in the retina or in primary retinal cortex, neurons will respond to the physical stimulus present in one or the other eye (that is, to either the upward or the downward motion, depending on which eye provides the dominant input; similar to the experiment carried out by Logothetis and Schiller 1989). Yet somewhere in the higher areas of the monkey visual systems, (in particular, in the superior temporal sulcus of monkeys), neurons should fire corresponding to what the monkey perceives, independently from the physical stimulus, thereby reflecting the percept or visual awareness of the monkey. Where are these neurons? Is there anything particular about these neurons? Do they fire in a special manner? Thus, we would like to know what sort of neuronal activity is responsible for awareness.

In general, at least three, nonexclusive criteria exist for the neuronal adequacy of awareness:

1. A sensory event causes awareness if a certain minimum number of neurons, say one million, become active. There is certainly some validity to this argument.
2. Neurons in a special part of the brain must become activated, say in IT or in the frontal lobes. Thus, if these neurons fail to become activated, the animal can still make a behaviorally relevant response, but the sensory information will fail to cause awareness. Conversely, one may never become aware of the activity of neurons, say in the brainstem or in the cerebellum. We currently believe that there must be some truth to this notion and that neurons have to be located in the "cortical" system (including neocortex, hippocampus and olfactory cortex) to cause awareness.
3. The neuronal activity has to be of a particular kind, say a high-frequency burst, an oscillatory response, or a phase-locked activity. This possibility is quite elegant, since it provides for an explanation of why certain events (such as blindsight) fail to cause awareness yet lead to meaningful behavioral motor output. In this case, the visual stimulus induces neuronal activity over the cortical system (inducing a motor output) yet fails to cause this special type of firing. Thus, it will not cause awareness. Crick and Koch (1990) currently postulate that phase synchronized oscillations in particular parts of the cortical system are necessary to cause awareness, while the absence or reduction of synchronized oscillation will cause a loss or a reduction of awareness.

Cautionary Comment

Changeux expressed surprise that some of the recent speculations about the mechanics of conscious awareness are based on experiments carried out with animals in which

the "state of consciousness" is under control (e.g., anesthesia). Livingstone and Hubel, in the early 1980s, carried out experiments relevant to this point. They followed the state of activity of defined neurons from striate cortex for extended periods of time (hours) during which the cat wakes up (becomes "conscious") and falls asleep (drowsiness). Interestingly, oscillations are recorded during the sleep phase—disappear when the cat wakes up. Most characteristic is the "sharpening" of the evoked signal upon stimulation e.g., by moving bars with reduction of "noise" activity. Such enhanced signal-to-noise ratio may facilitate traumatic coherency activity between neurons, through the elimination of redundant "static" oscillations.

Future research might perhaps first be focused on the delimitation of the compartments of the brain (if they exist) which are the most susceptible to the state of wakefulness (and or attention) and *then*, but only then, on the analysis of the properties displayed by the relevant cells, ensemble of cells, or even receptors.

Cautionary Comment (A.K. Engel)

Engel commented briefly on the relationship between oscillatory activity and state of alertness. In the discussion of this issue, there was certainly some confusion about the frequency band of the "oscillations" various speakers were referring to. In their 1981 paper, Livingstone and Hubel investigated *slow* EEG waves in the frequency band below 10 Hz, and it is certainly true that these slow waves disappear during arousal of the animal. However, the oscillations that have been considered relevant for visual cortical information processing (see Singer et al., chapter 13) have frequencies between 30 and 80 Hz. These *fast* oscillations are not dependent on the state of alertness and can well be observed in awake cats and monkeys. In this respect, Changeux's statement may be somewhat misleading.

What Happens During Sleep? (Roland)

What happens in the brain during sleep, perception, and when keeping something in mind in the awake state?

When the brain is asleep or anesthetized it is by definition not conscious. Sleep in the form of deep sleep (in stages II–IV) reduces the metabolic rate of glucose diffusely all over the brain by 25–35% (Heiss et al. 1987). Rapid eye movement (REM) sleep is associated with a general level of metabolic rate for glucose which is identical to that in awake subjects during rest. Also the pattern of metabolic activity during REM sleep is the same as that in awake subjects who are just resting with their eyes closed.

An important faculty which is a subfunction of consciousness is the ability to keep something in mind for a short time (working memory). In an experiment subjects were shown two pairs of visual stimuli, of which the first differed in orientation and the second in grating frequency; they did not know which pair to discriminate until either a green or a red lamp was lit. This gave an activation of V1 and immediate visual areas

plus a strong activation of the frontal polar cortex and adjacent part of middle frontal gyrus. Discrimination of the gratings or the orientation alone did not give this particular prefrontal activation (Gulyas and Roland 1991).

Actively keeping something in visual perceptual awareness was associated with activation of a subset of the visual areas plus activation of frontal polar cortex. During visual perception, V1 and a subset of visual association areas are active, most often in combination with activation of a few district fields in the prefrontal cortex and activation of the frontal eye fields. The subset depends on the nature of visual stimulation (Roland et al. 1990; Gulyas and Roland 1991; Haxby et al. 1989).

Similarly, during somatosensory perception, S1 and a subset of somatosensory association areas are active. Again the subset which is activated will depend on the nature of the stimulation (Seitz et al. 1991, Roland and Seitz 1991). In addition, distinct fields in the prefrontal cortex (different from these active in visual perception) are active. These fields each cover at least the space of 10^7 neurons.

CONCLUSIONS

Our goal, as stated in the introduction, was to consider the prospects for establishing theoretically fruitful connections between computational models of intelligence and neural models of the brain. At a detailed level we must admit that we have not ourselves succeeded in establishing actual connections between specific models. Our inquiry has been focused more on foundational issues than on comparing specific models of general intelligence with specific data culled from behavioral and physiological studies.

Yet there was a quiet optimism, albeit restrained, that a constructive relation between computationalist and neuroscientist may be forged in the not too distant future. The modest success achieved in the study of vision is encouraging, as are some of the links emerging between lesion studies and computational models even as universal as Newell and Rosenbloom's Soar. Perhaps the fields are indeed within shouting distance of one another.

REFERENCES

Altman J.S., and J. Kien. 1989. New models for motor control. *Neural Comp.* **1**:173–183

Barlow, H.B., and Foldiak, P. 1989. Adaptation and decorrelation in the cortex. In: The Computing Neuron, ed. R.M. Durbin, C. Miall, and G.J. Mitchison, pp. 54–72. Wokingham: Addison-Wesley.

Cowey, A., and Stoerig, P. 1991a. The neurobiology of blindsight. *TINS* **14**:140–145

Cowey, A., and Stoerig, P. 1991b. Reflections on blindsight. In: The Neuropsychology of Consciousness, ed. D.Milner, pp. 11–37. London: Academic.

Crick, F., and C. Koch. 1990. Some reflections on visual awareness. *Symp. Quant. Biology Cold Spring Harbor* **55:**1039–1047.

Damasio, A.R., H. Damasio, D. Tranel, and J.P. Brandt. 1990. Neural regionalization of knowledge access: Preliminary evidence. *Symp. Quant. Biology Cold Spring Harbor* **55**:1039–1047.

Damasio, A.R. 1989. Concepts in the brain. *Mind Language* **4**:24–28.

Damasio, A.R. 1990. Category-related recognition defects as a clue to the neural substrates of knowledge. *TINS* **13**:95–98.

Damasio, H., and A.R. Damasio. 1989. Lesion Analysis in Neuropsychology. New York: Oxford Univ. Press.

Damasio, A., D. Tranel, and H. Damasio. 1989. Amnesia caused by herpes simplex encephalitis, infarctions in basal forebrain, Alzheimer's disease, and anoxia. In: Handbook of Neuropsychology, ed. Y.F. Boller and J. Grafman, vol. 3, (ed. L. Squire), pp. 149–166. Amsterdam: Elsevier.

Damasio, A., D. Tranel, H. Damasio. 1990. Face agnosia and the neural substrates of memory. *Ann. Rev. Neurosci.* **13**:89–109.

Dasser, V., I. Ulbaek, and D. Premack. 1989. The perception of intention. *Science* **42**:162–164.

Durbin, R.M., and G.J. Mitchison. 1990. A dimension reduction framework for cortical maps. *Nature* **343**:644–647.

Engel, A.K., P. König, A.K. Kreiter, and W. Singer. 1991a. Inhemispheric synchronization of oscillatory neuronal responses in cat visual cortex. *Science* **252**:1177–1179.

Engel, A.K., P. König, and W. Singer. 1991b. Direct physiological evidence for scene segmentation by temporal coding. *Proc. Natl. Acad. Sci.* **88**, in press.

Evans, G. 1982. The Varieties of Reference. Oxford: Oxford Univ. Press.

Gulyas B, and P.E. Roland. 1991. Cortical fields participating in form and color discrimination in the human brain. *Neuroreport* **V2 N10**:585–588

Haxby, J.V. 1989. Neurospsychological evaluation of adults with Downs syndrome: Patterns of selective impairment in nondemented old adults. *J. Mental Deficiency Res.* **33**:193–210

Kirsh D. 1990. Preparation et Improvisation. *Reseaux* **43**:111–120.

Kirsh, D. 1991. Today the earwig, tomorrow man. *Art. Intell.* **47(1–3)**:161–184.

Linsker, R. 1986. From basic network principles to neural architecture. *Proc. Natl. Acad. Sci.* **83**:7508–7512, 8390–8394, and 8779–8783.

Livingstone, M.S., and D.H. Hubel. 1981. Effects of sleep and arousal on the processing of visual information in the cat. *Nature* **291:**554–561.

Logothetis, N.K., and Schall, J.D. 1989. Neuronal correlates of subjective visual perception. *Science* **245 (4919)**:761–763.

Miller, K.D., J.B. Keller, and M.P. Stryker, M.P. 1989. Ocular dominance column development, analysis and simulation. *Science* **245**:605–615.

Newell, A. 1991. Unified Theories of Cognition. Cambridge MA: Harvard Univ. Press.

Nützel, K., J. Kien, K. Bauer, J.S. Altman, and U. Krey. 1992. Dynamical attractor neural networks as models for motor control. *Neural Com*p., submitted.

Pasik, P., and T. Pasik. 1982. Visual functions in monkeys after total removal of visual cerebral cortex. *Contrib. Sensory Physiology* **7**:147–200

Premack, D. 1976. Intelligence in Ape and Man. Hillsdale, NJ: Erlbaum.

Premack, D. 1984. Pedagogy and aesthetics as sources of culture. In: Handbook of Cognitive Neuroscience, ed. M.S. Gazzaniga. New York: Plenum.

Premack, D. 1986. Gavagai or the Future History of the Animal Language Controversy. Cambridge, MA: MIT.

Premack, D. 1988a. "Does the chimpanzee have a theory of mind? revisited." In: Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkey, Apes, and Humans, ed. W. Byrne and A. Whiten, pp. 286-322. Oxford: Oxford Univ. Press.

Premack, D. 1988b. Minds with and without language. In: Thought without Language, ed. L. Weiskrantz. Oxford: Clarendon.

Premack, D. 1990. The infant's theory of self-propelled objects. *Cognition* **36**:1–16.

Premack, D. 1991. The aesthetic basis of pedagogy. In: Cognition and the Symbolic Processes: Applied and Ecological Perspectives., ed. R.R. Hoffman and D.S. Palermo. Hillsdale, NJ: Erlbaum.

Premack, D., and G. Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* **1**:515–526.

Rivest, R.L., and R.E. Schapire. 1990. A new approach to unsupervised learning in deterministic environments. In: Machine Learning, An Artificial Intelligence Approach, ed. Y. Kodratoff and R.S. Michalski, vol. 3, pp. 670–684. Morgan Kaufmann.

Roland P.E. et al. 1990. Asymmetric pneumatization of the petrous apex. *Otolaryngology - Head and Neck Surgery* **103 (1)**:80–88

Roland, P.E., and R.J. Seitz. 1991. Positron emission tomography studies of the somatosensory system in man. *Ciba Found. Symp.* **163:**113–120; discussion 120–124.

Rosenbloom, P., et al. 1991. A preliminary analysis of the Soar architecture as basis for general intelligence. *Art. Intell.* **47(1–3)**:289–326.

Schildberger, K. 1987. Acoustic communication in crickets: Behavioral and neuronal mechanisms of song recognition and localization. In: Nervous Systems in Invertebrates, ed. M.A. Ali, pp. 603-619. New York: Plenum.

Seitz, R.J. et al. 1991. Somatosensory discrimination of Shape — tactile exploration and cerebral activation. *Eur. J. Neurosci.* **3(6)**:481–492

Simon, H. 1962. The architecture of complexity. *Proc. Am. Phil. Soc.* **106**:467–482.

Spelke, E. 1985. Perception of unity, persistence, and identity: Thoughts on infants' conception of objects. In: Neonate Cognition, ed. J. Mehler and R. Fox, pp. 89–113. Hillsdale, NJ: Erlbaum.

Stoerig, P., and A. Cowey. 1991. Increment-threshold spectral sensitivity in blindsight. *Brain* **114**:1487–1512.

Stoerig, P., and A. Cowey. 1992. Blindsight and perceptual consciousness: Neuropsychological aspects of striate cortical function. In: Functional Organization of the Human Visual Cortex. Oxford: Pergamon, in press.

Tranel, D. and A. Damasio. 1985. Knowledge without awareness: An autonomic index of facial recognition by prosopagnosics. *Science* **228**:1453–1454.

von der Malsburg, C. 1973. Self-organisation of orientation sensitive cells in the striate cortex. *Kybernetik* **14**:85–100.

Weiskrantz, L. 1991. Outlooks for blindsight: Explicit methods for implicit processes. *Proc. Roy. Soc. Lond. B* **239**:247–278

Wellman, H. 1991. Children's Theories of Mind. Cambridge, MA: MIT.

Wimmer, H., and J. Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* **13**:103-128.

Willshaw, D.J., and C. von der Malsburg. 1976. How patterned neural connections can be set up by self-organisation. *Proc. R. Soc. Lond. B* **194**:431–445.

Winston P. 1992. Artificial Intelligence. Reading, MA: Addison Wesley.

Standing, left to right:
G.S. Stent, H. Damasio, P. Stoerig, A.K. Engel, R. Durbin, P.S. Rosenbloom, P.E. Roland

Seated, left to right:
J.-P. Changeux, D. Premack, D. Kirsh, A.R. Damasio