Cortical architecture and self-referential control for brain-like processing in artificial neural systems

Edgar Körner, Gen Matsumoto

1998

Artif Life Robotics (1998) 2:170–178

ORIGINAL ARTICLE

Edgar Koerner · Gen Matsumoto

Cortical architecture and self-referential control for brain-like processing in artificial neural systems

Received: March 23, 1998 / Accepted: April 23, 1998

Abstract Progress in understanding the way the brain processes information while it is constantly interacting with the sensory environment is hampered by inadequate models and theories. Current models and theories of brain computing are, obviously, still not completely correct when confronted with so-called real-world problems. Sensory recognition and the subsequent selection and optimization of a proper behavior are basically constraint satisfaction problems. Both conventional AI and current formal neural network systems operate with set constraints: the architecture and parameters are defined a priori, and then the input data are structured according to these set constraints on the learning process. However, as long as the constraints are set from outside the system (by the programmer, designer), the system has no ability for self-organization. There is the ability for adaptation within these a priori defined limits, but not the ability to include new knowledge into the consistent relational framework of existing knowledge beyond the prespecified constraints. Therefore, self-organization of constraints in complex systems is the key problem for getting self-organization of knowledge representation under real-world conditions. We show that a value system and self-referential control in a modular architecture are crucial prerequisites for both robust recognition of sensory input and the ability to integrate new knowledge into the already acquired knowledge representation. Finally, we outline a philosophy and propose a model approach that is a first step toward implementing those capabilities in artificial neural systems.

E. Koerner (✉)
Honda R&D Europe (Deutschland) GmbH, Future Technology Research Division, Carl-Legien-Str. 30, Offenbach, 63073, Germany
Tel. +49-69-89011730; Fax +49-69-89011749
e-mail: Edgar.Koerner@f.rd.honda.co.jp

G. Matsumoto
Riken Brain Science Institute, Brainway Group, 2-1 Hirosawa,
Wako, Saitama 351-0198, Japan

Key words Artificial neural systems · Self-referential control · A-priori knowledge · Top-down self-organization · Emotion-based value system

Need for a novel approach to brain-like computing

Neural architectures represent algorithms that organize computation

In recent years, neuroscience has made big leaps forward in both investigation methodology and insights into local mechanisms of processing sensory information in the brain. However, we still do not really know what happens in the brain when one recognizes a familiar person, or moves around navigating seemingly effortlessly through a busy street. Because of the great complexity of biological systems, extended chains of mathematical reasoning have less importance in biology than in mathematics, physics, and engineering – unless corroborated at every pertinent step by empirical data. However, many current artificial neural network models and brain dynamics theories seem more motivated by mathematics and simulation technology than by a genuine interest in understanding brain function. "Neurally inspired" formal neural networks and connectionist semantic models avoid a direct relation between their elementary processing nodes and real neurons in terms of what they actually represent, and how they may be embedded in a purposive architecture to allow the flexibility of processing we want our artificial systems to have, but such questions have to be tackled if we are to develop a realistic theory of the brain. Nonphysiological models may provide interesting metaphors to explain isolated phenomena, but are unlikely to make a major impact.

Fitting current formal neural networks to represent certain aspects of the architecture and function of neural structures is surely not the way to understand the brain. Formal neural networks are graphical notations of known approximation algorithms, developed in the context of the von Neumann computer metaphor. That the brain probably does the job in rather a different way is strongly suggested by our failure to get anywhere near to the performance of our brain in doing things that are very easy in everyday life, but hopelessly tricky for the way we are accustomed to organize computation in current computer technology.

Computer programs are axiomatic systems, since all subsequent procedures and world models will be derived from the set of abstractions (symbolic categories, rules of combination) produced by the designer. The computer does not have the background to update or modify the evaluations that led to these abstractions, since this knowledge resides in the brain of the designer. These abstractions are the "result of thought, but not the basis of thought".[1] Assuming that neurobiological processes have no other properties than those that current computing already replicates (e.g., assembling, matching, and storing of signal or symbol constructs) limits any models of brain computing to what current computers can do, and consequently prohibits insights into how the brain can avoid the inflexibility of rule-bound mechanisms, how it can escape the trap of combinatorial search, and how it can solve the problems of smooth transition between signal and symbol processing, efficient real-time coordination, etc. The brain is not organized like the hardware of current computers: the wiring is highly variable, and representations change over time. Behavior (and hence, also the computation that generates just that behavior) in neural systems seems to some extent self-generated in loops. Sensory input leads to brain activity that results in behavior, which leads to further sensation, which leads to further brain activity, ... This rearrangement of neural systems to generate appropriate behavior is not produced by a process comparable to a software compilation process. Each new perceptual categorization and sensory–motor coordination links hardware components together in new ways, creating new structures within the population of physical elements available for further activation and recombination.

That means that to obtain a better understanding of the type of computation probably utilized by the brain, we have to contrast neural computation with the programmed devices that adapt within fixed, predesigned constraints, instead of explaining brain processing in terms of those systems.

To understand the properties, and develop models, of mechanisms that today exist on earth only as biological systems, the right questions have to be asked, and the key problem seems to be to understand how the brain acquired its architecture and how it updates the memory that is expressed in this architecture, i.e., autonomous learning, including evolution and development.

Close reference to physiology is a necessary, but not sufficient, prerequisite. Neural systems have to be investigated within the purposive organization of a behavior. Even a bird's wing would be a mystifying structure if one did not know its purpose. Models that simulate the processing of isolated procedures in neurally inspired architecture will not take us beyond the von Neumann paradigm. Defining a truly hard problem of recognition or behavior, and then trying to understand how the system could solve this problem under the constraints provided by the functional architecture and the existing neural hardware (but not necessarily by already existing algorithmic structures) should result in some progress. For a complex system like the brain, functional organization matters, and it cannot be discarded without rendering the model obsolete.

To understand brain computing, one has to accept the "natural law" provided by the functional architecture and the sound experimental data instead of settling comfortably upon methodologies and models that involve only the perfect simulation of arbitrary software laws. The solution may rest in understanding the development of the system from the bottom up. For instance, flapping wings and having feathers were not the crucial steps toward the development of the aeroplane. It was necessary to follow the "phylogenetically" correct order of implementation of behavioral abilities and their related controls: first the airfoil principle, then dynamic balance, and last propulsion.[2] That means, the order of acquiring certain abilities matters. We should not focus on a final version of a sophisticated fixed architecture, but instead find the way this architecture evolved from showing simple to showing complex behavioral capabilities.

The crucial role of a value system for the development of self-referential architecture and control

Complex systems like the brain must have an incredible number of control mechanisms which are directly related not to the *content* of sensations, categorizations, etc., but to the *process of making* these sensations and categorizations at all, and to maintaining a unitary organization of the system that allows its co-ordination according to the changing environment, e.g., synchronizing, managing, and arbitrating among different functions. The brain has evolved within these constraints of global control. Hence, without knowing at least the general characteristics of this control, and the order in which the controls developed, we may easily fail to understand the true key points of the system's function. Sensory processing is never a purpose by itself, but serves for behavior control. The smarter the sensory system, the smarter the sensory-guided behavior. However, the definition of what the sensory input means is always decided at the highest instance of each brain, in direct reference to the "limbic system," which belongs to the phylogenetically (and also ontogenetically) oldest parts of the brain, and which can be considered as the "value system" of the brain. The limbic systems deal with emotions and serve for a coarse judgment of the current relative state of the creature and its environment: it evaluates a state as "good" or "bad" for survival, or also simply for the present desire of the creature. Only recently has it been understood that emotions are directly related to the definition of the relation {input pattern → fundamental types of behavior}.[3] This is not a fine discrimination between slightly different alternatives, but a definition of very basic behavioral modes that are selected by that part of the brain (the limbic system) which is the "watchdog" of the essential interests of the living creature (survival should be at the top of the list). Philosophically, this value system may be most closely related to the concept of "self" compared to other parts of the brain.

The brain is acting in response to changes in its environment, based on an evaluation of the situation from its "self"-interest, protecting its own existence, well-being, etc. Probably it is this "being a self", not only a reactive automaton, that is at the root of our flexibility in processing and behavior. The system need not be told any categorization. It is always interpreting objective syntactic sensory information from the viewpoint of its subjective internal description of the outer world, including itself as part of that world.

The internal representation of the outside world is defined by the value system selectively according to the behavioral needs of the organism that is the host of the "self". This permanent subjective redescription of objective sensory information into a consistent internal representation of the world seems to be a key process allowing robustness and flexibility, for instance solving the symbol grounding problem, and allowing rapid coordination of available resources for smart real-time behavior. Enforcing and keeping the consistency of this internal representation requires powerful control of self-reference, since any new information has to be evaluated by the already acquired knowledge, and then integrated into the existing relational architecture of the subjective knowledge representation.

Self-referential control allows self-organization of knowledge representation

Biological systems represent the world (including themselves) in a way that allows the optimal control of behavior by means of prediction, and the iterative tuning of this prediction to minimize prediction errors.

"Self-reference" is a crucial control principle in this type of cognitive ability: any modification of the sensory input states produces a prediction error (since the previous prediction was based on the previous sensory inputs) which the supervisory control ("self") has to compensate for by a modified prediction. If this characteristic of modification cannot be compensated for by an existing repertoire of predictions (which represent the already acquired knowledge on the subject's action within that environment), the prediction error and the sites in the representational framework where those errors occur define what is new (hence, what has to be learned). The complete situation that caused the steady prediction error does not have to be learned, but only the difference between the best existing prediction for that situation and the required prediction accuracy which eliminates an in tolerable prediction error sufficiently. In that way the system always makes a self-reference when behaving in the environment: necessary new knowledge is defined based on the impossibility of predicting the system's control state from existing knowledge. Hence, only that part is added to the existing knowledge which is needed to tune-up the prediction. In this way the new knowledge is always integrated into the relational framework of the existing knowledge.

Therefore, "self-referential control"

- enforces a consistent knowledge representation,
- allows the utilization of existing knowledge to make predictions even in an unknown environment, since to any sensory input, the brain responds first with the best fitting prediction (recall of acquired memory), which already has a value-based behavioral quality.

Hence, the system can react even to unknown events by trying to find the best analogy memorized from its experience. It can also start learning without a teacher, since it can apply a gradient strategy for improving its predictions based on a fairly good "initial hypothesis."

Therefore, to repeat the conclusion drawn above, it is not the facts which are memorized in the course of sensory experience which are the basis of the cognitive abilities emerging in the flexible response of the brain to its environment, but the architecture and self-referential control that forces the brain to make these representations. The architecture of the brain does not reflect the knowledge stored there, but the control that provided the constraints for its acquisition.

This is the crucial difference between the neural system's organization and that of a conventional computer. In systems with this type of control architecture, acquired knowledge is put in the place where it is needed, and that place is decided by semantic evaluation via self-reference, and not by the syntactic structure of the sensory input itself.

The decisive questions are: What type of self-referential control is required to allow the self-organization of a consistent relational knowledge representation?, and How could we implement such capabilities into our artificial neural systems?

Self-referential architecture and control must exist from the beginning of the self-organization process. Hence, they should be encoded genetically, and their key elements should be expressed by the phylogenetically oldest structures of all brains with a cortex. Next, the bootstrapping of a relational knowledge representation enforced by such self-referential control could plausibly be explained as being guided by previously acquired knowledge that may serve as an initial hypothesis, even if not a very sophisticated one, but some a priori knowledge must be in the system to initialize the process. Also, since this a priori knowledge is a prerequisite for integrating sensory information into a consistent internal representation, it must be there before any sensory experience can shape the brain. We conclude that not only the control architecture, but also the initial knowledge must be genetically encoded.

Below we propose a first approach that could provide the orientation for the development of non-von Neumann, brain-like computation. In the next section we propose a way to start the self-organization process, and also that a set of intrinsic values (as expressed in the emotional system) probably represents the necessary a priori knowledge. The following section describes a cortical architecture that might support the self-referential control of formation and recall of knowledge representation. We show that this cortical architecture performs sensory recognition in an analysis-by-synthesis way, which is what experimental data strongly suggest for visual recognition.[4]

Self-organization of semantic constraints for knowledge representation guided by a value system

The neocortex developed top-down

The failure of top-down designed systems of early AI to ensure flexibility and robustness when confronted with real-world problems in recognition and control encouraged the bias toward purely bottom-up schemes of self-organization of knowledge representation, as in D. Marr's approach for vision, or the subsumption architecture in robotics proposed by R. Brooks. However, when starting at the lowest sensory and behavioral level, and developing increasingly more complex sensory and behavioral repertoires, the direction of the self-organization of the architecture must be defined by trial and error. Neither statistically based learning, nor stochastic learning (e.g., genetic algorithms or evolutionary optimization rules) are sufficiently effective for such a purpose, and especially not for handling more complex problems in sensory categorization. To get more than simple reactive behavior, the designer has to specify the jobs, and the conditions for performing the jobs when adding a higher level to the systems hierarchy. The philosophy behind those approaches is the assumption that the brain has developed bottom-up, and that the emerging higher levels of the system's architecture subsequently enslaved the lower ones.

However, recent insights into the phylogenetic development of brains with a cortex suggest that there is a top-down process.[3] The phylogenetic development of the neocortex started from two prime moieties: the paleocortical moiety (temporal pole) tied to holistic sensory analysis, and the archicortical moiety (hippocampal cortex) tied to processing the effective behavior. From these moieties, all other cortical areas developed step by step, with interconnections between areas of the same level of the emerging hierarchy.[5] Hence, for both sensory analysis and behavior generation, the highest level of the hierarchy dealing with the most general evaluation of the sensory situation from the subject's position were in place first, and controlled the subsequent correlated development of lower levels of analysis and behavior generation. The phylogenetically youngest levels (last in development) are the primary sensory and motor areas. Only the oldest part of the cortex (the origin of its phylogenetic development) is reciprocally connected to the amygdala (AM), which is an old subcortical part of the limbic system related to emotion, and which has been shown to trigger elementary behavioral categories (e.g., fear conditioning, arousal, and escape behavior). Based on experimental evidence, we propose that this top-down development proceeded under the control of the AM and related structures, and that they constitute a value system for the subject (the living creature). Furthermore, we also conclude that in the primate neocortex, top-down control should still be dominant in defining the direction necessary for understanding a sensory situation, and setting the stage for the interpretation of the sensory input by the lower-level systems (Fig. 1).

General scheme for self-organization of knowledge representation

With the development of systems serving an ever more detailed analysis of sensory inputs (the hierarchy of which developed in a top-down direction, see above), the sensory input to the cortex may have been moved downward together with the emerging lower hierarchical levels to the current primary sensory areas, while the "old" direct input to the oldest part of the sensory analysis is still present via the AM (Fig. 2). In the well-investigated ventral visual pathway of the cat, there are still direct inputs from the sensory thalamus to the inferotemporal (IT) cortex and all lower levels of visual filtering, despite the fact that the main stream of visual input to the visual cortex has been moved to V1, the phylogenetically youngest area, which is the lowest level of the cortical hierarchy of visual processing. To start the self-organization process of sensory categorization and purposive behavior generation at all, some a priori knowledge has to be available within the system. As a candidate to store such a basic set of rules about how to react to sensory input states, we identify the AM based on recent experimental evidence.[3]

There is sufficient experimental evidence to indicate that AM neurons code a set of behaviorally important coarse sensory situations,[6–8] and that the shortcut sensory input to the highest level of unimodal sensory processing, the IT cortex, is formed before the developing cortical afferent input via the primary sensory areas reaches the same target.[9] Coarse representations in the IT cortex are formed as early as 10 days old in infant monkeys, a time when the developing afferent connections upwards in the neocortical filtering hierarchy have not yet reached the IT cortex. These representations are definitely generated by subcortical inputs and are not subject to modification by ongoing sensory experience later on when the cortical afferent pathway to the IT cortex is fully developed.[10] The candidate source for a subcortical input is direct input from the thalamus via the AM. Projections from the AM to the phylogenetically oldest part of the IT cortex are reciprocal and topographically organized,[11] but this does not apply to other sensory areas.[12] Based on these sound facts, we propose that the AM may serve as a teacher (pointer) for establishing representations of the sensory environment in the IT cortex, which could then be refined and diversified – but not basically modified – by the rich details subsequently delivered by processing along the cortical hierarchy. This means that since, for any rush of sensory input, the established coarse representation in the IT cortex is always activated first before the afferent wave of filtered details arrives at the IT cortex via the cortical hierarchy, these representations can actually guide the afferent filtering to this target by sending feedback to lower levels of sensory filtering which meets the bottom-up

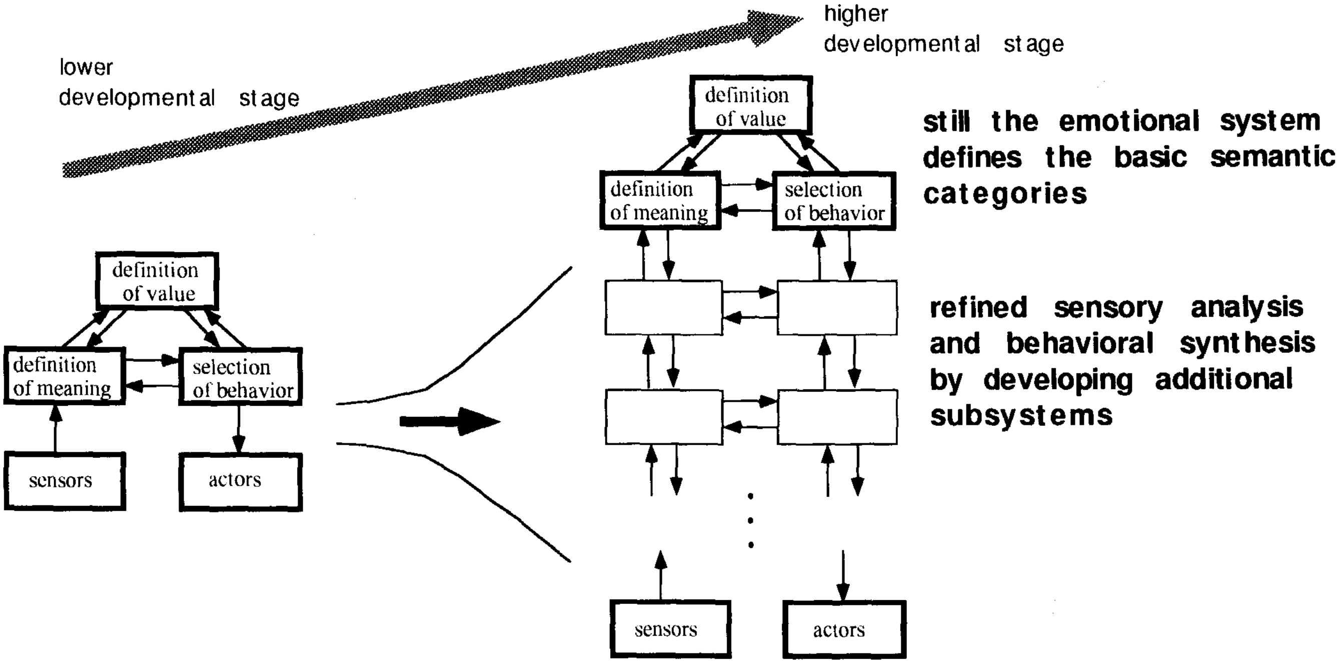
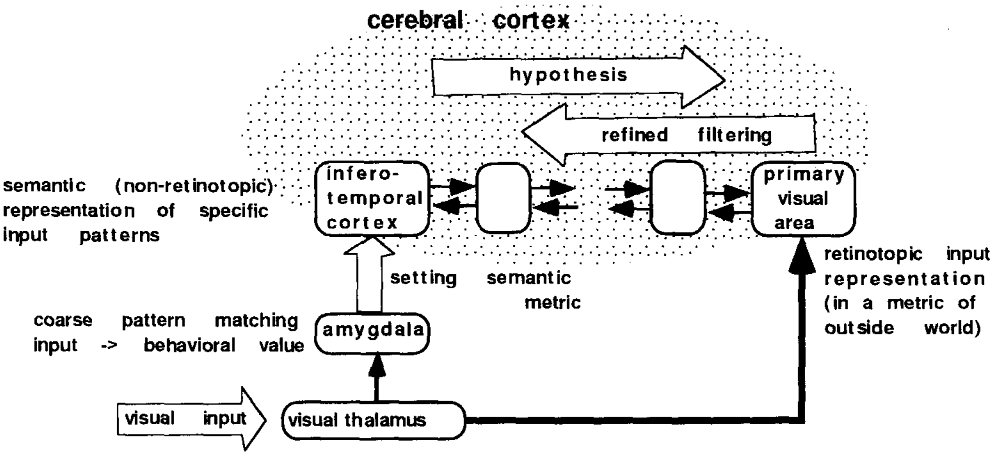


Fig. 1 According to our hypothesis, the phylogenetic development of the cortex is a top-down process, and a basic set of coarse rules on sensory–behavioral pattern evaluation mediated by the emotional system constitutes the origin of the development of knowledge representation. The subsequent development of both a more refined sensory analysis and a behavioral synthesis *in between* the value system that still remains at the top of the processing hierarchy and the sensor/actor level utilizes the semantic metric provided by the emotional system as the control for the emergence of refined constraints at any subsequent lower level of knowledge representation

Fig. 2 The shortcut via the amygdala activates the highest level of unimodal sensory processing, the inferotemporal (IT) cortex, before the (both phylogenetically and ontogenetically) developing cortical hierarchy of elaborate filtering can give inputs to the IT cortex


cerebral cortex
hypothesis
refined filtering
semantic (non-retinotopic) representation of specific input patterns
infero-temporal cortex
primary visual area
retinotopic input representation (in a metric of outside world)
setting semantic metric
coarse pattern matching input -> behavioral value
amygdala
visual input
visual thalamus

stream and takes control of it. The experimental evidence cited above also implies that a basic set of such patterns and related behavior should be genetically imprinted in the AM. We conclude that these rules could serve to start building a knowledge representation to explore the environment and define the coarse constraints within which the system can self-organize depending on its sensory experience.

Therefore, the AM seems to be the site that defines the semantic metric at the IT cortex for cortically preprocessed sensory patterns. The term "semantic metric" is used here to describe the relation between a certain situation in the environment and the activation of an appropriate behavior. The AM as a key element of the emotional system is proposed to define the fundamental categorization of sensory situations (good or dangerous, pleasant or annoying, etc.), which then is diversified during the ongoing experience of the living creature into a more and more elaborate hierarchy of sensory representation and a suitable behavioral repertoire.

Moreover, since input to the IT cortex via the shortcut pathway thalamus → AM → IT is naturally much faster than via refined cortical filtering, this shortcut may serve not only as a semantic pointer for learning, but also for any recognition it may preactivate a coarse representation in the IT cortex as a hypothesis about new sensory input. From the evidence assembled by LeDoux,[13] we further propose that this activated hypothesis may also define a semantic evaluation of the sensory input, and set the semantic constraints top-down for the detailed processing of the same sensory signals when they later arrive bottom-up in the lower levels of the cortex.

One-shot learning: now-print command by the amygdala

Having a teaching input from the AM that defines the site at the top of the hierarchy of sensory representation where the input pattern arriving at the lowest level has to be connected is a proper prerequisite for the self-organization of knowledge representation. However, sensory situations are highly variable, and do not repeat sufficiently often to facilitate statistical learning. For some sensory situations the subject should not need a second chance to recognize it (danger, etc.). Learning to represent such short manifestations and transient states in a heterarchy is not an easy problem.

We propose the following hypothesis which is a possible scheme for one-shot learning in the brain.[3] In cases of strong emotional activation (arousal), some of the nuclei of the AM and further subcortical nuclei triggers the nonspecific activation and supply of a transmitter that supports the formation of LTP (long-term potentiation) in cortical visual and auditory areas.[13-15] Therefore, the part of the AM which is topographically connected to the oldest part of the IT cortex may define where a certain sensory situation has to be represented within the internal metric of semantic categorization, while other nuclei of the AM may serve to distribute a "now-print command" for one-shot learning nonspecifically into all cortical areas in cases of a sensory situation that caused arousal. This is either because the subject wants to learn that situation and raises its arousal by attention, or because the subject failed to respond properly to a sensory input since its internal representation of this situation has not yet been sufficient to elicit the required behavior, and disappointment or fear provide the arousal (Fig. 3).

Existence of a value system is a prerequisite for any self-organization

The proposed model hypothesis of how a complex system can start the self-organization of knowledge representation implies that emotions are probably not only some "side-effect" of brain processing, but are actually the crucial basic set of coarse pattern matching {sensory input pattern → behavioral archetype} that defines the coarse semantic metric for the evaluation of sensory situations, and supports its refinement with sensory experience. Our conclusion is that for any development of an autonomous system, regardless what class of performance is selected, a value system has to be designed first, setting the coarse potential characteristics of behavior within which the system can self-organize. Refined sensory analysis should develop in accordance with the need for refinement of behavior based on this coarse semantic metric. Self-organization of sensory categorization cannot be understood in terms of a bottom-up philosophy.

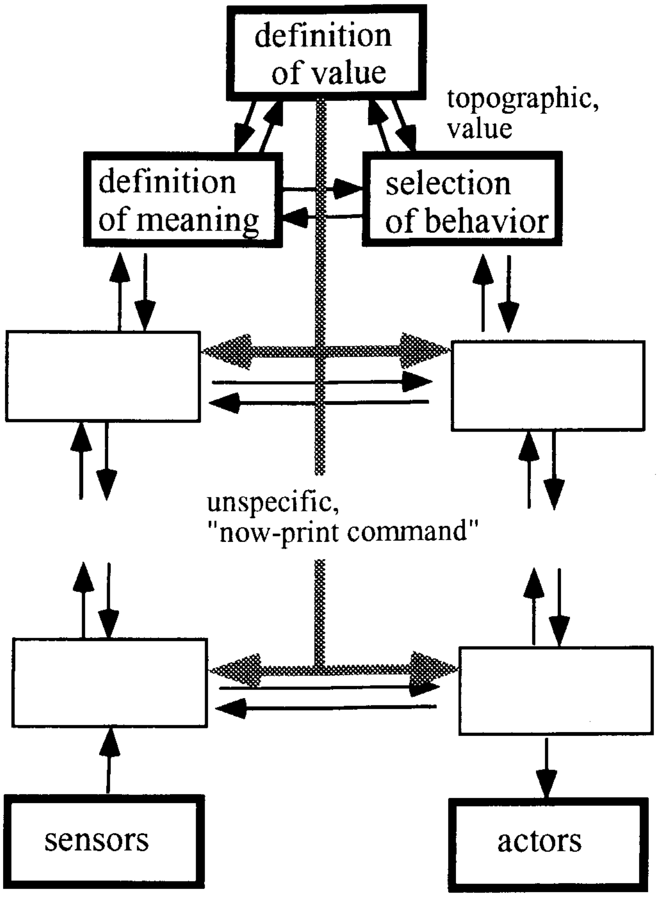


Fig. 3 It is proposed that a now-print command is sent from the amygdala to all subsystems of the heterarchy of sensory analysis and behavioral synthesis

The primacy of top-down control for relating sensory events to meaningful behavior, and the existence of "supervisory control" by a value system (like the emotional system in living creatures) to coarsely set the semantic metric and watch its refinement through ongoing experience to keep it consistent are two key requirements for creating artificial systems that are expected to show the flexibility and robustness we admire when considering the performance of living creatures' brains.

Robust and flexible recognition: analysis by synthesis

Hypotheses constitute dynamic constraints for top-down control of sensory interpretation

The top-down bias of control that guides the refinement of the system's architecture during phylogenetic development, as shown above, should later also dominate recall and the learning of facts within that architecture in a similar manner. Recognition of complex sensory situations must be an active process involving both bottom-up and top-down processing. The first stage in visual recognition seems to be a coarse holistic classification of the input,[16] which is subsequently identified in detail.[17] Accumulated experimental evidence in psychophysics supports the assumption that conscious experience involves an active modelling rather than a passive perception of incoming information.[18]

A self-referential system will respond to a sensory input by trying to find structures and regularities already represented in the system, and to assemble the known aspects into the largest possible contextual frame. Aspects already understood should be removed from the data stream to make it easier to detect new regularities.[4,19] For this purpose, the neocortex has to construct a working model of the environment to compare it with the sensory input, remove those parts that match known items, and analyze the residue. Therefore, any recognition must be some kind of generation of internal hypotheses and verifying them by eliminating the respective aspects from the sensory input. We conclude that interpretation of sensory input seems to be not only the activation of a higher-level internal representation, but also a process of re-creation of the internal representational architecture which best relates to the configuration of the sensory input, and which serves as an emerging dynamic constraint for the further refinement of the recognition.

We now consider image recognition, since the visual system is by far the best investigated neural system in the cortex. Only for vision do we have sufficient experimental data to verify our hypothesis, and this comes from morphology, physiology, and psychophysiology. We also know from more than 30 years of computer vision research which problems are the most difficult.

It is widely agreed that a proper global hypothesis about sensory input will speed up an image interpretation process, and avoid the problem of combinatorial explosion which is inherent in the bottom-up local filtering approach. However, in the case of real-world problems, to get the proper holistic hypothesis at the top of a hierarchic recognition system in order to control the filtering at lower levels, the recognition problem should already have been solved by the forward filtering process, i.e., a hen-and-egg problem.

From the experimental evidence, we hypothesize that the brain avoids that dilemma by a parallel but coarse input to higher hierarchic levels of visual representation in the ventral pathway via the amygdala, triggering a coarse holistic decision on the behavioral meaning of the sensory input (see Fig. 2). We propose that this shortcut input of raw visual data to the highest level of visual representation activates a coarse hypothesis which feeds forward an expectation to the lower levels that sets the dynamic constraints for the refinement of this holistic, but coarse, hypothesis.

Neurons in the amygdala show visual stimulus-selective responses, but they have a stronger tendency to respond to a specific category of stimuli; some responded only to a limited category of stimuli, such as an angry human face, a particular person, a dry food, the threatening face of a monkey, etc.[8] After that coarse setting of a semantic metric, the most salient aspects of the input description that are transmitted fastest along the neocortical filtering hierarchy may then specify a definite, but still coarse, holistic description at the IT cortex within about 100 ms in a forward processing based on the first spike.[20,21] The rapid activation of a holistic initial hypothesis on the raw input data at the top level for a pattern representation of the visual input (for a simple object this may be the IT cortex, for a scene the perirhinal cortex may be the best location) can impose content dependent constraints on the afferent input description of the lower levels of the neocortical hierarchy. We assume that the refinement of the description by the lower level is only possible if the higher level has a sufficiently entrained hypothesis to feed back a definite support to allow the activation of related details of the description, limit the breadth of search, and prevent a combinatorial explosion of possible alternatives in local filtering.

This assumption is supported by the fact that object representations in the IT cortex are activated within 5 ms after the arrival of the afferent information and do not change,[21] while at V1 and extrastriate visual areas the ensemble of activated feature detectors may vary within a 200-ms window of sustained cortical activation after the arrival of the afferent information.[22,23] We propose that the, 200-ms response duration reflects the refinement of the internal hypothesis in a process of hypothetical reasoning, and that the subsequent activation of the less salient parts of the input description is biased under feedback from the more global evaluation at the higher level,

Generation and verification of hypotheses require complex elementary processing nodes

To implement the self-referential control for recall and learning, the prediction generated by the activated hypothesis must be compared with the true forward input description (see above). The comparison is made at any processing node in the system's architecture, since the system has to decide exactly where and by what difference the knowledge has to be updated to allow a correct prediction of the input in question the next time it is received. This difference must be defined both explicitly and instantly. Usually, there is no chance of keeping the input constant for enough time to perform any kind of backpropagation, and rapid modification of the hypothesis is a key point for any learning without a teacher under real-world conditions. For comparison, the input description and the merging hypothesis should be separately represented at any elementary processing node in the system. Then, any of those nodes has to decide locally – according to its comparison of its state relative to the state of the global system reflected by the feedforward and feedback inputs – which of its possible states it should instantiate, and which of its representations has to be updated in the process of learning. The elementary processing node of networks capable of this kind of hypothetical reasoning must have a sufficiently complex organization both for local control and knowledge representation to act as the required "local agent," and this basic organization should be roughly the same across all nodes of the system regardless of their position in the hierarchy.

We propose that the neocortical columnar module should be the required complex elementary processing node. Our system architecture consists of a hierarchy of homogeneous nets, the elementary processing nodes of which are modular units.

Functionally, the system is composed of two parts that are locally connected at any modular unit: the hypotheses are generated from afferent (sensory) signal flow in a forward processing hierarchy, and the hypotheses are firmly established, verified, and updated in a recurrently connected upper system part which serves to re-create that hierarchy of dynamically linked modular units that are expected to be active if the emerging hypothesis in question is true. Verification of the hypothesis is performed at any modular unit of the network system by the feedback prediction. Those nodes (modular units) of the artificial neural system which have a forward-activated hypothesis consistent with the predictive feedback can strongly entrain this hypothesis, while switching off the afferent signal filters that generated the hypothesis. The local hypothesis is kept alive by mutual excitation of all local hypotheses that form a consistent global one within the recurrently connected hypothesis system part (Fig. 4), otherwise it cannot serve as the "context" for evaluating subsequent inputs. We have demonstrated the generation of an initial hypothesis, and its subsequent refinement under increasing top-down control of the emerging global hypothesis in simulating an object recognition in this architecture.[24]

Why the cortex needs to have a modular, multilayered organization

A system like the cortex which is in continuous interaction with the environment, and which responds to any input by a prediction based on its previously acquired knowledge, cannot have simple processing nodes. It needs a kind of firmware-level organization to enable the local processing node to adjust its behavior within the organization, as proposed in the previous section. If we flatten the cortex we find an almost homogeneous layer composed of complex nodes and columnar units, the basic internal organization of which does not depend on the information represented there. We propose that this unitary architecture of columnar units represents not the structure of the knowledge stored there, but the control that forces the system to make those representations.

This firmware may represent the difference in the flexibility of processing, and especially in advanced learning, between creatures with and without a cortex. Creatures with a neocortex dominate the world because they know more about it than other animals, and it is the neocortex that is responsible for this.[4] The neocortex should be studied as the system that makes use of acquired knowledge to decompose the stream of sensory data into (for the system) meaningful pieces, and to eliminate the expected from the input description, isolating those aspects of the input not yet accounted for by the internal description generated.[4,19] This would explain the enormous selective advantage of creatures with a neocortex for speedy and reliable recognition and learning. In this paper, we proposed an architecture that can implement the required quality of processing. We tried to define a general architecture by "abducting" crucial characteristics of the neocortical type of processing and of its neural hardware organization, which, as we claim, is different from lower-level types of signal filtering and approximation which are implemented well by formal neural network architectures composed of simple formal neurons as the elementary processing nodes. Neural networks are graphical notations of the classes of algorithms which they support. Hence, different description levels for the encoding of information flow and its inherent control must also be reflected by an appropriate neural network architecture. Signal filtering and approximation in lower (subcortical) sensory processing stages are surely different in their algorithmic structures from internal simulation at different abstraction levels, which we propose for the cortical processing stages.

Conclusion

Fitting formal neural networks to represent certain aspects of the architecture and function of neural structures is

Fig. 4 A simplified scheme of bidirectional processing according to our proposed general neocortical-type architecture for self-referential control of recognition and learning. The subsystems at any hierarchic level are composed of modular units which have the same general local control for intra- and intermodular communication. See Koerner et al.[24] for details of architecture and functional implementation. *Black arrows*, reentrant connectivity within the hierarchy of feature representation; *shaded arrows*, upstream signal description in a forward hierarchy

surely not the way to understand the brain. That the brain probably does the job in a rather different way is strongly suggested by our failure to get anywhere near the performance of the brain in doing things that are very easy in everyday life, but hopelessly tricky for the current generation of computers.

To advance beyond the well known paradigms of current computational theory, we need a more functional grasp of brain-type computation. As we suggested, the controls that allow the self-organization of the system may be the same ones that govern the adaptability and flexibility we admire when considering the performance of living beings.

Physiology matters, because behind these structures is the architecture of control we need to understand. Simple architectures generate simple behavior. Without including a proper firmware organization in artificial neural systems, any such system may be able to perform a very limited job (since the relevant algorithm is represented in the designed architecture), but it will not have the robustness and flexibility we are looking for when trying to abstract some useful idea from the biological original for sensory processing and behavior control. In this paper, we have tried to show that architectures in the brain do not reflect the result of thought, i.e., a ready-made algorithm for solving a problem, but they reflect the control that generates the constraints to select a proper algorithm for a specific problem posed by the input – or to create a new one if application of the ones already acquired does not result in an adequate solution. Understanding the nature of knowledge representation in the brain first requires an understanding of the control that forces the system to make representations. To date, this intrinsic self-referential control has not been the focus of work in neurocomputing and cognitive neurobiology. We strongly suggest that a value system, a priori knowledge, and neocortical firmware (columnar architecture) are crucial elements of artificial neural systems that are expected to show both the robustness and the flexibility we admire when considering the performance of a brain.

References

1. Edelman GM (1992) Bright air, brilliant fire: on the matter of the mind. Basic Books, New York
2. Pollack JB (1993) On wings of knowledge: a review of Allan Newell's unified theories of cognition. AI 59:355–369
3. Koerner E, Koerner U, Matsumoto G (1996) Top-down selforganization of semantic constraints for knowledge representation in autonomous systems: the role of a value system for autonomous systems. Bull Electrotech Lab 60(7):1–5
4. Barlow H (1994) What is the computational goal of the neocortex? In: Koch C, Davis JL (eds) Large-scale neuronal theories of the brain. MIT, London, pp 1–22
5. Pandya DN, Yeterian EH (1990) Prefrontal cortex in relation to other cortical areas in rhesus monkey: architecture and connections. In: Uylings HBM, Van Eden CG, De Bruin JPC, Corner MA, Feenstra MPG (eds) Progress in Brain Research, vol 85, Elsevier, Amsterdam, pp 63–94
6. Spiegler BJ, Mishkin M (1981) Evidence for the sequential participation of the inferior temporal cortex and amygdala in the acquisition of stimulus–reward associations. Behav Brain Res 3:303–317
7. Allman J, Brothers L (1994) Faces, fear and amygdala. Nature 372:613–614
8. Adolphs R, Tranel D, Damasio H, Damasio A (1994) Impaired recognition of emotion in facial expressions following damage to the human amygdala. Nature 372:669–672
9. Romanski LM, LeDoux JE (1992) Equipotentiality of thalamo-amygdala and thalamo-cortico-amygdala circuits in auditory fear conditioning. J Neurosci 12:4501–4509
10. Rodman HR (1994) Development of inferotemporal cortex in the monkey. Cerebral Cortex 5:484–498
11. Amaral DG, Price JL, Pitkaenen A, Carmichael ST (1992) Anatomical organization of the primate amygdaloid complex. In: Aggleton JP (ed) The amygdala: neurobiological aspects of emotion, memory, and mental dysfunction. Wiley-Liss, New York, pp 1–66
12. Iwai E, Yukie M (1987) Amygdalofugal and amygdalopetal connections with modality-specific visual cortical areas in macaques (*Macaca fuscata, M. mulatta*, and *M. fascicularis*). J Comp Neurol 261:362–387
13. LeDoux JE (1993) Emotional memory systems in the brain. Behav Brain Res 58:69–79
14. Steele GE, Weller RE (1993) Subcortical connections of subdivisions of inferior temporal cortex in squirrel monkeys. Visual Neurosci 10:563–583
15. Weinberger NM (1993) Learning-induced changes of auditory receptive fields. Curr Opinion Neurobiol 3:570–577
16. Rosch E, Mervis CB, Gray WD, Johnson DM, Boyes-Bream P (1976) Basic objects in natural categories. Cognit Psychol 8:382–439
17. Cavanagh P (1991) What's up in top-down processing? In: Gorea A (ed) Representations of vision. Cambridge University Press, Cambridge, pp 295–304
18. Picton TW, Stuss DT (1994) Neurobiology of conscious experience. Curr Opinion Neurobiol 4:256–265
19. Koerner E (1994) Autonomous recognition and selforganization of knowledge representation in neural networks. Part 1. From structuring data to constraint generation by self-referential control. Holonics 4:3–34
20. Celebrini S, Thorpe S, Trotter Y, Imbert M (1993) Dynamics of orientation coding in area V1 of the awake primate. Visual Neurosci 10:811–825
21. Oram MW, Perrett DI (1992) Time course of neural responses discriminating different views of the face and the head. J Neurophysiol 68:70–84
22. Dinse HR (1994) A time-based approach towards cortical functions: neural mechanisms underlying dynamic aspects of information processing before and after postontogenetic plastic processes. Physica D, 75:129–150
23. Freeman RD (1995) Space and time in the central visual pathway. Proceedings 5th International Symposium on Bioelectronic and Molecular Electronic Devices, Okinawa, Japan, November 28–30, B1-01, pp 1–4
24. Koerner E, Tsujino H, Masutani T (1997) A cortical-type modular neural network for hypothetical reasoning. Neural Networks 10:791–814