Development of a Large-Scale Integrated Neurocognitive Architecture

Part 1: Conceptual Framework

James A. Reggia, Malle Tagamets, Jose Contreras-Vidal, Scott Weems,
David Jacobs, Ransom Winder, Timur Chabuk

University of Maryland*

June, 2006

TR-CS-4814, UMIACS-TR-2006-33

Abstract: The idea of creating a general purpose machine intelligence that captures many of the features of human cognition goes back at least to the earliest days of artificial intelligence and neural computation. In spite of more than a half-century of research on this issue, there is currently no existing approach to machine intelligence that comes close to providing a powerful, general-purpose human-level intelligence. However, substantial progress made during recent years in neural computation, high performance computing, neuroscience and cognitive science suggests that a renewed effort to produce a general purpose and adaptive machine intelligence is timely, likely to yield qualitatively more powerful approaches to machine intelligence than those currently existing, and certain to lead to substantial progress in cognitive science, AI and neural computation. In this report, we outline a conceptual framework for the long-term development of a large-scale machine intelligence that is based on the modular organization, dynamics and plasticity of the human brain. Some basic design principles are presented along with a review of some of the relevant existing knowledge about the neurobiological basis of cognition. Three intermediate-scale prototypes for parts of a larger system are successfully implemented, providing support for the effectiveness of several of the principles in our framework. We conclude that a human-competitive neuromorphic system for machine intelligence is a viable long-term goal, but that for the short term, substantial integration with more standard symbolic methods as well as substantial research will be needed to make this goal achievable.

*Affiliations: UMCP: Dept. of Computer Science, UMIACS, Dept. of Kinesiology, CASL; UMB: MPRC and School of Medicine

Correspondence: James A. Reggia, Dept. of Computer Science, A. V. Williams Bldg, University of Maryland, College Park, MD 20742 Email: reggia@cs.umd.edu

Acknowledgement: This work is supported by DARPA BICA Award FA87500520272.

Contents: Part 1

I. Introduction

The idea of creating a general purpose machine intelligence that captures many of the features of human cognition goes back at least to the earliest days of artificial intelligence (AI) and neural computation. In spite of more than a half-century of research on this issue, there is currently no existing approach to machine intelligence that comes close to providing a powerful, general purpose human-level intelligence. For example, while general cognitive architectures such as SOAR [Rosenblum et al, 1993] and ACT-R [Anderson et al, 2004] have been studied for many years and have been used successfully to model many specific aspects of human behavior, they have been less successful in scaling up to real world applications, and are limited by being rooted in rule-based (production system) processing. There have also been fairly general AI models of knowledge representation and inference, such as those based on first-order predicate calculus (e.g., resolution-based refutation systems) and state space search methods [Brachman & Levesque, 2004; Russell & Norvig, 2003; Sowa, 2000]. While these general AI methods are widely applicable, they are sometimes called "weak methods" because they have proven less effective in applications and are computationally expensive. General purpose neural network methods such as backpropagation and self-organized feature maps have also been very successful in specific applications involving learning, such as pattern recognition, data visualization, and autonomous vehicle control, but have not been extended to many aspects of cognition. Many more symbolic, neural, and probabilistic methods have been studied in cognitive science, AI and neural computation, but the common experience seems clear: success has come in specific, focused domains, and not in the form of a general, human-like ability to solve problems and learn.

In spite of this limited success, we believe that a renewed effort to produce a general purpose and adaptive machine intelligence is timely, likely to yield qualitatively more powerful approaches to machine intelligence than those currently existing, and certain to lead to substantial research progress in cognitive science, AI and neural computation. Our optimism in this regard comes from the convergence of three advances:

- Experiments and discoveries in cognitive science and neuroscience are revealing key aspects of human memory, reasoning and learning mechanisms and their neurobiological basis, e.g., via the use of fMRI, MEG, and other functional measurements.

- Methods for constructing intermediate-scale modular neural systems have become increasingly effective and refined; the task now is to expand these systems, and to assemble and integrate them in a single framework.

- Progressively more powerful and less expensive computer hardware is becoming available, including non-standard high-performance computing architectures that make possible highly parallel computations.

These advances suggest that fundamental progress in creating a powerful, general purpose machine intelligence will come from creating a modular but integrated cognitive architecture that is inspired by human brain organization and supported by a high-performance computing platform.

When one considers the broad range of problems faced by people on a routine basis, or how people address formalized situations such as potential Decathlon and Challenge problems, it quickly becomes evident that we bring to bear a remarkable range of abilities during problem solving in an *integrated* fashion. Such integration will be essential for a situated, general-purpose machine intelligence to exhibit human competitive (or better) intelligence. In the following, it is important to recognize that this integration will need to occur along at least two related but largely orthogonal dimensions. The first dimension of integration, *behavioral tasks*, spans the broad range of tasks an intelligent agent must perform, often concurrently. These include processing multi-modal sensory input, determining the current situation from these past and present sensory events, controlling actions (motor control, arm manipulation, etc.), navigating through a changing environment, recognizing threats and opportunities, processing written and spoken natural language, pursuing self-determined goals in a rational fashion, adjusting to unexpected events, and learning from experience. The second dimension of integration, *cognitive mechanisms,* spans the underlying information processing algorithms required to support these individual behaviors/tasks. These include a variety of memory and representation mechanisms for both long-term memory (semantic, procedural, episodic, etc.), and short-term working memory, a broad range of reasoning algorithms (deductive inference, causal/explanatory or abductive inference, probabilistic pattern classification, etc.), methods for generating and/or interpreting temporal sequences of events, learning procedures at multiple levels that lead to improved performance, and top-down control mechanisms that coordinate all of these memory, reasoning, and learning methods. While there are many computational systems today that can produce a reasonable level of performance on one or a few aspects of such behavioral tasks and cognitive mechanisms, no single existing system encompasses the broad array of behaviors and algorithms listed above. Further, it is not enough just to include all of these specific abilities within a single system: they must also act together in an effective and coordinated fashion.

In this context, we believe that the *long-term goal* of creating a general-purpose machine intelligence will best be served by pursuing a computational model that is directly based on the hierarchical and modular organization, dynamics, and plasticity of the human brain, especially the neocortex and its interactions with subcortical structures. Why pursue a neuromorphic/brain-inspired architecture? One reason is that the human brain is currently the only known entity capable of exhibiting robust general intelligence in the form of integrated problem solving, language processing, planning, creative design, and learning. In short, the brain provides the only proof-of-existence that such an integrated intelligent entity is possible, and it is the only known system that encompasses information processing mechanisms sufficient to produce human-level cognition. These mechanisms are based on an underlying neural foundation that inherently supports massively parallel computations, something that is necessary for real-time operation and robustness to damage. Our judgment is that a large-scale computer system modeled after the human cerebral cortex (neocortex), the part of the human brain most closely related to problem solving and cognition, as well as closely integrated non-neocortical brain structures (thalamus, hippocampus, basal ganglia, cerebellum, etc.), is currently the best bet for a truly qualitative advance in machine intelligence over the long term. The following sections of this report present a conceptual framework in which to develop a large-scale neurocognitive architecture of the sort we envision, along with some preliminary results supporting the plausibility of this framework.

While this long-term goal provides a clear target for a successful, general-purpose machine intelligence, it raises the question of what the optimal strategy is for attaining that goal while simultaneously making progress over the short term of the next five years. One strategy would be to immediately commence implementing a large-scale neuromorphic architecture that spans all of cognition. However, our current knowledge of brain function still contains substantial gaps and uncertainties, and our understanding of how to use contemporary neural computation methods effectively to capture some aspects of cognition is also limited. Accordingly, over the next five years, we believe that the *optimal short-term strategy* is to develop a hybrid architecture that combines neurobiologically-inspired methods and cognitively-inspired methods within a unified framework. By "cognitively inspired methods", we mean more conventional symbolic and numeric methods from cognitive science and AI rather than neural computational methods. Part 2 of this report, *Design and Implementation*, provides a more detailed look at the motivations for a hybrid architecture, and then extends our framework to encompass cognitively-oriented symbolic and other methods in a unified setting.

II. Design Principles for a Neuromorphic Framework

Given that the human brain is the only known system capable of general cognition, it seems prudent to base the design of a general-purpose machine intelligence on the brain's organizational and computational principles, and this is the approach that we take here. Of course, there are widely recognized barriers to such a neurobiologically-inspired methodology, and these have deterred past work in this area. The human brain is highly complex, and we currently have an incomplete understanding of the neurobiological basis of many aspects of human cognition. Those aspects of brain function that we do understand reasonably well seem to be primarily low-level sensorimotor and reflex functions, while higher-level cognitive functions are much less understood. Further, the size and complexity of an artificial large-scale neurocognitive architecture would appear to make its implementation very difficult. We believe that these barriers can largely be overcome. The design of complex systems can be facilitated by modularity, and there is continuing steady progress in understanding the biological basis of cognition, led in part by functional imaging and modern electrophysiological methods. Existing neurocomputational models of individual brain systems show that the technology is there for many of the parts needed for a full-scale system, and the difficult challenge now is how to put those parts together effectively into a large-scale and coordinated whole. Further, contemporary high-performance electronic computing hardware and emerging non-standard computing resources indicate that the needed computational substrate is or will soon be available, and will lead to very efficient implementations by ultimately capturing the natural parallelism of neural computations at the hardware level.

In this and the following sections, we present a computational theory of human cognition that is tightly grounded in the hierarchical and modular structure, dynamics, and plasticity of neocortex and other closely coupled subcortical brain structures. While the inspiration for our approach comes directly from the brain, we are *not* trying to develop a veridical model of the brain. Rather, we are extracting the fundamental organizational and processing principles of the nervous system and applying them to create a neuromorphic machine intelligence. These principles include locality of computation, massively parallel processing, hierarchical and modular structure, decentralized control, and a fundamental role for learning and adaptation. Our theory will subsequently serve as the basis for designing a large-scale integrated model of cognition founded primarily upon neurobiological principles, and this will be described in a

future Part 2 of this report. While there are many previous theories/models of brain subsystems, to our knowledge no one has ever created an architecture with the broad scope and integrated coverage of brain and cognitive functions that we are considering here. Our neurobiologically-oriented approach focuses on the critical issue of bridging the gap between neuromorphic systems and cognition.

In the rest of this section we describe and elaborate upon some of the basic design principles used in our theory. Subsequently, we present the results of two initial experiments developing intermediate-scale systems that demonstrate the feasibility of some aspects of our approach (Section III below). We then return to the broader issue of how multiple systems can be integrated and controlled in a full, large-scale neurocognitive architecture (Section IV).

A. Top-Level Overview

Our neuromorphic theory is based upon an underlying architecture having a network of hierarchically organized modules whose structure and function is directly inspired by human neocortical and subcortical organization and brain relationships to cognition. While there are important gaps in our knowledge [Uttal 2001], a great deal is currently known about the mapping of behavior in general and cognitive functions in particular to human brain regions. We thus summarized the results of our recent efforts to compile a listing of important known function-to-brain relationships as a separate report [Tinerella et al, 2006]. Cataloging these relationships between cognitive functions and brain regions proved to be an ambitious goal, given the uncertainties and even disagreements about the representation of some aspects of memory, language, and other cognitive functions in the brain. Further, the mapping is not really one-to-one in that some cognitive functions are distributed over multiple brain regions, and some regions contribute to multiple functions [Mesulam, 1990].

The basic conclusion that comes from critically examining current knowledge of human brain structure and function is that the brain's architecture can best be viewed as composed of repeating and nested functional modules. The hierarchical organization is roughly

brain → systems → areas/nuclei → local circuits → neurons.

For example, in the neocortex the local circuit modules are cortical columns whose inter-columnar connectivity is extensively (but only partially) documented in the voluminous neuroscientific data that is available. These columns are often viewed as the basic functional units of cortex [Mountcastle, 1998]. At the next level up, modules correspond to cortical areas that are interconnected by various neuroanatomical pathways and tracts. Concrete examples of histologically-distinguishable cortical areas would be the Brodmann areas 1, 2, 3, 4, … which are also labeled in ways related to their functionality (S1, M1, Wernicke's area, prefrontal eye fields, etc.) or anatomical features (supramarginal gyrus, angular gyrus, etc.). These areas can sometimes themselves be divided into subregions, e.g., primary somatosensory cortex region S1 can be viewed as partitioned into hand/arm/trunk regions. Examples of specific pathways/tracts connecting cortical areas are the arcuate fasciculus between Wernicke's and Broca's areas, and callosal connections between corresponding left and right mirror image cortical areas. At the next highest level, interconnected areas are integrated into identifiable functional systems such as the inferior temporal-frontal visual system, the spoken language system, the sensorimotor system, and so forth. Finally, these systems are integrated into a top-level network via the pathways between their components and/or overlapping components.

Implicit in this organization are feedforward, feedback, and recurrent connectivity. A similar hierarchical structure can be identified for subcortical regions such as the cerebellum, thalamus and basal ganglia.

In this context, the primary features of our framework for creating a large-scale and general-purpose neurocognitive architecture can be summarized as follows.

- Our architecture is a hierarchical network of nested and iterated modules, inspired by the neurobiological structures outlined above. These modules have spatial relationships to one another, unlike with many neural models, and this has significant implications for connectivity, functionality and learning.

- Functionality in our architecture is provided by the activation dynamics of its modules, occurring simultaneously at multiple levels of the structural hierarchy. In other words, our framework is based on a dynamical systems perspective rather than the primarily logical/symbolic approach used in many mainstream cognitive models in psychology and AI. Cognition is viewed as an emergent property of self-organizing neural processes, not something that is directly "programmed in".

- Both the structural architecture and the neurobiologically-inspired functional mechanisms are guided not only by the need for good performance but also by a drive to minimize costs (energy use, connectivity, etc.). In part, cost minimization is based upon the strength and nature of functional interactions between brain regions, and is informed by recent human functional imaging data (fMRI, PET, etc.) and electrophysiological data (EEG, MEG, etc.).

- Working memory, executive control functions, and sequential behavioral processing are represented in multiple ways in our theory, including competition between neural modules for activation that influences global control of activity (one aspect of attentional mechanisms), sustained patterns of neural activity in cortical regions, and recurrent connectivity between regions that can gate one another's activity.

- Functions of modules are largely learned, not pre-programmed, so that a module's functionality is determined in part by its location and connectivity, and in part by a "learning agenda" during which different components of the model learn independently in a prescribed, multi-stage fashion before being integrated and trained further collectively, much as occurs in human brain and childhood cognitive development.

- Finally, learning is a continuous process, implying among other things that our architecture can reorganize after damage and partially recover via dynamic reallocation of functionality.

We now turn to making this top-level perspective operational by considering some of the basic design principles of our architecture's structure, dynamics and learning mechanisms in more detail.

B. Structure

Paralleling the hierarchical organization of the human brain summarized above, i.e.,

brain → systems → areas/nuclei → local circuits.

the structural aspects of our framework are

architecture → systems → regions → cells/voxels.

For clarity, the correspondences and terminology used in the following are summarized in Table 1. An important emphasis in our approach is that conceptually one is focused on specifying an architecture more at the level of assembling regions into systems and less on specifying low-level details of neurons and their connectivity than in most past neurocomputational work. In other words, while neurocomputational models are often viewed as a "bottom-up" approach to machine intelligence, our conceptual framework takes a "top-down" view of their design.

Table 1: Terminology and Correspondences

Biological Structure	Model Structure	Interconnection Name
brain	architecture	relation
system	system	coupling
area/nucleus	region	pathway
local circuit	voxel/cell	connection
neuron	-	-

The lowest level of detail in this framework is the neural *cell* that is loosely intended to model a local volume element, or *voxel,* and its included local neural circuitry, such as a cortical column. The term "cell" here is not related to the concept of a biological cell; it refers instead to a cell of space and its contents in the same way that the term "cell" in computational systems like cellular automata refers to an atomic processing unit. A distinguishing feature of our neuromorphic architecture is that individual neurons within a cell are generally not explicitly represented – the atomic elements used in our model are the cells/voxels and their interconnections. This differs from most neurocomputational models where neurons (or even smaller elements such as dendritic compartments) are explicitly viewed as the atomic units of computation. Our position is that if one wants to develop a large-scale integrated machine intelligence, individual neurons (dendritic trees, molecular structures, etc.) provide too low a level of abstraction at which to start. Some implications of this choice are that the functionality of local neural circuits must be captured in the internal dynamics of a voxel/cell, and that the dynamics of a cell does not in general match that of an individual neuron. Cells communicate locally in our model via weighted *connections* and have one or more internal activation levels.

Cells in our framework are assembled into regularly structured *regions* that roughly correspond to areas in the cortex, or subcortical neuroanatomic structures such as nuclei in the thalamus or basal ganglia. As illustrated in Figure 1, these cellular arrays or regions have an explicit spatial organization. In the following, we will generally view these arrays as being 2D structures, but there is no reason that other dimensionalities (1D, 3D, etc.) cannot be used, and all that we say below applies equally well in such situations. The regular repetitive cells in arrays provide a simple, uniform base upon which to construct an architecture and define its

computational properties, and this uniformity will facilitate a hardware implementation over the long term should that become appropriate. Some implications of an explicit spatial representation are that real-valued distance metrics are relevant, that intra-array connectivity can be an explicit function of geometric (versus topological) distances, and that self-organizing topographic and feature map formation becomes an important functional issue. As illustrated in Figure 1, a region receives inputs and sends outputs to other regions via *pathways*, collections of individual inter-cell connections analogous to identifiable tracts in the central nervous system. Regions also generally have substantial internal recurrent connectivity.

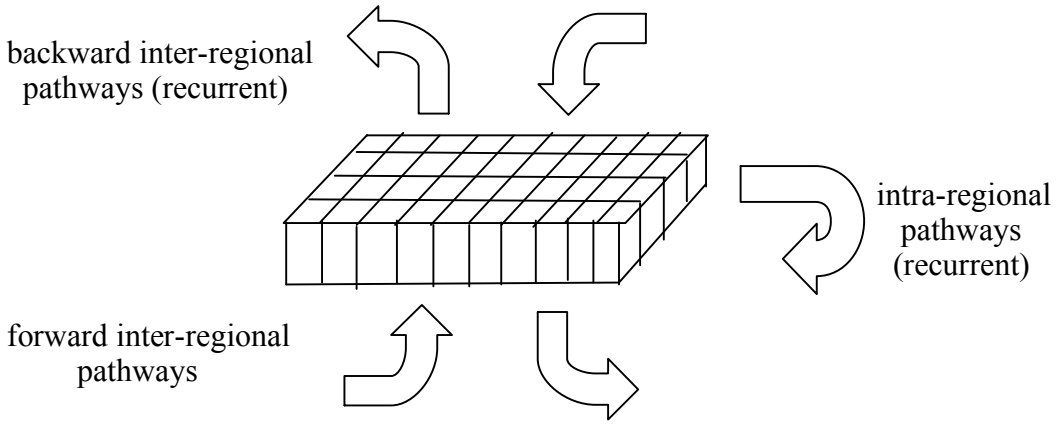


Figure 1. Schematic representation of a generic 2D *region* where each element is a cell/voxel (volume element) whose functionality captures the dynamics of local neural circuits such as those of a cortical column. Arrows indicate forward (bottom), backward (top) and internal (on the right) connectivity, which is highly recurrent. ■

A *system* in our framework is the analog of a brain system that is devoted to some class of behavioral function, such as vision, motor control, memory, language, etc. As illustrated below in Figure 2, a system is composed of a network of regions that are interconnected via pathways. The explicit spatial organization of regions means that such pathways can be specified as geometrically-meaningful projections or mappings of one region onto another, rather than connection-by-connection. Further, each region like those pictured in Figure 2, viewed as a whole, has one or more associated activation levels distinct from those of its component cells, and each pathway has one or more associated weights distinct from those of its constituent connections. These activation values and weights serve as part of the top-level control mechanism, as will be explained below.

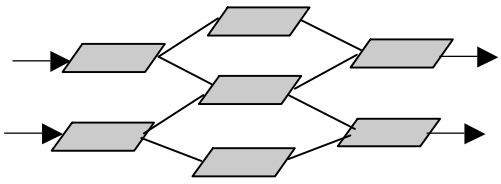


Figure 2. A *system* within our architecture is a network of interconnected, functionally-related regions like that shown in Figure 1, seven of which are shown here. The regions are connected via bidirectional pathways (lines without arrowheads). ■

Finally, in an analogous fashion, the resultant neurocognitive *architecture* can be viewed as composed of a network of interconnected systems that provide the structural basis of the entire model. Each individual system, viewed as a whole, may have one or more associated activation values, distinct from those of their component regions, and one or more associated weights on their interconnected *couplings* that are distinct from the weights on their inter-regional pathways. While it is beyond the scope of this report, "social interactions" between multiple neurocognitive architectures can be influenced by *relations* as indicated in Table 1, supporting the formation of social networks.

C. Dynamics

At the level of cells, activation dynamics in our model incorporate many features of methods used in contemporary neural networks, and these features are not intended as innovations of this work. Each cell has one or more real-valued activation levels that are repeatedly updated based on incoming activity from other cells in their local neighborhood, or from other regions. Activation rules that govern the updating of a cell's activity are generally expressed as non-linear differential equations, and the behavior of a cell is viewed as a dynamical system having various attractor states. The cells forming a region act collectively, producing region-level attractor states that emerge from the numerous non-linear interactions between activated cells in that region, something that can be viewed as an analog of the "mass action" occurring in the nervous system. Cognitively-relevant information is thus encoded in a region using a distributed representation/encoding (coarse coding). Put otherwise, working memory is represented by sustained activity patterns across regions, where these patterns are the attractor states. Long-term memory is represented in inter-cell connection weight values, or intra-cell parameter values.

In addition to these fairly conventional computational mechanisms, our approach encompasses a number of innovations, or at least non-standard features. One fundamental organizing principle that distinguishes our theory is that neural architectures should be based not only on obtaining good performance, but just as importantly on minimization of costs such as energy use and structural connectivity. Such cost minimization, or parsimony, appears to be an important constraint on brain evolution [Gibbons, 1998], has proven very effective in some of our past work in explaining neocortical dynamics and specialization [Reggia et al, 1992; Shkuro et al 2003], and creates neural architectures that scale up in size better if eventually implemented in hardware. We now describe two ways that this basic parsimony principle is incorporated within our framework.

First, as the cells/voxels that are atomic elements of our model are not neurons, they can exhibit behaviors that are quite different from typical biological neurons in past neural models. For example, a cell in our framework may retain specific details of previously seen input patterns and base its output on such patterns in novel ways. This allows one to capture within a cell's dynamics the functionality of neural circuitry used in some past models of working memory [Tagamets & Horwitz, 1998]. Most relevant here is that a cell/voxel may also exhibit competitive activation dynamics [Reggia et al, 1992] that can substantially reduce intra-regional recurrent connectivity. For example, neocortex has long been recognized to exhibit a Mexican Hat pattern of activation due to a localized stimulus: a region of evoked activity is surrounded by an annulus of suppressed activity. This is captured in many computational models of cortex by intra-region connections: relatively few short-range excitatory connections

and relatively many longer range inhibitory connections. In our model, cells/voxels can competitively distribute their activity, something that is implausible for an individual neuron but is perfectly legitimate for a voxel (neural circuitry) to do. The result is that a Mexican Hat activity pattern is produced without the need for numerous inhibitory connections, greatly simplifying intra-regional circuitry. Distributing neural activity in this competitive fashion implies synaptic connections whose strengths not only change slowly during learning as in most neural models, but also change very rapidly to direct the spread of activity. Such "fast weights" have become increasingly plausible in recent years with the growing evidence that rapid (on the order of milliseconds) changes in biological synaptic strengths are a common and important computational mechanism in the brain [Abbott & Regehr, 2004].

A second, more cognitively interesting use of competitive dynamics within our framework is at the higher level of regions and their interconnecting pathways (Figure 2). As noted earlier, regions and pathways also have activation levels and weights associated with them that are distinct from those of their components. The higher-level activation values associated with regions can either be derived from the activations of the region's component cells, such as a time-averaged mean activity level, or imposed by other regions or external entities as top-down control information. Similarly, each inter-region pathway has one or more associated weights distinct from the weights on the individual inter-cell connections that compose the pathway. For example, one weight associated with a pathway is its *gain* indicating the magnitude of its inter-regional effects; dynamically adjusting such a gain alters effective network structure. The key idea is that, in integrating regions into systems, and systems into an architecture, these high-level activations/weights allow regions to turn one another on/off, and for one region to "gate" (enable/disable) the flow of activity between other regions. Such gating is believed to occur, for example, between cortical and subcortical brain regions during motor control and during performance of working memory tasks. We view these high-level inter-regional effects as the basis for implementing competitive and cooperative effects between regions, just as they occur between cells within a region, and for parsimoniously distributing activity. In this way, there is a distributed global control of the flow of activity throughout the overall architecture, and this control process forms one aspect of attentional mechanisms. While we have previously used competitive activity distribution between columns as the basis of a theory of neocortical dynamics [Reggia et al, 1992], and also as a control mechanism for non-neurobiological cognitive/AI models of print-to-sound transformation and diagnostic problem solving, this will be the first time that it will be used as part of an attention mechanism based on thalamocortical interactions.

Finally, for a situated cognitive architecture to function effectively, it must be able to process events as they unfold sequentially in time. Processing of temporal/sequential events is supported within our framework by recurrent intra-region connections and recurrent inter-region pathways. This recurrent connectivity with its inherent delays leads to attractor states that are generally not fixed points, i.e., to quasi-periodic and chaotic attractors, and to switching between such attractor states as the basic mechanism for cognitive operations over time.

D. Learning: A Developmental Approach

The ability to learn is a critical aspect of human intelligence and thus a fundamental part of our theoretical framework. The needs in this area are extensive. Learning is required across a range of levels, from low-level sensorimotor processing and control through high-level cognitive functions and executive decision making, and across a range of contexts (supervised, reinforcement, and unsupervised scenarios) and modalities. We address these needs by integrating multiple learning algorithms in our framework, some of which are off-the-shelf methods and others of which are innovations that address specific needs. These algorithms act at different levels of our structural hierarchy, from individual cells and their connections to entire regions and their inter-regional pathways. The functional operations acquired by an initially generic region during learning are based on that region's unique position in an architecture's network as well as its intrinsic properties, just as is postulated to occur for functional localization in the cerebral cortex [Passingham et al, 2002]. As we explain below, the modular nature of our architecture allows learning to proceed in a multi-stage, incremental fashion that we refer to as a *learning agenda*. This approach is inspired by human neurobiological and cognitive developmental stages, and makes the training of a large scale cognitive system tractable. We now consider some of the details of the learning mechanisms, starting with the most conventional.

As with activation dynamics, at the level of cells and their connections we incorporate a variety of existing learning methods within our framework that are not intended as innovations of this work. These include unsupervised methods such as Hebbian learning, reinforcement methods such as temporal difference learning [Sutton & Barto, 1998], and supervised methods such as contemporary versions of error backpropagation like RPROP [Reidmiller & Braun, 1993] and methods for learning with recurrent networks. However, even at this lowest level we adopt some non-standard methods to address the broad range of learning methods needed by a general purpose machine intelligence, and give two examples of these here.

First, as noted earlier, processing temporal events is a fundamental requirement for a situated autonomous/semi-autonomous cognitive agent. At a minimum, the ability to learn to both recognize and generate temporal sequences is needed. There are a variety of effective supervised learning methods for temporal sequences, but unsupervised methods for distributed representations are much less developed. For the latter, recent discoveries of temporally asymmetric Hebbian learning in neocortex and other brain structures [Bi and Poo, 2001; Markram et al, 1997] have led to suggestions that this may be an important mechanism for learning temporal sequences [Rao and Sejnowski, 2000]. We recently created a specific implementation of temporally-asymmetric Hebbian learning and used it successfully with recurrent neural networks to "discover" an effective distributed representation for different temporal sequences of phonemes representing words [Schulz & Reggia, 2004]. This approach should generalize to analogous sequential tasks (e.g., learning to recognize an opponent's strategies). Our more recent experiences integrating and adopting sequence processing methods in larger, system-level models are encouraging, as we describe in Section IIIA below.

A second non-standard approach to learning at the level of cells that is incorporated into our framework is the learning of activation dynamics. Most neural network learning methods assume an a priori, fixed activation dynamics that is at least loosely modeled after how individual neurons process information, with learning occurring primarily by changing weights on connections. However, since the atomic units in our framework (cells/voxels) are not

restricted to behave like individual neurons as long as they retain local information processing, our approach permits the activation function of cells (as well as connection weights) to be learned. For example, cells can learn novel ways to combine their individual inputs (rather than just as a linearly weighted sum), internal parameter values, whether to distribute their output activity in the usual non-competitive fashion or in a competitive fashion, and so forth. We have previously used this approach successfully in simple networks [Grundstrom & et al, 1996], and believe that it will generalize readily to the neural architecture described here, greatly increasing the flexibility and effectiveness of learning.

Learning at higher levels in the structural hierarchy, such as learning activity and weight values at the level of regions and their pathways, is largely unexplored in past neurocomputational systems. We believe that reinforcement learning methods are very promising at this level. In addition, fMRI data may provide useful guidance for setting pathway parameters such as the functional connectivity between regions. By *functional connectivity*, as opposed to structural connectivity, we mean the dynamic relationships between regions that exist during cognitive tasks. These relationships are associated with the covariance of regional activities as observed during functional imaging and often represented using structural equation modeling. Our initial attempts to guide task-specific pathway gain learning using fMRI data have been encouraging and are described below in Section IIIB.

Finally, from a more global perspective, our framework recognizes that one cannot assemble a large-scale neurocognitive system all at once and simultaneously learn everything that is needed in one step. Thus, a central aspect of our methodology is that it incorporates a *developmental approach* that leverages our framework's inherently modular architecture. This is inspired by developmental processes shaping the human brain during childhood. Different brain systems have distinct developmental time courses, with synaptogenesis and synaptic elimination reaching peaks at different ages for different systems [Neville and Bavelier, 2000]. Behaviorally, children go through a sequence of stages in which psychological competencies appear in a fairly typical order, and these stages are loosely correlated with developmental changes in the brain [Kagan and Baird, 2004]. For example, children learn to recognize some aspects of phonemes of their native language well before they learn to produce spoken phonemes [Vouloumanos and Werker, 2004]. Our intent is not to model accurately the details of human childhood development, but to use this natural multi-stage process as a guide in assembling a large-scale neurocognitive architecture.

In our framework, the practical implementation of a developmental approach takes the form of a *learning agenda*. A learning agenda specifies a plan or procedure for the incremental construction and training of parts of a system, their assembly and further training, and so forth, until a complete and fully trained architecture is achieved. In a sense, this is a specific instantiation of a long-standing philosophy of how to go about creating a general machine intelligence that can be dated back to the early days of AI [Turing, 1950] and that continues to have its advocates today. This philosophy argues that one should initially aim to produce a machine intelligence with the abilities of a young child, and then allow such an artifact to learn additional abilities. Our initial experiment in using a simple multi-stage learning agenda as part of building an intermediate-scale neuromorphic architecture is encouraging, and we turn to that in the following section.

III. Intermediate-Scale Prototype Experiments

Here we describe the results of three exploratory computational experiments that we have just completed to establish the plausibility of some of the concepts introduced above and to examine their implications. The general hypothesis being tested in these experiments is that one can efficiently create intermediate-scale neuromorphic models based on the principles of our theoretical framework that not only perform nontrivial cognitive tasks but also are consistent with existing neuroscientific data obtained in intact and brain damaged individuals. Each of the three intermediate-scale models developed in this work roughly corresponds to what we have referred to as a "system" above: they are a collection of functionally related regions that are formed into a network by inter-regional pathways.

A. Associative Word Learning Model

Our first experiment focused on the assembly of an *associative word learning model***.** This model is based on the "classic" and highly influential Wernicke-Lichtheim-Geschwind (WLG) theory of language processing that is widely known in neuropsychology and clinical neurology [Brown & Hagoort, 1999]. Accordingly, it differs from most past computational models related to language in that past models have largely simulated the cognitive processes involved and generally did not intend to represent the underlying cortical regions and their interregional connections explicitly. We use the model to address two specific questions. First, beginning with an untrained model consisting of interconnected neocortical regions spanning both cerebral hemispheres, is it possible to create a left-lateralized computer simulation of the primary regions and pathways of the WLG theory that can learn to recognize heard words that are object names, repeat them, and associate them with the appropriate objects? Second, assuming that one can successfully implement a computational simulation of the WLG theory, to what extent does it behave in ways reminiscent of the classic aphasia syndromes following focal cortical lesions?

The basic functional-anatomic framework of the WLG theory is illustrated in Figure 3. Broca's area (BA) and Wernicke's area (WA) are the most prominent language processing centers in almost all theoretical models of language processing, including the WLG theory. Although the relative importance of these areas to different language functions is a long-standing question, there is no doubt they each serve separate roles. WA receives input from primary auditory cortex (A1), among other areas. The language deficit known as Wernicke's aphasia is closely associated with WA loss, and is characterized by impaired comprehension and repetition ability, but with some spared ability to produce fluent, but often meaningless, verbal utterances. In contrast, BA is believed to be responsible for more expressive aspects of language, playing an important role in grammatical speech production. Destruction of BA along with surrounding frontal cortex is associated with Broca's aphasia, an impaired ability to produce linguistic output despite retained comprehension. The arcuate fasciculus (AF) is the pathway connecting WA and BA. There are some linguistically impaired patients, said to have conduction aphasia, who are capable of both comprehending and producing speech, but incapable of repeating heard words. Historically the proposed underlying deficit in these individuals is blockage of the AF 's "conduction" of information from WA to BA. Currently it is believed that communication between these areas is mediated by more extensive anatomical routes, including regions such as the supramarginal gyrus in the parietal lobe. Another parietal

area, the angular gyrus, appears to play an important role as functional center of linguistic and visual object comprehension as part of a distributed semantic system. Lesions to the angular gyrus and surrounding cortex have been shown to lead primarily to multimodal comprehension deficits and have been classically associated with transcortical sensory aphasia. Finally, inferior cortical association areas have also been linked with recognizing and naming visual objects (confrontation naming). Object recognition is believed to take place through a ventral visual pathway, leading from V1/V2 to inferior temporal cortex (IT), with IT representations being more complex and not retinotopic. While classical WLG theory did not include IT, lesions along this pathway lead to loss of object recognition.

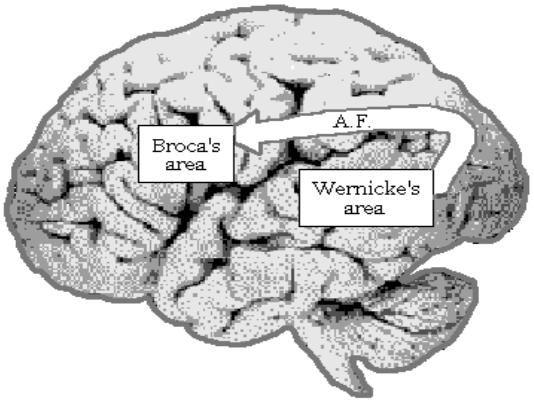


Figure 3. Central aspects of the Wernicke-Lichtheim-Geschwind theory. Wernicke's area (responsible for receptive language processing) is connected via the arcuate fasciculus (AF) to Broca's area (expressive processing). Inferior parietal regions such as the supramarginal gyrus and the angular gyrus are viewed as important tertiary association cortex but are not labeled here. The cortical areas supporting language are assumed to be only present in the dominant left hemisphere. ■

While the WLG theory is clearly inadequate to account for all language phenomena and it does not incorporate some important concepts from contemporary psycholinguistics [Caplan, 2003; Poeppel and Hickok, 2004], it is a powerful organizing heuristic for understanding the neurobiological basis of language that has influenced most contemporary theories of language. To our knowledge, no one has previously developed a neurocomputational model of language functions based on this traditional neurological theory.

The architecture that we assembled is illustrated in Figure 4 and consists of a network of perisylvian cortical regions forming the core of the WLG theory. This implementation is also informed by the results of studies over the last few decades that were not available to the founders of the WLG model, such as functional imaging. Some of these regions, their activation dynamics, and their learning procedures are inspired by earlier, simpler neural network models. Broca's area (BA) and primary motor cortex (M1) are modeled after an earlier phoneme sequence generation model [Reggia et al, 1998], while primary auditory cortex (A1) and superior temporal gyrus (our rendition of Wernicke's area (WA)) are modeled after an earlier phoneme representation model [Schulz & Reggia, 2004]. The remaining four regions (visual cortex (V1/V2), inferior temporal cortex (IT), and supramarginal gyrus (SMG), and angular gyrus (AG)) use similar methods. Each region is a cellular array in the sense defined earlier, and in addition to the inter-regional recurrent connections, there are recurrent intra-

regional connections that are not shown in Figure 4.

While the classic WLG model is generally used to describe human left hemisphere language processing pathways only, more recent research has suggested that homologous right hemisphere processing circuits may also exist and contribute to right hemisphere language processing. Experimental observations that were largely unavailable to the founders of WLG theory suggest that both hemispheres have substantial *potential* for language processing initially, with (usually left) hemispheric specialization for language arising during childhood development and language acquisition. In this context, our computational model's structure includes two initially identical hemispheres. For each left hemisphere region, there is a homologous right hemisphere region homotopically connected to it via simulated corpus callosum connections. The exception is that only a single M1 output layer is present, with connections back to both left and right hemisphere BA areas. This ensures that only a single output is produced based on the input received from pathways of both hemispheres. Thus, there are a total of 15 simulated cortical regions in the model. Except for different random initial weights, homologous left and right regions are initially identical. In effect, two identical sets of mirror image hemispheric regions are present, with one being designated the left hemisphere and the other designated the right. The challenge is for left hemisphere dominance, an important explicit feature of the WLG framework, to emerge during learning even though both hemispheres receive the same input patterns. Our intent here is not to suggest that paired left and right hemispheric regions are ultimately necessary in a neurocognitive architecture; we are only trying to determine whether current methods for guiding which functions become acquired by which regions can scale up to a model of this size and complexity.

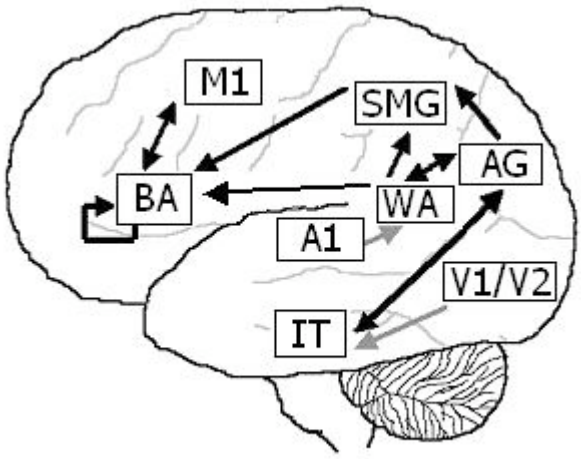


Figure 4. Diagram of the modules within the associative word learning model's left hemisphere, along with arrows representing inter-region pathways. Grey arrows indicate unsupervised learning pathways, and dark arrows indicated supervised learning pathways. Intra-regional recurrent connections exist but are not shown. Homologous regions and pathways are present in the right hemisphere but not pictured here. BA = Broca's area, WA = Wernicke's area, IT = inferior temporal cortex, SMG = supramarginal gyrus, AG = angular gyrus. ■

We limited the model to processing single word names for tractability, and because it is an important part of routine "bedside testing" in clinical neurology. Inputs to the model are fifty

"heard words" taken from the NetTalk corpus represented as temporal sequences of auditory phonemes in primary auditory cortex (A1), and images of objects from the Snodgrass-Vanderwart corpus in primary visual cortex (V1). Input of a spoken word was done by presenting its phonemes as a temporal sequence of patterns imposed on A1, as illustrated in Figure 5a. Because the model is instantiated in two hemispheres, input patterns are presented simultaneously to A1 areas in the left and right hemisphere. Each individual input phoneme is encoded as a unique distributed pattern of 34 auditory distinctive features (voicing, duration, nasality, etc.), normalized to unit length to prevent input patterns with many features from dominating learning. Visual input consists of 50 two-dimensional images, each corresponding to one of the words described above. Figure 5b gives an example. These images were taken from the line drawings of familiar objects in the normed Snodgrass and Vanderwart (1980) corpus, converted and scaled to a 50 x 50 bitmap format. The images (considered as vectors) are also normalized to unit length.

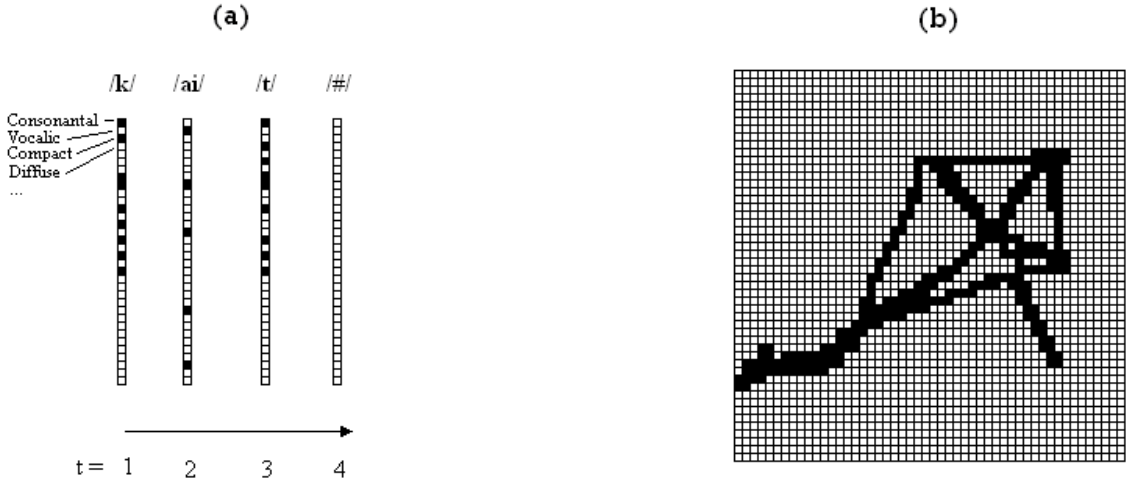


Figure 5. Coded representations of inputs in areas A1 and V1/V2. Each darkened neural element is "on" (activation value 1.0), while all others are set to "off" (0.0). **a.** In A1, the word "kite" is presented as a temporal sequence of three phoneme vectors, each represented as a vector of auditory distinctive features, followed by an all off end-of-word indicator /#/. **b.** In V1/V2, the corresponding picture is presented as a two dimensional, 2500 element image. ■

Outputs from the model are "spoken words" represented as temporal sequences of motor phonemes in primary motor cortex (M1) corresponding to the correct pronunciation of the given input word or picture. Each motor phoneme is encoded as a pattern of 20 articulatory distinctive features (using a different encoding than A1), so each neural element in M1 represents an articulatory distinctive feature.

During training, the model produces a sequence of output phonemes, ideally the same number as the number of phonemes in the target output, plus an output vector of all zeros called the stop phoneme and designated /#/. No specific functionality is assigned a priori to model cortical regions, other than that implicitly present due to their location and interconnectedness in the network. Initially, homologous cortical regions in the simulated left and right hemispheres are symmetric except for randomly assigned synaptic weights, so before training both hemispheres contribute equally to output and the model structure does not favor either left

or right hemisphere specialization.

Rather than trying to train all of the model's functions simultaneously, we adopted a learning agenda that consists of three stages. The goal of the first stage is to develop representations within the primary sensory association areas (IT and WA) using unsupervised learning. This phase corresponds to attentive viewing and listening to pictures and auditory stimuli without producing output, much as an infant experiences both visual and auditory stimulation following birth before language production occurs [Vouloumanos and Werker 2004]. Using the 50 stimuli described above as inputs, with each stimulus having both a visual (image in V1/V2) and auditory (temporal sequence of phonemes in A1) representation, learning proceeded separately for each stimulus modality. Each iteration consists of the stimulus information being set in the sensory cortical area (V1/V2 or A1), activity propagating to the sensory association area (IT or WA), and finally weights being adjusted using competitive Hebbian learning based on activity within the sensory association areas. Note that for heard words presented as a temporal sequence (e.g., for kite, /k/, /ai/, and /t/ plus stop phoneme), for any given auditory stimulus multiple inputs are received by the system, with learning occurring after each input phoneme using temporally-asymmetric Hebbian learning [Schulz & Reggia, 2004].

The goal of the second stage of training is to learn the bidirectional associations between word representations in WA and image representations in IT. This was accomplished using resilient error backpropagation [Reidmiller and Braun, 1993] where area AG served as a "hidden layer" between WA and IT. While error backpropagation is generally viewed as a form of supervised learning, note that the model is free to determine any representation (i.e., encoding) for the word-image associations that it learns in area AG.

The goal of the third stage of training is to have the model generate the correct output sequence of motor phonemes to name a seen picture or to repeat a heard word. Learning to repeat a heard word is especially challenging: generation of the output motor phonemes does not start until *all* input auditory phonemes for that word have been processed. Thus, the model must discover an internal representation for each temporal auditory sequence that persists and is adequate to generate the correct corresponding temporal sequence of motor phoneme features. Learning during this second phase occurs for all connections to and from areas SMG, BA, and M1 using resilient error backpropagation [Riedmiller and Braun 1993].

Hemispheric specialization is an important aspect of the WLG model. Past computational studies using simpler models than the one we are studying here have found that lateralized functionality can be consistently produced during learning when corresponding left and right cortical regions are asymmetric in size, excitability or synaptic plasticity [Shkuro, Glezer et al. 2000; Reggia, Goodall et al. 2001; Weems and Reggia 2004]. We elected to encourage left hemisphere specialization in our model by giving the left hemisphere a learning rate advantage throughout training (all three phases). Thus, while the two hemispheres were structurally identical and connected through a simulated corpus callosum for each area, the left hemisphere was a more rapid learner and therefore expected to become a better language processor.

We adopted the following four performance measures to assess model behavior. *Repetition* measures the percentage of correct output phonemes produced following presentation of auditory input words. *Naming* is measured in the same way as repetition, except that it reflects percent correct phonemes following visual stimuli. *Fluency* is a measure of the percentage of the expected number of phonemes that are produced following auditory input (unlike with repetition, the correctness of the phonemes produced is not considered). Finally, *recognition* is a measure of the number of correctly identified stimuli, regardless of the correctness of phonemic production. Identifying an equivalent of recognition in a neural network is a problematic issue since it is a purely cognitive construct. The angular gyrus has been identified at times as the location for storage of semantic information [Caplan 2003; Dronkers, Wilkins et al. 2004] and as a modality-independent association area [Binder, Frost et al. 1997; Booth, Burman et al. 2002]. In our model, it is the earliest area to receive information from both visual and auditory modalities, and thus is in a unique position to associate information received through these two stimulus input pathways. For these reasons, we defined recognition to be the extent to which the AG regions' activation patterns bilaterally, following a stimulus, could be used to determine correctly what the stimulus name had been. A value of 100% correct on this measure with the intact model implies that a *unique* activation pattern was created during learning in the AG's for each word in the training data. Following lesions to the WLG model, the value of this measure indicates the extent to which the original representations of learned words in the intact WLG model persist in the lesioned WLG model.

Model performance, as determined by our four performance measures, was assessed in ten independent simulations that were identical except for initially random weights. Figure 6 shows performance of the intact model before (a) and following (b) initial training. We see that the trained model performs nearly perfectly for each of the four dependent measures. Thus, the model developed unique internal representations (AG activity patterns) for the individual named objects. It was also successful in identifying the simulated visual and auditory input stimuli and mapping them onto the correct series of output phonemes. This is a substantial accomplishment, as the correct sequence of phonemes, ranging from three to ten in length, needed to be produced from two different forms of input based solely on learning synaptic connection strengths in a complex recurrent network. We also measured a laterality coefficient value [Shkuro et al, 2000] of -0.36, indicating that the left hemisphere had a much more influential role in determining phonemic output than the right hemisphere.

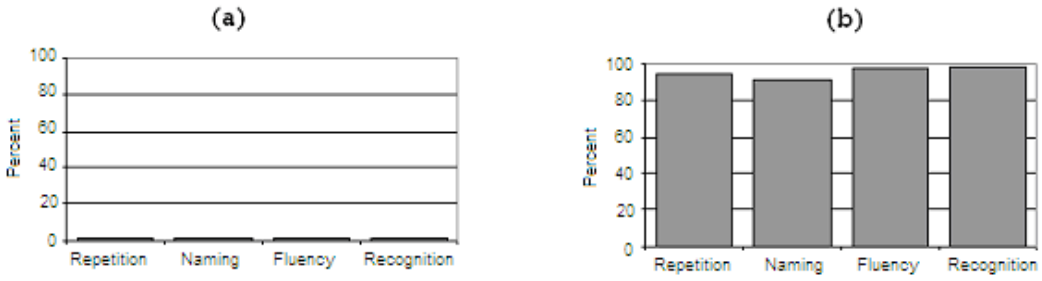
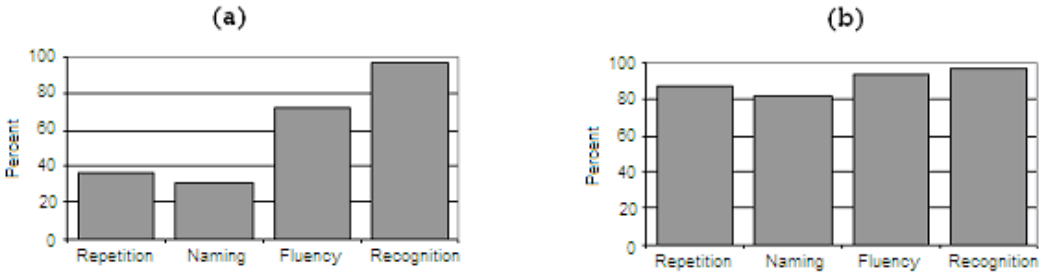


Figure 6. Unlesioned modeling performance, as assessed using the four dependent measures produced by the model. Before training (a), the model fails to identify or recognize the stimuli, but after training (b) the model performs consistently well, above 90% for each measure. ■

In addition to testing the intact model, we also examined model performance following simulated lesions to the regions WA, BA, AG, and IT, and to the AF pathway. Lesions consisted of "removing" 75% of the neural elements in a given area (or 75% of the connections in the case of the arcuate fasciculus) by permanently fixing their output to zero. The lesions roughly correspond to damage *classically* associated with Wernicke's, Broca's, and transcortical sensory aphasia, visual anomia, and conduction aphasia, respectively (Caplan 2003). However, correspondences between these biological lesion sites and aphasic syndromes are currently recognized to be imperfect at best.

Remarkably, simulated lesions to the individual regions of the model generally produced deficit patterns reminiscent of the corresponding classical aphasia syndromes seen in people [Caplan, 2003]. For example, Figure 7 shows performance following damage to the model's AF. When the left hemisphere AF was damaged (Figure 7a), both repetition and naming performance measures dropped below 40%. Fluency, although affected, dropped much less, and recognition ability did not drop at all. In contrast, damaging the right hemisphere AF (Figure 7b) had minimal effect on all four performance measures. Damage to the left arcuate fasciculus (AF) in humans is classically associated with impaired naming and repetition ability, but a retained ability to comprehend and produce some linguistic output (Anderson, Gilmore et al. 1999), consistent with the model's behavior. Comparable results were obtained with lesions to other model areas [Weems and Reggia, 2006].



Figures 7. Model performance following (a) left and (b) right hemisphere damage to the simulated arcuate fasciculus. ■

To summarize, the key finding of the current model was that it is capable of learning "from scratch" the visual image, auditory phoneme sequence representations (names), and motor phoneme sequence representations of fifty separate objects. We consider such results to be promising. Remember that we did not assign any functionality a priori to any cortical region in the model, nor did we devise any new neurocomputational methods in creating the model (i.e., we used off-the-shelf modules, activation dynamics, learning methods, etc.). The learned ability of the model to produce output corresponding to the correct phonemic representation of both auditory and visual input stimuli is not trivial, as both the auditory and motor phoneme distinctive feature representations were distinct and complex; associations had to be made at several processing levels via multi-layered neural networks. For example, in word repetition, the model did not begin to generate output motor phonemes until after *all* auditory phonemes had been processed for that word, so it had to retain an internal representation of the word from which to generate its correct pronunciation. The model had to learn to not only map to the correct sequence of motor output vectors representing phonemes from the input patterns, in

whatever form that input took (temporal auditory phoneme sequence or static image), but also had to know the correct temporal length of the appropriate output and cease output phoneme production at the correct time. This is considerably more complex than simple association learning, yet the model demonstrated near perfect performance on all performance measures in spite of the simplicity and small size of the cortical regions simulated relative to their biological counterparts.

Lesion analysis further supports the belief that word and picture recognition in our computational model was accomplished in a manner similar to that posited historically by the creators of the WLG theory. The primary finding was that the model, in general, demonstrated patterns of word processing deficits following left hemisphere lesioning much like those predicted by the WLG theory and often similar to those observed in human aphasic patients. None of these deficit patterns occurred with equivalent damage to the homologous regions of the right hemisphere, consistent with the model's acquisition of hemispheric specialization. While some of the model's deficits following lesioning would be readily predicted from the architecture of the model (just as the originators of WLG theory surmised), other results are much less straightforward. For example, it is not obvious why damage to the model's Wernicke's area should reduce repetition ability but have much less impact on fluency. This dissociation between repetition and fluency performance is an interesting aspect of this model, and is especially promising because it suggests that Wernicke's area is responsible for *meaningful* speech production, but not simply the ability to produce speech-like utterances. If Wernicke's area is simply an important part for the production of verbal output, both repetition and fluency should have both suffered following WA damage. Instead, verbal production remained high following WA damage, but that output was meaningless (high fluency), indicating Wernicke's area plays an important role in the management of performing *meaningful* speech.

B. Delayed Match-to-Sample Model

Executive functions are the high level cognitive abilities that allow manipulation of information. One major component of executive function is the ability to keep information in a short-term memory so that it can be manipulated and combined with other information. Both single-cell recordings in animals and imaging studies in humans suggest that this ability, called *working memory (WM)*, involves a network of interacting brain regions, with the frontal cortex playing a key role. WM is further divided into different components, which include the maintenance of information online, resistance to interference from irrelevant information, and operations that allow this memory to be erased, updated, and selected for further processing. Decision-making is also thought to involve operations that require comparisons of alternatives that are held in WM, and is closely linked to WM. However, the neural underpinnings of these functions are poorly understood.

Although functional magnetic resonance imaging (fMRI) has revealed brain regions that are involved in WM, there still are no techniques for relating fMRI activity to underlying neuronal circuit properties. In order to understand how the operations of WM are implemented in the brain, we have developed a large-scale systems-level model of WM that includes a method for relating the neural mechanisms to human fMRI data. The goals for the model are

that it be able to perform WM tasks that are typically used in fMRI studies, that its neural dynamics mimic those found in animal single-cell recordings, and that it reproduce human imaging results quantitatively in the brain regions included in the model. This approach makes it possible to begin to explain the human data in terms of underlying neuronal circuit dynamics. The model is composed of multiple brain regions and includes a working memory circuit that maintains representations of recently seen objects in short-term memory, and it performs a delayed match-to-sample task, in which it makes a decision about whether there is a match between a stimulus held in working memory and a stimulus that is presented after a delay, possibly with intervening stimuli. An attentional system that models the presumed effects of dopamine controls the performance. The model fulfills three major requirements for a working memory system: 1) maintaining representations in short-term memory; 2) resistance to interference; and 3) the ability to make a decision that initiates an update of the contents of the memory. Together, these features implement a form of executive control that is necessary for intelligent behavior.

The basic model that we created addresses the delayed match-to-sample task, and incorporates the ventral visual pathways. In the delayed match-to-sample task, subjects are asked to determine whether or not the current input stimulus matches a previously seen stimulus that is retained in working memory. Previously we studied a simpler model as shown in Fig. 8. Inputs are visual patterns (letters, simple geometric shapes, etc.) similar to those used in human fMRI experiments. Outputs from the prefrontal (PF) region are decisions (match or no match) about whether the current visual input matches the previous one stored in working memory. Prefrontal cortex is believed to play a critical role in working memory during this task, and this is captured in the model by cortical column circuitry inspired by electro-physiological data from this region [Tagamets & Horwitz, 2003]. This previous model serves as the basis for our new model as described in the following.

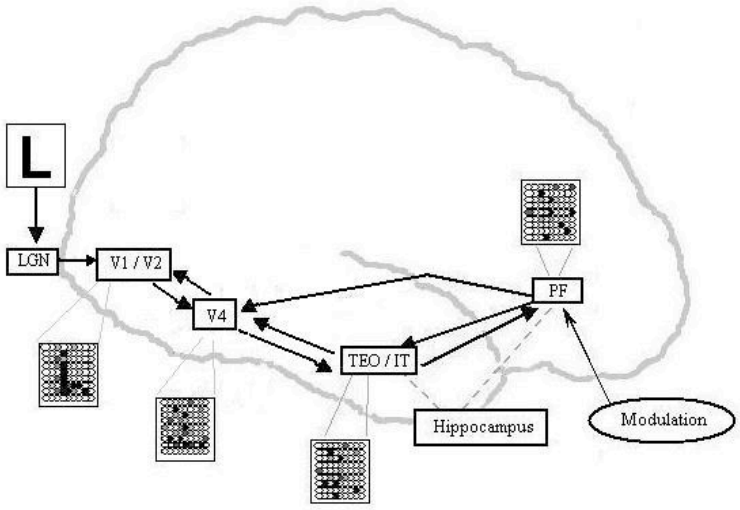


Figure 8: The match-to-sample model has connections based on neuroanatomic pathways. Only right regions and pathways are shown; left ones and inter-hemispheric pathways are also present. Grids illustrate activity patterns propagating through the network. LGN = lateral geniculate nucleus; V1/V2,V4 = early visual regions; TEO/IT = inf. temporal visual regions; PF = prefrontal cortex. ■

In new work, we explored the hypothesis that we can create an intermediate-scale neurocognitive system, but now one that includes regions from both hemispheres, working memory, and learning of interregional functional connectivity based on human fMRI data. Unlike the original functional imaging model (Figure 8), learning is now used heavily to acquire connection weights and pathway level inter-regional connectivity strengths instead of manually assigning such values. Most importantly, the new resulting model differs from most previous visual system models in being constrained to match quantitative fMRI data (some of

which we collected ourselves), in spanning two cerebral hemispheres, and by its integration with hippocampal regions and with prefrontal working memory regions. To our knowledge, no one has previously developed a neurobiologically grounded computational model of delayed match-to-sample human behavior having this scope and fidelity to behavioral, neural, and functional imaging data.

Figure 9 depicts the overall architecture of the new extended model. The task that we model, visual shape matching, involves mainly the occipitotemporal visual pathway. Single-cell recordings in primates have provided data about specific visual response properties in these areas [Tanaka 1993], as have imaging studies in humans [Sergent et al 1992; McIntosh& Gonzalez-Lima 1994; Courtney et al 1996]. This pathway includes areas V1, V2, V4, the TEO region of the inferotemporal cortex (TEO/IT), and lateral prefrontal cortex (PFC). The hippocampus (HC) is primarily associated with long-term memory (LTM) but is also thought to be involved in working memory.

Each region in the model is composed of 8x12 arrays that represent subpopulations of neurons with different types of response properties. The early visual cortices, areas V1 and V2, encode simple components of visual objects, such as line segments, their orientations, and intersections of lines that form angles. In the model, there are subpopulations that encode horizontal and vertical lines. Area V4 is the first region in the pathway that is considered to be association cortex, in which visual representations of basic shapes combine with other information, such as color (which enters the brain via a different pathway) and spatial relationships from the dorsal visual pathway. Area TEO/IT is thought to be a region of the brain that encodes whole objects, such as faces, trees, or words, with specialization for different types of objects in different populations and sub-areas of TEO/IT. The evidence for this is that damage to this region can result in selective loss of the ability to recognize specific classes of objects, such as faces, words, or even vegetables. The PFC has been implicated in executive function in general, and is thought to contain abstracted representations of objects and their context. Finally, the hippocampus (HC) has also been implicated in WM, thought its role in this is not clearly understood. Neurons in the regions V1, V2, and V4 are active only when a stimulus is in view. Neurons in regions further along in the pathway (including areas TEO/IT and PFC) have the capability of maintaining high levels of activity even when no item is currently in view. Thus these regions are likely to play a key role in WM function. However, one distinction between areas TEO/IT and PFC that has been observed in neurons from electrophysiological experiments in monkeys is that WM traces are maintained across intervening stimuli in PFC, whereas in TEO/IT an intervening stimulus replaces the current memory with a representation of the new stimulus [Miller et al, 1993]. This suggests that neuronal circuits in the PFC implement the property of resistance to interference in WM. In the model, the PFC contains the WM circuits, and feedback from the PFC to area TEO/IT enhances temporary memory maintenance for only the most recent stimulus in area TEO/IT.

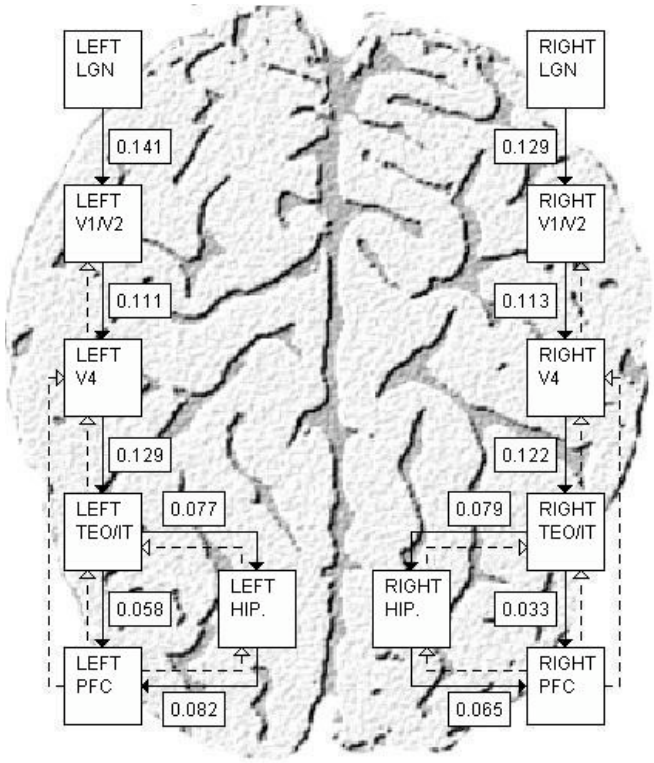


Figure 9: Architecture of the full model. Each block represents a brain region that has been implicated in visual working memory. Visual input enters the network through the lateral geniculate nucleus (LGN) and is passed forward through visual brain regions (V1, V2, and V4) that successively abstract the representations. Area TEO of the inferior temporal cortex (TEO/IT) is the region thought to be specialized for representations of whole visual objects. The prefrontal cortex (PFC) contains the working memory circuits, which maintain short-term memories of recent stimuli, and make decisions about whether they match the current stimulus.

Specialized circuits that maintain memory traces and decide on matching stimuli make up the populations in the PFC region of the model. The four different types of units in the WM circuit are based on distinct populations that have been identified in single-cell recordings in monkeys in delayed memory tasks [Funahashi & Kubota, 1994; Goldman-Rakic 1995]. Activity enters the circuit via the cue units C, from where it is passed on to D2 delay units. After the stimulus disappears, D1 delay units increase activity, and the memory is maintained by recurrent excitation between the D1 and D2 units. If a new stimulus matches the currently held memory, the response units R become active, indicating a match. Otherwise, if attention modulation is strong enough, the D1 and D2 units continue to maintain the representation of the remembered stimulus.

A separate circuit implements decision-making in the model (Figure 10). This circuit is composed of two units in each hemisphere: one that responds when a stimulus matches the one in WM, and another that responds when there is no match. These units receive inputs from all of the WM circuit units in the frontal cortex (see Figure 10), and thus collect the total response from all frontal WM circuits. The connection weights are determined by a supervised learning mechanism.

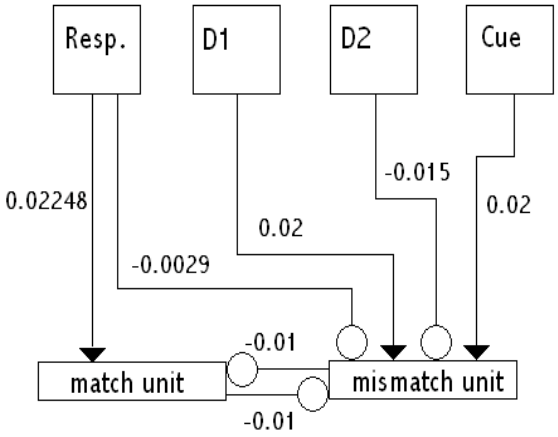


Figure 10. Decision-making circuits in the frontal cortex of the model. D1 and D2 units maintain delay-period activity if there is a sufficiently high level of attention. Response units R activate if a new stimulus matches that currently held in memory. The frontal cortex in the model is made up of an array of these circuits. All WM circuits in the frontal cortex converge onto this circuit, which includes a match unit that fires when a match has occurred, and a mismatch unit that fires when the current stimulus does not match the one held in memory.

The most critical issue with the extended model was matching model activity to fMRI, which is an indirect measure of neuronal activity. The relationship between fMRI and neuronal measures is complex. The responses of neurons constitute the computations that are performed by the brain: a high firing rate in a neuron suggests selectivity of that neuron to a particular state of the brain, e.g., it might represent that a face is in view. The connections between neurons, i.e. their strengths and patterns, determine the responses of the neurons. Changing how neurons interact changes their firing properties, and activity in these connections requires large amounts of energy. Energy requirements in local brain regions increase demand for oxygen. Finally, fMRI measures oxygen levels that change when blood flow responds to local changes in energy needs. The net effect is that fMRI is thought to mainly reflect the energy requirements of synaptic activity, and not the neuronal spiking that is commonly used as an index of encoding. We have previously demonstrated that the consequence of this is that there can be a dissociation between neuronal spiking activity and measured fMRI [Tagamets & Horwitz, 2001].

Our approach to modeling fMRI has been to focus on the numbers and attributes of the connections. Most connections in the brain do not extend beyond a fairly localized area, and this is captured in the architecture of the local circuits. Thus, incoming activities can have potentially large effects on the local circuits, depending on their configuration. fMRI data is modeled as the sum of all incoming and local circuit connections within a region integrated over time.

We developed a neural network learning algorithm, the *gains learning* algorithm, that can be used to find the strengths of interregional connections for the model to match activations in an arbitrary fMRI data set [Winder et al, 2006]. This problem differs from the usual supervised learning methods in neural networks in that there are target values for all regions in the

network, not just for an output "layer." This method allows estimation of functional connectivity while allowing for effects of the interaction of interregional and local circuits. It differs from other measures of functional connectivity for fMRI currently in use in two major ways. First, the model itself is a generative one that attempts to explain how imaging data such as rCBF and BOLD can be explained by neuronal behaviors. Second, the connection training method is based on matching average activations in the regions of interest (ROIs), as opposed to other methods such as structural equation modeling [McIntosh & Gonzalez-Lima, 1994], partial least squares [McIntosh, 1998; McIntosh et al., 2004; McIntosh & Lobaugh, 2004], and DCM [Friston et al., 2003; Mechelli et al., 2003; Penny et al., 2004], which derive effective connection strengths from covariances or correlations that are computed from the data. This is an advantage over the other methods, since in many cases, average within-task activations are computed more reliably than within-task covariances, which are difficult to compute in event-related fMRI data.

The gains learning algorithm is a gradient descent method that attempts to find solutions that will minimize the overall error between modeled and target activations. It was demonstrated empirically that this algorithm converges to unique solutions, and that it finds the correct solutions on data with known connectivity. We then applied it to an fMRI data set in order to examine connections in healthy control subjects as they performed a working memory task involving linguistic stimuli. The connection weights shown in Figure 9 provide an example. These specific weights were derived by using this learning method on the fMRI data. The learning algorithm was also applied to fMRI data from the same task in a group of volunteers with schizophrenia. The greatest differences between the groups were found in temporo-frontal connections between the groups, a result that is consistent with a number of other imaging results in schizophrenia.

Finally, we examined the effects of modifications in local prefrontal circuitry on changes in fMRI activations, functional connectivity, and performance of the task. The results of damaging the frontal circuitry suggest that functional connections are much more sensitive to these changes than BOLD activations, and the performance changes are suggestive of working memory deficits commonly found in schizophrenia, in that the memory is more susceptible to interference. Together with evidence for local circuit disturbances in prefrontal regions in schizophrenia, our results suggest that decreased recurrent excitation within prefrontal cortex can simultaneously explain the disconnection between frontal and temporal regions and the deficits in working memory in schizophrenia.

In summary, our work with modeling functional imaging data has been directed at gaining a better understanding of both the quantitative and qualitative properties of this data and in elucidating the underlying neuronal circuits that carry out cognitive operations. This method allows experimental results from the animal literature to be incorporated into explaining fMRI data. We conclude that this combined theory-driven and data-driven methodology extends current imaging analysis methods, and allows examination of properties other than total activations and functional interregional connection strengths that are currently in use for fMRI data analysis.

C. Adaptive Sensorimotor Control Model

A primary goal of research on the cognitive neuroscience of decision-making is to produce a comprehensive model of behavior that flows from perception to action (including decision-making) with all of the intermediate steps defined. The model should be able to generate not only simulated neural activity, fMRI and other functional neuroimaging data (as shown in Sections IIIA-B), but also behavioral performance (i.e., accuracy and reaction time data) data in both intact and neurological conditions. Although we and others (e.g., [Husain et al., 2004]) have developed models of perception, and models of action have also been put forward [Bullock et al., 1993; Guigon and Baraduc, 2002; Contreras-Vidal and Wen, 2003], integrating perception, decision-making, and action networks is still needed. To address this gap, we have recently integrated a model of adaptive frontal-parietal sensorimotor transformation with the Bullock et al (1993) model of redundant arm reaching. Importantly, our model now incorporates complementary parallel cortico-cerebellar-thalamo-cortical and cortico-striato-thalamo-cortical neural "loops" that are thought to be critical for motor adaptation learning in response to developmental and/or environmental changes.

The hypothesized cortical sensory integration and coordinate transformations required for controlling an arm reaching to visual targets (summarized in Figure 11) can be initially learned through simultaneous exposure to patterned proprioceptive and visual stimulation during self-produced movement [Bullock et al, 1993; Guigon and Baraduc, 2002]. These sensorimotor transformations or mappings can then later be updated (or new maps formed) with the help of fronto-parietal and/or parieto-cerebellar circuits. Recent motor control theories suggest that the brain uses internal models to learn these mappings, and to plan and control accurate movements. An internal model is thought to represent how the biomechanics of the arm interacting with the outside world would respond to a motor command; therefore it can be seen as a predictive model of the reafference that helps the system plan ahead [Imamizu et al., 2000]. For example, during adaptation to 'force fields' (external forces applied through a robotic manipulandum which alter the normal dynamic characteristics of arm motion), these adaptive internal models are thought to generate compensating torques which allow the arm to track an invariant reference trajectory to a specified target. In the case of a distorted kinematic environment (e.g., altered screen cursor-hand relationships), the internal model would represent the new inverse kinematics required to transform a desired movement vector in visuospatial coordinates into a joint-based motor command. Adaptive sensorimotor behavior therefore involves the problems of localizing the hand and the targets in space, trajectory planning (which involves computing the vector linking the hand to the target), coordinate transformation, and control, and the brain must solve these problems to bring the hand from the starting position to a desired target location. There are many different approaches to modeling adaptive sensorimotor behavior ranging from adaptive control techniques to biologically-inspired neural network approaches [Bullock et al, 1993; Contreras-Vidal et al, 1997]. A benchmark test performed by many researchers in motor learning is a reaching task between points usually lying along the circumference of a circle at equally spaced intervals (i.e., the 'center-out' task). The human operator is told to move 'as fast as possible', the cursor on the computer screen from point A to point B in a straight line using a robot manipulandum, or a computer mouse or pen as input devices. The experimenter then either distorts the kinematic mapping of the handle (or mouse) or programs the robot handle to exert environmental force disturbances on the subject. This allows researchers to examine how subjects react to various kinematic and dynamic

perturbations, thus furthering their understanding of any adaptive processes that might be occurring in parallel.

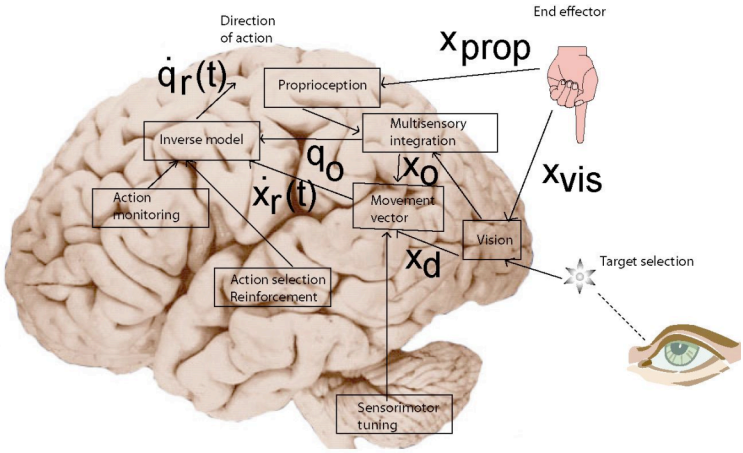


Figure 11: Adaptive sensorimotor transformations for redundant reaching. Visual (x_{vis}) and proprioceptive (x_{prop}) signals are integrated to form a multimodal representation of initial hand position (x_o), which can then be compared to the desired target location (x_d) to compute the movement vector in visuospatial coordinates. Next, inverse kinematic computations are performed in a parieto-premotor network to transform changes in end-effector position (dx_r/dt) into changes in joint angles (dq_r/dt) that specify the desired trajectory of movement, that is, the intended direction of action. Deviations between desired and actual movements, detected by an action monitoring system, causes progressive recruitment of basal ganglia and cerebellar networks, respectively, for updating the sensorimotor transformation networks. It is hypothesized that action monitoring is performed by the anterior cingulate cortex, whereas action selection and reinforcement may be related to basal ganglia function and sensorimotor tuning may result from cerebellar involvement. Notation adapted from Sober and Sabes (2003). ■

An interesting result of these studies has been that despite large differences in models, they often display three common features: a trajectory generator, sensory feedback and control loops, and an adaptive process. Models of reach planning based on either kinematic (e.g., velocity command model) or dynamic (e.g., torque command model) variables have been proposed in the literature. In the velocity command model, the motor command is specified as joint angle velocities, whereas in the torque command model, the trajectory generator outputs a set of nominal joint torques for the arm which is learned over a lifetime of executing such tasks. The second feature is that humans will use visual and kinesthetic feedback to correct for the arm motion when it drifts from this nominal path. Finally, humans will adapt to environmental disturbances or distortions to the sensory mappings to keep the arm moving along the desired trajectory [Shadmehr and Mussa-Ivaldi, 1994; Contreras-Vidal and Buch, 2003]. Researchers such as Bullock et al (1993), Sanner and Kosha (1999) and Thoroughman and Shadmehr (2000) have used neural network based approaches to attempt to model these adaptation processes. Their results indicate that the brain constructs motor commands using computational elements that are a superposition of primitives that have gaussian-type tuning functions that encode hand velocity. However, these models cannot account for the effects of neurological lesions (e.g., Parkinsonism or cerebellar lesions) nor the functional and kinematic changes resulting from environmental changes such as screen cursor rotations or force field perturbations.

To address these gaps, we have extended Bullock et al (1993) model of redundant reaching by complementing the cortical circuit with two sub-cortical networks, namely, a fronto-striatal loop and a parieto-cerebellar loop. The fronto-striatal network is modeled as an adaptive search element, guessing new sensorimotor transformations and reinforcing successful guesses while punishing unsuccessful ones [Grosse-Wentrup and Contreras-Vidal, 2006]. This system uses an error (evaluative) signal to drive the selection and the reinforcement/punishment mechanisms. The parieto-cerebellar component is modeled as an adaptive error-correcting module that continuously updates a correction term to drive the error of actual versus desired movement to zero [Contreras-Vidal, Grossberg, Bullock, 1997]. Simulations of a kinematic visuomotor adaptation task using a redundant arm moving in the horizontal plane (see Figure 12) and with learning processes disabled in the cortico-striatal network resulted in error curves resembling those of Parkinson's disease patients [Contreras-Vidal and Buch, 2003; Grosse-Wentrup and Contreras-Vidal, 2006]. Simulated PET data have also been computed to assess the functional activation of various brain regions of interest [Contreras-Vidal and Wen, 2003]. The model's patterns of simulated constant and variable errors were found to match the learning curves seen in the experimental data. In agreement with experimental studies (Inoue et al, 2000), for example, the simulated PET signal of superior parietal lobe showed an increase in functional activation due to the introduction of the visual feedback distortion. Our simulations also showed increased activation in the lateral cerebellum as reported by Imamizu et al (2000) in a similar study involving adaptation to rotated screen cursor-hand relationships.

IV. Towards a Large-Scale Neuromorphic Architecture

In the previous sections, we outlined some basic design principles for organizing a neurocognitive architecture, and showed that one can readily build substantial portions of system-level modules of such an architecture at present using many of these principles. In this section, we focus on specifying the behavioral capabilities required for a large-scale architecture representing a situated and embodied agent functioning in a naturalistic setting. Our intent is not to exhaustively list and describe all of these, but rather to address those that we see as critical in the initial development of the architecture. Specifically, we consider the range of systems needed in a large-scale architecture, including

Sensory Systems
Motor Control
Memory
Language
Executive Functions

We outline the scope and functional requirements of each system, and provide where possible an indication of important neurobiological inspiration for functionality. Implicit in all of this is that learning is involved in each system listed above, and that each system is based on the principles outlined in Section II.

A. Sensory Systems

A full-scale neurocognitive architecture will need to process, interpret and act upon a broad range of sensory inputs. At a minimum, these include analogs to human vision, audition and proprioception, but in specific situations might also include more exotic senses such as infrared detection or chemical recognition. Here we consider general issues concerning sensory processing, illustrating them with specific details for the visual system, the most highly elaborated sensory system in humans and non-human primates.

While the goal of sensory systems is to convey and interpret environmental information, in a broader sense this information must also be evaluated in terms of three factors: 1) the need for action; 2) its unexpectedness/novelty; and 3) its usefulness. If action is required, the sensory information needs to facilitate an estimate of its urgency. Urgent needs have to be routed via faster pathways to more automatic response systems, while non-urgent information can be further evaluated for novelty and usefulness. Novel stimuli convey less certainty about utility and need for action, but more potential for unexpected results. A full architecture will need sensory systems that encode features that address each of these three factors. One key factor that especially needs to be accounted for is the amount of time it takes to process information that has varying degrees of immediate utility. Each of the factors outlined above takes increasingly more time for processing, and thus they will be encoded and represented in separate but integrated subsystems. Toward this end, the full architecture should have parallel pathways that accommodate fast reflexive actions and increasingly slower evaluative processes.

The organization of biological sensory systems provides useful insights for an artificial architecture. Biological systems carry out sensory processing by means of hierarchical pathways that successively transform incoming stimuli from specific features to progressively more abstract representations that include encoding of context. When sensory information reaches the neocortex, it passes through a network of cortical regions that provide increasingly abstract representations. Each modality has a primary sensory cortical area containing different populations of neurons that respond selectively to very specific properties of the incoming stimuli. For example, in the visual system, the primary visual cortex has separate populations of neurons that encode features such as orientation of line segments and edges, direction of motion, and color. These primary region populations forward their encoded information via multiple separate pathways that each process distinct types of visual information. A ventral system that deals with object identification (the “what pathway”, as in the match-to-sample model earlier) and a dorsal system (“where pathway”) that processes spatial information, were identified some twenty years ago [Ungerleider and Mishkin 1982]. Other identified visual pathways include those for color perception and motion detection. These multi-area pathways are also highly interactive with one another. Both monkeys and humans have specialized regions for face and hand recognition, and for distinguishing between biological and non-biological motion. Humans have also been found to have specialized brain regions for specific objects besides faces and hands, such as words, places, tools, and other classes of objects. The price of such complexity is relative slowing of operations, since visual information must pass through multiple levels of analysis before a full representation is built.

In the object vision pathway, edges and oriented line segments are successively composed into groupings that represent parts of objects, and finally reach the inferior temporal

cortex, where visual representations of whole objects are thought to be stored in the brain (at least in large part). Initially separate from each other, the pathways exchange increasingly more information further along in the hierarchy, so that, for example, the color and shape of objects are represented together in the inferior temporal cortex. Beyond this area, the visual representations are increasingly integrated with other sensory systems in multimodal association regions in the superior and anterior temporal cortex. The abstracted representations finally converge in two main high-level brain regions, as follows. Abstracted, possibly multimodal representations converge in the frontal cortex, where evaluation of the incoming information is integrated into the context of the current state. Other pathways lead to the medial temporal lobes, which mediate integration of new information into long-term memory. Presumably each of these endpoints requires somewhat different representations of the content. We have previously implemented models of sensory systems for the object vision pathway as described above [Tagamets and Horwitz 1998] and the auditory pathway [Husain et al. 2004]. These models have a hierarchical structure, and include a short-term memory module that maintains representations of recent items and makes decisions about the similarity of new information to that currently in short-term memory.

Humans also possess fast sensory pathways that lead almost directly to the orienting motor output system, allowing for reflexive responses that do not depend on the more extensive, slower evaluation in the main pathways. In the visual system, there is a pathway that leads directly from the retina to the superior colliculus, which, in turn, is connected directly to the motor output system. This pathway is particularly sensitive to motion, and is presumably useful for quickly responding to potentially threatening movements. Similar fast pathways exist for the auditory and somatosensory systems, mediating the startle reflex to sudden sounds and the somatic reflex arc that mediates reflex reactions to sudden pain on the surface of the body. These fast pathways do not process the level of detail that is available to the other pathways, and thus are not necessarily available for conscious perception. Rather, they encode existence of sudden changes in the environment, such as sudden movement or a sudden loud noise, favoring speed over detail.

B. Motor Control

The organization of biological motor control systems in human and non-human primates provides useful insights that may be adopted in an artificial sensorimotor control system. Anatomical, physiological and clinical studies suggest multiple spatially segregated cortico-striatal and cortico-cerebellar loops, with each loop involved in a distinct aspect of cognitive-motor behavior as summarized in Figure 12 [Hoover & Strick, 1993; Middleton and Strick, 2000; Doyon et al, 2003]. Each basal ganglia loop originates in a cortical area, and enters the basal ganglia through cortical projections to the sensorimotor (skeletomotor) striatum or the associative (cognitive) striatum. The basal ganglia output to individual cortical areas appears to originate from separate regions in the internal segment of the globus pallidum (GPi), which in turn projects through specific thalamic areas to cortical areas known to control distinct aspects of behavior. Thus, in reaching tasks involving hand movements to a chosen visual target after an instructed delay period, cortico-basal ganglia activation would be characterized by neuronal activation of dorsomedial regions of the GPi which project to the supplementary motor area (SMA) involved in motor preparation, as well as to prefrontal areas responsible for spatial working memory during a delay period (area 46). Moreover, ventrolateral GPi neurons and their cortical targets in the ventral premotor (PMv) area, which is involved in target

selection, and coordinate transformations for movement, would also be recruited. As the movement command is released and the movement starts to unfold, primary motor cortex and related basal ganglia and thalamic areas would be recruited to control movement parameters such as movement speed and size rescaling.

In this scenario, cerebellar associative and motor channels originating in distinct regions of the cerebellar dentate nucleus would be recruited in parallel to reduce the discrepancy between desired and actual (or predicted) states. Thus, while the basal ganglia may be involved in the selection of appropriate movements and/or control strategies based on external cues, the cerebellum may be involved in the recalibration of motor commands through the adjustment and optimization of movement parameters [Jueptner and Weiller 1998]. Thus, it appears that functional cortico-basal ganglia engagement is crucial in tasks that are initially effortful and in which correct responses are self-selected through trial-and-error. However, once the appropriate action has been found and stabilized, the cortico-cerebellar networks can fine-tune the internal model through practice until the task can be performed automatically. This updating is likely to be accomplished through internal models representing the forward and inverse computations related to sensory-motor mappings and the interaction of the body and the environment [Kawato & Wolpert, 1998].

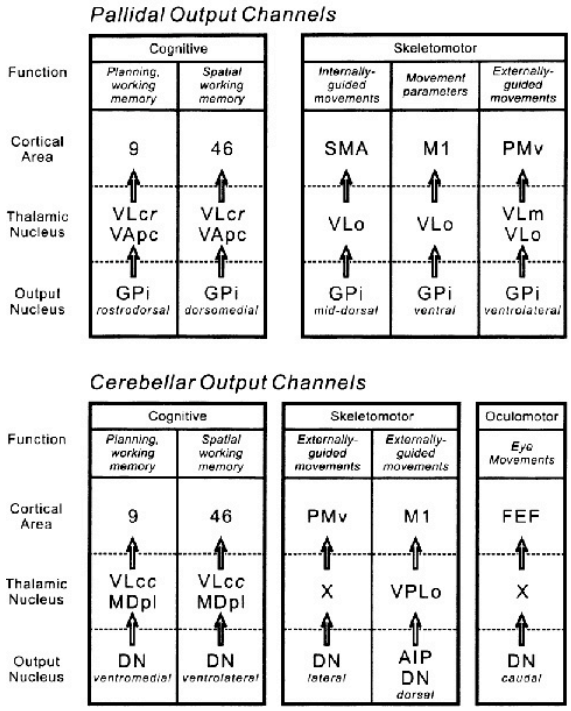


Figure 12: Multiple spatially separate cortico-basal ganglia-thalamo-cortical and cortico-cerebellar-thalamo-cortical networks involved in distinct cognitive and motor functions. AIP, anterior interpositus nucleus; DN, dentate nucleus; GPi, globus pallidum pars interna; SMA, supplementary motor area; PMv, ventral premotor cortex; M1, primary motor cortex; Vlo, ventrolateral nucleus, oral division; X, area X; VLc, ventrolateral nucleus, caudal division; VA, ventroanterior nucleus; VPLo, ventral posterior nucleus, oral division; pl, posterior lateral; m, medial; cr, caudal-rostral; pc, posterior-caudal. Note that basal ganglia and cerebellar input areas (e.g., striatum and cerebellar cortical areas) are not shown. Adapted from Middleton and Strick (2000).

Importantly, the internal (forward) model contributes a speed advantage by using predicted sensory reafference instead of waiting for actual reafference. However, the system's predictive ability is adaptively recalibrated by new experiences with actual kinesthetic consequences of motor signals. These kinesthetic signals provide a non-visual basis for measuring the evolving configuration of the body parts and their positions vis-à-vis the visible or remembered location of environmental features. Interestingly, although internal parallel forward models can actually improve upon kinesthesia, they are also parasitic upon it, and can be expected to degrade if kinesthesia degrades [Contreras-Vidal and Gold, 2004].

The importance of having internal models of movement control is best exemplified by behavioral studies of adaptation/learning in the presence of kinematic and/or dynamic perturbations during limb movements. The generation of descending neural commands appropriate for required actions implies that the internal representation includes a mechanism for selection of motor commands that can produce the desired kinematic output [Dounskaia 2005]. Moreover, these adaptation studies demonstrate the existence of aftereffects after extended practice under distorted environments that represent learning and/or updating of internal kinematic or dynamic representations of the interaction of the limb and the environment [Kagerer & Contreras-Vidal, 1997; Contreras-Vidal & Kerick, 2004]. Neuroimaging studies suggest both cortical and subcortical substrates for the acquisition of these internal models. For example, fMRI and PET studies of adaptation to kinematic distortions (e.g., alterations of hand-cursor relationships) in human subjects show a highly interconnected network comprising the putamen, preSMA, PMv, posterior parietal cortex (PPC), and the lateral cerebellum [Imamizu et al. 2000; Krakauer et al., 2004].

Recent studies indicate that these internal models are not necessarily exact replications of complex dynamic equations that account for interaction torques, gravitational forces, or inertial characteristics of the arm [Dounskaia 2005]. Rather, approximations of internal representations for proximal and distal components of movement are used to simplify the computational burden. Specifically, it appears that for planning and control purposes, the 'leading' joint responsible for launching and directing the movement towards the desired target is initially controlled without regard to interaction torques, whereas the function of the 'secondary' or joint 'slaves' is to take advantage of interaction torques generated by the leading joint to take the end-effector to the target.

The principles underlying internal model formation and representation can be generalized to learning of internal representations of the spatial or contextual environment for navigation. The hippocampus of the rat has been hypothesized to host a spatial representation of the animal's surrounding environment [O'Keefe & Nadel, 1978], as the firing of hippocampal 'place' cells is strongly correlated with the location of a freely moving rat in its environment. If the environment changes (e.g., configuration, orientation and color of objects), remapping of the hippocampal map occurs leading to distinct maps for distinct environments [Lever, Willis, Cacucci, Burguess, & O'Keefe, 2002; Bostock, Muller & Kubie, 1991; Cressant, Muller & Poucet, 2002]. It has been suggested that a dynamical spatial and temporal representation of the space and task environment based on the encoding of transitions provides a natural solution for switching from a spatial cognitive map to its motor implementation [Banquet et al, 2005]. The so-called 'place' cells integrate visual and movement related information during navigation. In this scenario, transitions are associated with their movement vector by convergence of place information and path integration as navigation takes place. Transitions are computed using current direct and delayed indirect visual inputs, and spatiotemporal contiguity between successive place fields is ensured by Hebbian learning of a contextual map during exploration [Banquet et al, 2005]. The transition cells in this cognitive map are the building blocks of the neural representations of temporospatial sequences, graphs, and contextual maps putatively stored in parietal or prefrontal cortices. They appear to correspond to the internal representations for inverse/forward kinematics used during arm reaching in which changes in end-effector location are associated with changes in joint angles and vice versa. Thus, internal models of sensorimotor transformation or coordinate

transformations (e.g., from changes in visual space to changes in joint space) may represent general design principles used for movement planning and control during navigation and reaching [Grosse-Wentrup and Contreras-Vidal, 2006].

C. Memory

The range of memory functions needed in a full-scale architecture is illustrated in Figure 13. In this hierarchy, which is typical of how neurospychologists view human memory organization today (e.g., [Baddeley, 1997]), different types of memory are often distinguished via dichotomies: long-term memory versus short-term memory, implicit versus explicit memory, semantic versus episodic memory, etc. Of these, the distinction between long and short term memory is arguably the most important. Information must not only be stored in a long term sense, becoming available days, months, or years after it is integrated with existing knowledge, but must also be readily retained over shorter periods of time, typically down to the range of seconds. Whereas the former is essential for the gradual accumulation of knowledge about one's environment, the latter is essential for the routine availability of that information both during and after storage. Working memory is sometimes used synonymously with short-term memory, although it refers to that information within short-term memory that is also manipulated as part of active cognitive processing. Implicit memory is that which does not require conscious awareness for recall, whereas explicit memory requires attention. Mirroring the distinction between implicit and explicit memory is procedural and declarative memory, with the former involving attention-independent motor skill learning, and the latter involving attention-dependent accumulation of knowledge. Within different kinds of explicit or declarative memory there also exists semantic and episodic memory. Semantic memory involves recall of meaning and other general knowledge not necessarily linked with the learning event itself, and episodic memory involves the recall of information along with the context and other environmental factors that were present at the time of learning.

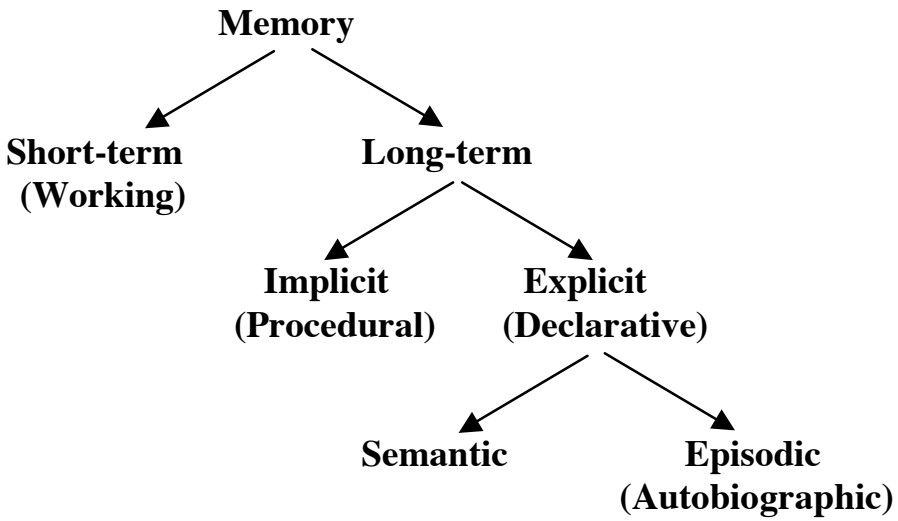


Figure 13: The range of memory functions needed in a full-scale architecture. Approximate synonyms are indicated parenthetically. ■

Several cortical areas have been identified as having especially important roles in memory storage and recall, and provide inspiration for model development. First, the medial temporal lobe, which includes hippocampus, entorhinal, perirhinal, and parahippocampal cortex, is highly involved in memory storage and consolidation [Tulving and Markowitsch, 1998]. Explicit memories appear to be stored in neocortex in a distributed manner. Memories are also stored only in relation to existing knowledge, so that both storage and recall of new information leads to a restructuring of one's representation of the world. Second, premotor cortex and subcortical structures, such as basal ganglia and cerebellum, are important for procedural memory, a type of implicit memory involving motor function [Schacter, 1987]. Third, limbic structures are important in making salient events more likely to be stored in memory [Damasio, A., 1996]. Finally, prefrontal cortex plays a key role in maintaining information in working memory, and for providing information about context that is important for episodic memory [Cohen et al, 2004].

There are important implications of the hierarchy of memory types in Figure 13 and of the many brain regions playing roles in human memory. The memory system of a full-scale architecture should be highly distributed and highly overlapping with other systems. Clearly, memory storage and recall does not occur independently of other aspects of cognition. Using the neurobiological mechanisms of human memory as a starting point, it appears that model regions will fall into two classes: regions in many systems (sensory, motor, language, etc.) will need to be involved in memory storage and recall, while separate regions should be responsible for consolidation (analogy to hippocampus), for linking stored information with specific information about the learning event in which it was acquired, for providing context in episodic memory (prefrontal cortex), etc. For example, the memory system must be closely linked with the motor system for that part of implicit memory that involves movement and navigation. Finally, mechanisms are required for recognizing particularly salient and novel information so that it may be given priority for storage.

D. Language

The WLG model (associative word learning model) of Section IIIA provides a useful starting point for building a full-scale language processing system. That model learns to recognize, repeat and produce names represented as phoneme sequences for a limited number of object images. While the language processing facilities of a full architecture must extend qualitatively beyond such abilities, a full human competitive natural language processing system seems an improbable target for the short-term, given that a half century of work with this goal has met only limited success. A more reasonable goal is to produce an architecture that learns a restricted but representative core vocabulary, perhaps comparable to that of a young child, and that has the ability to learn additional task-specific words when appropriate. Here we assume that input/output is spoken in the form of phoneme sequences. Contemporary speech processing methods could be integrated with such discrete phoneme processing if needed.

To achieve language sufficiency, the information learned by a full-scale architecture will need to go beyond the WLG word association model in at least three fundamental ways. First, a much larger vocabulary is needed, and must include words naming non-objects (verbs describing states and actions, conjunctions, negation, prepositions, etc.). Word sense

disambiguation will become very important in this context. Second, the full architecture must be able to interpret heard word sequences, such as multi-word names and complete sentences, extracting their meaning in terms of the current situation, external events, and internal goals. This includes basic tense information (past, present, future, etc.). Third, the architecture will need the ability to map its internal representations of situations, events, goals, etc. onto multi-sentence spoken utterances in appropriate situations. From a pragmatics perspective, the architecture must be able to determine when it should ignore/attend to input communication, and when it is appropriate to initiate spoken sentences. This latter ability transcends language in a narrow sense, overlapping with the executive functions and control issues discussed later in this report. To assure generality of the mechanisms used to support all of these language capabilities, a reasonable requirement to impose on the architecture is that its language learning mechanisms should work independently of the specific natural language used for training (English is assumed here, but in principle the use of any other natural language should be learnable just as well).

While our understanding of the neurobiological basis of language is quite limited, there are many insights available from lesion data and contemporary experimental studies (fMRI, EEG, MEG, etc.) that provide hints and guidelines for implementing a language system, including those summarized earlier in Section IIIA. Example sources of further biological inspiration include the following. First, the neurobiological basis of word meanings (semantics) is heterogeneous and widely distributed across a number of neocortical areas bilaterally, including secondary/association and motor areas [Saffran and Sholl, 1999]. This suggests that a substantial part of learning involving the language system will need to occur in the context of at least partially trained sensorimotor systems, providing both a basis for grounding word meanings as well as important constraints on a learning agenda. Many intriguing psycholinguistic disruptions following localized brain damage (category-specific deficits, loss of specific details vs. category information, dissociation of abstract versus concrete word impairments, etc.) provide both clues as to the organization of stored information and the potential patterns of behavioral disruption that could be useful should one wish to assess the neurobiological plausibility of an implemented artificial architecture. In contrast to semantic processing, the neurobiological mechanisms underlying syntactic processing tend to be more localized and largely confined to the language-dominant hemisphere. Both lesion studies and functional imaging data have fairly consistently indicated that Broca's area is most strongly associated with grammatical encoding and parsing operations, although there is some conflicting evidence, and the left anterior temporal and other regions have also been implicated [Hagoot et al, 1998]. Perhaps the most surprising aspect of this observation is that the use of syntactic information, even during heard sentence interpretation, seems to be tied to a "speech output area" (Broca's area) rather than receptive temporo-pareital regions, as might be expected from classic WLG theory. At the very least, these observations of distinctly distributed neurobiological mechanisms suggest that syntactic and semantic processing should be viewed as separate, concurrently active operations that jointly constrain sentence interpretation and generation.

E. Executive Functions

The executive functions of planning and goal-seeking are high-level cognitive operations that bestow capabilities often thought to be uniquely human. From a neurocomputational perspective, planning tasks are among the most difficult to solve. A

number of nontrivial issues need to be addressed for a successful planning system: 1) What is an optimal representation for a plan? 2) What regulates the sequencing of operations, both when creating a plan and when executing it? 3) How is progress evaluated as a plan is carried out? 4) How are corrections and revisions made when things do not proceed according to plan? 5) What role, if any, do emotion and motivation play in human plan formulation? and 6) How does success or failure in formulating and/or carrying out plans relate to learning new or better strategies? While sophisticated symbolic AI software for automated planning is available today [Ghallib et al, 2004], it is generally not robust to noise or unexpected events, and is generally hand-coded rather than learned.

Planning and goal-seeking behaviors are generally referred to as executive functions in the cognitive literature. Other aspects of executive function are voluntary attention, working memory, and inhibition of inappropriate or irrelevant thoughts or actions. These functions all play a role in planning. However, there is evidence that there are other, possibly subconscious, aspects of cognition that play an important role in strategic behaviors. A recent study examined the influence of deliberation on decision making [Dijksterhuis 2006]. Not surprisingly, this study found that people made better decisions after thinking about pros and cons in a simpler decision task, when there were up to 4 or 5 factors in the choice. But in complex decisions, when there were more factors, better decisions were made when people were not given an opportunity to deliberate about the options, but rather used their "gut" feeling to make a choice. This finding supports the view that the human brain is a stochastic machine that has limited resources for logical manipulation of complex situations.

Recent theories of the organization of prefrontal cortex address how these limited resources might be optimized by hierarchical structuring of goals. In this view, the most anterior part of prefrontal cortex generates and holds general goals and schemas [Ramnani 2004]. Pathways from this region enter other frontal regions that process successively more specific information, with the endpoint being the most posterior part of frontal cortex, i.e. the motor cortex, which initiates actions. In between the two extremes lie parts of the frontal cortex that have been associated with working memory, which is thought to be the limiting factor in complexity of cognitive operations. Working memory is generally viewed as a temporary store in which contextual information can interact with incoming information. For example, in operations for summing a series of numbers, each successive number must be added to the currently held sum, that sum must be updated, and maintained until the next number is available. The contextual information is usually a combination of content retrieved from long-term memory (e.g. knowledge that 4 + 2 = 6) and new information (i.e. the current task). This organization scheme suggests that chunking into subgoals might be a continuum, and that working memory is a key player in selecting, scheduling, and evaluating the organization of subgoals. Another salient aspect of this scheme is that the anterior frontal cortex has often been associated with affective states, or emotional aspects of decision-making. This suggests that the most abstract level of goal formation is likely to be strongly affected by the internal state of the individual, by factors such as mood, motivation, and arousal. Finally, information certainly goes in the opposite direction, thus "informing" the anterior frontal cortex of the consequences of the operations. Presumably this information is abstracted as it proceeds in the forward direction, and is ultimately reintegrated into the internal state. Theories of reward and reinforcement learning rely on the

salience of outcomes to an internal motivational state, and this may be a mechanism by which goals are revised and new strategies can be learned.

V. Conclusions and Implications

In this report, we presented a conceptual framework for the long-term development of a large-scale machine intelligence that is based on the modular organization, dynamics and plasticity of the human brain. Some basic design principles were presented along with a review of some of the relevant existing knowledge about the neurobiological basis of cognition. Three intermediate-scale prototypes for parts of a larger system were successfully implemented and evaluated, and these provide support for the effectiveness of several of the principles in our framework. We conclude that a human-competitive neuromorphic system for machine intelligence is a viable long-term goal. For the short term, however, substantial integration with more standard symbolic methods as well as substantial research will be needed to make this goal achievable, and we will consider such issues in Part 2.

The two intermediate-scale models that we studied both learned to perform relatively simple cognitive tasks. The WLG model dealt with processing single spoken words and images of the objects named by these words. The match-to-sample model learned to make decisions about whether a visual pattern was the same as the preceding one. In both cases the model was composed of a network of multiple regions with interconnecting pathways that could be directly related to neocortical and subcortical brain structures. Further, training these models was computationally tractable: learning times were measured in hours using contemporary laptop computers. The primary conclusion from these results is that one can readily build substantial portions of system-level models of basic aspects of cognition at present using the framework and principles that we described in Section II. This conclusion is supported by other experiences building system-level models of lower-level sensorimotor mechanisms such as limb control.

More specifically, the results that we obtained establish the following important aspects of our conceptual framework including the following:

1. It is possible today to routinely assemble networks of regions whose functionality is not pre-assigned or programmed-in, but is determined during learning by their location within a network of interconnected regions.

2. Temporal sequences can be recognized and generated appropriately by such networks following training, based on recurrent connectivity between regions.

3. A learning agenda can be used to divide the learning process into manageable pieces, allowing an entire system to learn in a multi-step process that resembles the occurrence of multiple stages during human childhood development.

4. Working memory can readily be implemented as regional activation attractor states, activation patterns that persist across multiple input/output events.

5. Learning of higher-level pathway weights (gains) can be guided effectively via data about functional connectivity of brain areas collected during experimental fMRI studies.

6. The complexity of these systems makes it very difficult to monitor their dynamics and changes during learning; a graphic interface permitting visualization of model states is very informative and increasingly necessary as the size of a system increases.

Of course, the exploratory systems done to date are quite limited in the size of their regions and the generality of the cognitive tasks that they address, and they are not integrated with other systems. This implies that some key issues that remain to be specified are how to scale systems up to full human-level functionality, the range of systems that should be developed in a complete system, how individually-developed systems will interact and become integrated, and how top-level control will work.

V. Literature Cited

Abbott L & Regehr W. Synaptic Computation, *Nature*, 431, 2004, 796-803.

Anderson J et al, Integrated Theory of Mind, *Psych. Rev*, 111, 2004, 1036-60.

Anderson, J., R. Gilmore, et al. (1999). Conduction aphasia and the arcuate fasciculus: A reexamination of the Wernicke-Geschwind model. *Brain and Language* 70(1): 1-12.

Baddeley, A. *Human Memory: Theory and Practice*, Psychology Press, 1997.

Banquet J, Gaussier Ph, Quoy M, Revel A & Burnod Y. A hierarchy of associations in hippocampal-cortical systems : Cognitive maps and navigation strategies, *Neural Computation* 17, 2005, 1339-1384.

Bi G and Poo M. Synaptic Modification by Correlated Activity, *Annual Review of Neuroscience*, 24, 2001, 139-166.

Binder, J., J. Frost, et al. (1997). Human brain language areas identified by functional magnetic resonance imaging. *Journal of Neuroscience* 17(1): 353-362.

Booth, J., D. Burman, et al. (2002). Functional anatomy of intra- and cross-modal lexical tasks. *NeuroImage* 16(1): 7-22.

Bostock, E., Muller, R. U., & Kubie, J. Experience-dependent modifications of hippocampal place cell firing. *Hippocampus*, 1, 1991, 193–206.

Brachman R, Levesque H. *Knowledge Represent. & Reasoning*, Morgan-Kaufmann, 2004.

Brown C & Hagoort P. *Neurocognition of Language*, Oxford Univ. Press, 1999.

Caplan, D. (2003). Aphasic syndromes. *Clinical Neuropsychology*. K. Heilman and E. Valenstein. New York, Oxford University Press**:** 14-34.

Cohen, J., Aston-Jones, G., and Gilzenrat, M. (2004). A systems-level perspective on attention and cognitive control. *Cognitive Neuroscience of Attention*. M. Posner, Ed. Guilford.

Courtney, S. M., Ungerleider, L. G., Keil, K., & Haxby, J. V. (1996). Object and spatial visual working memory activate separate neural systems in human cortex. *Cerebral Cortex, 6, 39-49.*

Cressant, A., Muller, R., & Poucet, B. Remapping of place cell firing pattern after maze rotations. *Exp. Brain. Res.,* 143, 2002, 470–479.

Contreras-Vidal JL & Gold DR. (2004) Dynamic estimation of hand position is abnormal in Parkinson's disease. *Parkinsonism and Related Disorders,* 10(8):501-506.

Contreras-Vidal JL & Kerick S. (2004). Independent component analysis of dynamic brain responses during visuomotor adaptation. *Neuroimage*. 21(3): 936-945

Contreras-Vidal JL & Schultz S. (1999). A predictive reinforcement model of dopamine neurons for learning approach behavior. *J Comput Neurosci.* 6(3):191-214.

Contreras-Vidal JL, Grossberg S, Bullock D (1997) A neural model of cerebellar learning for arm movement control: cortico-spino-cerebellar dynamics. *Learning and Memory*, 3(6):475-502.

Contreras-Vidal J & Buch E (2003). 'Effects of Parkinsons disease on visuo-motor adaptation'. *Exp Brain Res* 150:25–32.

Contreras-Vidal J & Wen J (2003). Functional Activation, Proc Intl Graph. Soc., 72-76.

Dijksterhuis A et al. On making the right choice: the deliberation-without-attention effect, *Science,* 311, 2006, 1005-1007.

Damasio, A. (1996). The somatic marker hypothesis and the possible functions of prefrontal cortex. *Philosophical Transactions of the Royal Society of London*, 251, 1413-1420.

Dounskaia N (2005) The internal model and the leading joint hypothesis: implications for control of multi-joint movements. *Exp Brain Res* 166:1-16.

Doyon, J, Penhune V, Ungerleider LG. (2003). Distinct contribution of the cortico-striatal and cortico-cerebellar systems to motor skill learning. *Neuropsychologica* 41:252–262.

Dronkers, N., D. Wilkins, et al. (2004). Lesion analysis of the brain areas involved in language comprehension. *Cognition* 92: 145-177.

Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, 19, 1273-1302.

Funahashi, S. & Kubota, K. (1994). Working memory and prefrontal cortex. *Neurosci.Res.*, 21, 1-11.

Ghallib M, Nau D & Traverso P. *Automated Planning,* Morgan-Kaufmann, 2004.

Gibbons A. The Brain's Energy Crisis, *Science*, 280, 1998, 1345-7.

Goldman-Rakic, P. S. (1995). Cellular basis of working memory. *Neuron*, 14, 477-485.

Grosse Wentrup M & Contreras-Vidal JL. (2006). The role of the striatum in adaptation learning: A computational Model. Submitted to *Biological Cybernetics.*

Grundstrom E & Reggia J. Learning Activation Rules, *Int. J. Neural Sys*, 7, 1996, 129-47.

Guigon E & Baraduc P (2002). A neural model of perceptual-motor alignment. *J Cogn*

Neurosci 14:538–549.

Hoover J, Strick P (1993) Multiple output channels in the basal ganglia. *Science*. 259:819-821.

Husain FT, Tagamets MA, Fromm SJ, Braun AR, Horwitz B (2004) Relating neuronal dynamics for auditory object processing to neuroimaging activity: a computational modeling and an fMRI study. *NeuroImage*, 21**:** 1701-1720.

Imamizu H, Miyauchi S, Tamada T, et al. (2000) Human cerebellar activity reflecting an acquired internal model of a new tool. *Nature* 403(6766):192-5.

Jueptner M, Weiller C (1998) A review of differences between basal ganglia and cerebellar control of movements as revealed by functional imaging studies. *Brain* 121:1437–1449.

Kagerer FA, Contreras-Vidal JL, Stelmach GE (1997) Adaptation to gradual as compared with sudden visuo-motor distortions. *Exp Brain Res* 115:557–561.

Krakauer JW, Ghilardi MF, Mentis M, Barnes A, Veytsman M, Eidelberg D & Ghez C (2004) Differential cortical and subcortical activations in learning rotations and gains for reaching: A PET Study*, J of Neurophysiology*, 91:924-933.

Kawato M, Wolpert D (1998) Internal models for motor control. *Novartis Found Symp* 218:291–304.

Kagan J & Baird A. Brain and Behavioral Development During Childhood, in *The Cognitive Neurosciences III,* M. Gazzaniga (ed.), MIT Press, 2004, 93-103.

Lever, C., Willis, T., Cacucci, F., Burgess, N.,&O'Keefe, J. (2002). Long-term plasticity in hippocampal place-cell representation by environmental geometry. *Nature*, 416, 90–94.

Markram H, Luebke J, et al. Regulation of Synaptic Efficacy by Coincidence of Post-synaptic aps and epsps, *Science*, 275, 1997, 213-215.

McIntosh, A. R. (1998). Understanding neural interactions in learning and memory using functional neuroimaging. *Ann.N.Y.Acad.Sci.*, 855, 556-571.

McIntosh, A. R., Chau, W. K., & Protzner, A. B. (2004). Spatiotemporal analysis of event-related fMRI data using partial least squares. *NeuroImage*, 23, 764-775.

McIntosh, A. R. & Gonzalez-Lima, F. (1994). Structural equation modeling and its application to network analysis in functional brain imaging. *Human Brain Mapping*, 2, 2-22.

McIntosh, A. R. & Lobaugh, N. J. (2004). Partial least squares analysis of neuroimaging data: applications and advances. *NeuroImage*, 23 Suppl 1, S250-S263.

Mechelli, A., Price, C. J., Noppeney, U., & Friston, K. J. (2003). A dynamic causal

modeling study on category effects: bottom-up or top-down mediation? *J.Cogn Neurosci.*, **15**, 925-934.

Mesulam M, Large-Scale Neurocognitive Networks, *Ann Neurol*, 28, 1990, 597-613.

Middleton FA, Strick PL. (2000) Basal ganglia and cerebellar loops: motor and cognitive circuits. *Brain Res Brain Res Rev*. (2-3):236-50.

Miller, E. K., Li, L., & Desimone, R. (1993). Activity of neurons in anterior inferior temporal cortex during a short- term memory task. *Journal of Neuroscience*, 13, 1460-1478.

Mountcastle V. *The Cerebral Cortex*, Harvard Univ. Press, 1998.

Neville H & Bavelier D. Specificity and Plasticity in Human Neurocognitive Development, *The New Cognitive Neurosciences,* M. Gazzaniga (ed.), MIT Press, 2000, 83-98.

O'Keefe, J.,&Nadel, N. (1978). *The hippocampus as a cognitive map*. Oxford: Clarendon Press.

Passingham R, Stephan K & Kotter R. The Anatomical Basis of Functional Localization in the Cortex, *Nature Reviews Neuroscience*, 3, 2002, 606-616.

Penny, W. D., Stephan, K. E., Mechelli, A., & Friston, K. J. (2004). Modeling functional integration: a comparison of structural equation and dynamic causal models. *NeuroImage*.

Poeppel, D. and G. Hickok (2004). Towards a new functional anatomy of language. *Cognition* **92**: 1-12.

Rao R & Sejnowski T. Predictive Learning of Temporal Sequences in Recurrent Neocortical Circuits. In Solla S et al (eds.), *Advances in Neural Information Processing Systems*, MIT Press, 12, 2000, 164-171.

Reggia J et al. Competitive Distribution in Neocortex, *Neural Comp.,* 4,1992, 287-317.

Reggia J, Goodall S, & Shkuro Y. Computational studies of lateralization of phoneme sequence generation, *Neural Computation,* 10, 1998, 1277-1297.

Reggia, J., S. Goodall, et al. (2001). The callosal dilemma: Explaining diaschisis in the context of hemispheric rivalry via a neural network model. *Neurological Research* **23**: 465-471.

Riedmiller, M. and H. Braun (1993). A direct adaptive method for faster backpropagation learning: the RPROP algorithm. *Proceedings of the IEEE Conference on Neural Networks*.

Rosenbloom P, Laird J & Newell A, *The Soar Papers*, MIT Press, 1993.

Russell S & Norvig P, *Artificial Intelligence*, Prentice Hall, 2003.

Sanner R, & Kosha M (1999). A Mathematical Model of the Adaptive Control of Human Arm Motions. *Biological Cybernetics*, 80:369-382

Schacter, D. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology, Learning Memory and Cognition*, 13, 501-518.

Schulz, R. and J. Reggia (2004). Temporally asymmetric learning supports sequence processing in multi-winner self-organizing maps. *Neural Computation* **16**(3): 535-561.

Sergent, J., Ohta, S., & Macdonald, B. (1992). Functional neuroanatomy of face and object processing: A positron emission tomography study. *Brain*, **115**, 15-36.

Shadmehr R, Mussa-Ivaldi F (1994). Adaptive representation of dynamics during learning of a motor task. *J Neuroscience* 14:3208-3224.

Shkuro, Y., M. Glezer, et al. (2000). Interhemispheric effects of simulated lesions in a neural model of single word reading. *Brain and Language* **72**: 343-374.

Shkuro Y & Reggia J. Cost During Evolution…, *Cognitive Sys Res*, 4, 2003, 365-83.

Sober SJ, Sabes PN. (2003) Multisensory integration during motor planning. *J Neurosci*, 23(18), 6982-6992.

Sowa J. *Knowledge Representation,* Brooks/Cole, 2000.

Sutton R & Barto A. *Reinforcement Learning*, MIT Press, 1998.

Tagamets MA, Horwitz B (1998) Integrating electrophysiological and anatomical experimental data to create a large-scale model that simulates a delayed match-to-sample human brain imaging study. *Cereb. Cortex*, 8**:** 310-320.

Tagamets M & Horwitz B. A model of working memory, *Neural Networks*, 13, 2000, 941-952.

Tagamets, M. A. & Horwitz, B. (2001). Interpreting PET and fMRI measures of functional neural activity: the effects of synaptic inhibition on cortical activation in human imaging studies. *Brain Res.Bull.*, **54**, 267-273.

Tanaka, K. (1993). Neuronal mechanisms of object recognition. *Science*, **262**, 685-688.

Thoroughman K and Shadmehr R (2000). Learning of action through adaptive combination of motor primitives. *Nature*, 407: 742-747.

Tinnirella M, Tagamets M, Weems S, Contreras-Vidal J and Reggia J. *A Behavior-to-Brain Map,* CS-TR-4803/UMIACS-TR-2006-24, University of Maryland, 2006.

Tulving, E. and Markowitsch, H. (1997). Episodic and declarative memory: role of the

hippocampus. *Hippocampus*, 8(3), 198-204.

Turing A. Computing Machinery and Intelligence, *Mind*, 59, 1950, 433-460.

Ungerleider LG, Mishkin M (1982) Two cortical visual systems. In: Ingle DJ, Goodale MA, and Mansfield RJW, eds. *Analysis of Visual Behavior*. Cambridge, MA: MIT Press, 549-586.

Uttal W. *The New Phrenology*, MIT Press, 2001.

Vouloumanos, A. and J. Werker (2004). Tuned to the signal: The privileged status of speech for young infants. *Developmental Science* 7(3): 270-276.

Weems, S. and J. Reggia (2004). Hemispheric specialization and independence for word recognition: A comparison of three computational models. *Brain and Language* 89: 554-568.

Weems S & Reggia J. Simulating single word processing in the classic aphasia syndromes based on the Wernicke-Lichtheim-Geschwind Theory, *Brain and Language*, 2006, in press.

Winder R, Cortes C, Reggia J & Tagamets M. A learning method for matching experimental fMRI to a model of visual working memory, 2006, under review.