Integrating the Brain Parts into a Coherent Cognitive Architecture

Randall C. O'Reilly

Department of Psychology and Neuroscience

University of Colorado Boulder

345 UCB

Boulder, CO 80309

randy.oreilly@colorado.edu

May 15, 2015

Supported by: ONR grants N00014-13-1-0067, D00014-12-C-0638, N00014-14-1-0670, and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract number D10PC20021. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained hereon are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the U.S. Government. R. C. O'Reilly is CTO at eCortex, Inc., which may derive indirect benefit from the work presented here.

Abstract

This chapter builds on David Marr's original ideas about how the neocortex and hippocampus functioned separately, and in their mutual interaction, to develop a large-scale biologically-based cognitive architecture. This architecture also includes a critical role for the prefrontal-cortex and basal ganglia systems, which provide an overall control function, based on the ability to maintain tasks, goals, and other such information in an active state robustly over time, and the ability to update this information dynamically under influence of reward-based learning, to guide cognitive functioning over time. Detailed computational models of these brain systems and their interactions have been developed, and shown to support complex cognitive function. Lessons from these initial models are enumerated, which shape the ongoing development of new biologically-based computational models that should support even more powerful cognitive abilities.

Keywords: computational models, hippocampus, neocortex, prefrontal cortex, basal ganglia, neural networks, computational cognitive neuroscience

Introduction

David Marr's early theoretical work established a paradigm of analyzing the neural anatomy of different brain areas, and advancing computational theories for how the biology supports useful functions. The trio of classic papers, Marr (1969, 1970, 1971), showed how the cerebellum, neocortex, and hippocampus (archicortex) could support different learning and processing functions, which made sense of their known importance for those functions. My own work has followed a similar approach, starting with computational modeling of the hippocampus and neocortex (O'Reilly & McClelland, 1994; McClelland, McNaughton, & O'Reilly, 1995; O'Reilly, 1996, 1998), and then focusing on the unique properties of the prefrontal cortex and basal ganglia system (O'Reilly, Braver, & Cohen, 1999; Frank, Loughry, & O'Reilly, 2001; O'Reilly & Frank, 2006; O'Reilly, 2006). Having all of these brain models available, the obvious next step is to integrate them into a single brain-scale model, which though easier said than done, is indeed what we have been working on in the past several years (OReilly, Hazy, & Herd, 2015; Herd, Krueger, Kriete, Huang, & O'Reilly, 2013; Ziegler, Chelian, Benvenuto, Krichmar, O'Reilly, & Bhattacharyya, 2014). This chapter attempts to trace the links between these individual brain models and Marr's original ideas, and then provides an overview of our work on integrating these brain parts together into a unified biologically-based *cognitive architecture*, including a number of current directions that we're currently exploring.

In re-reading the original papers on the theory of neocortex (Marr, 1970) and hippocampus (Marr, 1971) in the light of the considerable modern development of the central principles developed in these papers, one is struck by how much of the most important, core ideas Marr was able to articulate, and even formalize in fairly rigorous mathematical terms. Presaging his later focus on information processing based approaches to cognition, his paper on the neocortex is based on information theoretic principles, and the essential insight that the job of the neocortex is to develop an efficient language for encoding all that we experience. Specifically, this internal representational language involves the development of concepts or categories that apply across a wide range of different memories, and thus provide a more efficient way of recoding these memories compared to some kind of raw storage of sensory inputs. He anticipates the considerable work on self-organizing category formation (e.g., by Kohonen, 1977, 1982) in developing the

notion of clusters or "mountains" of probability that drive the formation of new concepts. He is very astute in recognizing that there is no "free lunch" set of prior biases for such a learning system (Geman, Bienenstock, & Doursat, 1992), and that our neocortex must be wired to take advantage of the empirical statistics of our particular world. After developing this computational-level information-processing framework, he then considers how neural hardware can implement it, based on the basic unit of an R-theta codon, which is just a thresholded integrate-and-fire neuron in essence. The comprehensive range of considerations (including consideration of some Bayesian principles) is truly impressive.

The second paper on hippocampus starts to paint a more complete cognitive architecture involving the dynamic interactions of neocortex and hippocampus, in a manner that remains essentially current. Specifically, the hippocampus is envisioned as a sparse, rapid-learning system for taking "snapshots" of the state of the neocortical system, and holding on to these *simple memories* long enough for the relevant information to be incorporated (consolidated) into the neocortical system. These same principles were elaborated in our widely-cited paper (McClelland et al., 1995), in light of considerable additional data and computational modeling results.

This consideration for the larger structure of the full neural information processing system of the brain provides an early indication of the importance of this process of thinking about the larger cognitive architecture and the ways in which the different brain systems interact with each other to actually perform complete cognitive functions. One of most important aspects of the way my colleagues and I have approached this problem also owes a strong debt to Marr's legacy (Marr, 1982): we are working on models at multiple levels of abstraction, and attempting to synthesize across the insights from these different levels of analysis. Specifically, we have a long-standing collaboration with the developers of the ACT-R cognitive architecture, which operates at a symbolic and subsymbolic level, to develop a synthesis of the key features of that architecture and those of our own cognitive architecture (Jilk, Lebiere, O'Reilly, & Anderson, 2008). Thus, our work spans the full range of Marr's classic levels of analysis (computational, algorithmic, and implementational), and, like his early papers, seeks to find those ideas that have strong convergent support across levels. Although some have taken Marr's levels as a justification for ignoring the more detailed properties of the biology, we instead find that biological details often strongly constrain and inspire computational ideas, just as they did for Marr in his early papers.

The next section provides a brief overview of the core ideas behind our biologically-based cognitive architecture, and how it connects with some of Marr's original ideas. Then, we describe our first attempts to create large-scale integrated models of all of these interacting brain systems. Finally, new directions building upon this initial work are described, including considerable in-progress ideas about how to take these models to the next level of functionality.

Large-Scale Biologically-Based Cognitive Architecture

Insert Figure 1 about here.

One of the most essential and enduring contributions of Marr's work was the notion that one could look at the neuroanatomy of a given brain area, and, when equipped with appropriate neuro-computational and information processing principles, deduce the major function of that area. In continuing this approach, my colleagues and I have elaborated the cognitive architecture diagrammed in Figure 1 (O'Reilly & Munakata, 2000; O'Reilly, Munakata, Frank, Hazy, & Contributors, 2012; OReilly et al., 2015). This architecture includes the neocortex and hippocampus as envisioned by Marr, plus the prefrontal cortex and basal ganglia system, which has unique specializations that enable it to support essential cognitive functions that cannot otherwise be supported by the other two systems. In the following sections, the major features of the separable components of these systems are elaborated.

Posterior Neocortex

Consistent with Marr's essential insights (Marr, 1970), the posterior neocortex in this architecture supports learning mechanisms that enable the development of powerful, efficient representations of the external world, going from perception to semantics. While Marr's ideas here focused on somewhat symbolic-level *codon* representations that learned to encode concepts or categories, my own approach has been shaped by the subsequent work in the PDP framework (Rumelhart, McClelland, & the PDP Research Group, 1986; McClelland, Rumelhart, & the PDP Research Group, 1986) emphasizing the importance of overlapping distributed representations, and the power of the error backpropagation learning mechanism.

The very tricky (and essentially intractable) problems of when to cleave off a new concept representation or not that Marr wrestled with in Marr (1970) are nicely managed by these two principles.

First, with distributed representations, a pluralistic approach to concepts and categories can be taken, where many different ways of carving up the world are applied *in parallel*. Thus, the strong tradeoffs associated with any one choice are mitigated. Second, the virtue of error-driven learning is that it shapes representations across multiple levels in a hierarchy to encode information in whatever way is necessary to solve the problems at hand. Thus, instead of attempting to articulate, and implement, a sufficient set of broad *a priori* constraints on the nature of representations that should be generally useful, error-driven learning instead just learns those representations that are actually important in solving the problems that the organism needs to solve. This circumvents any number of vexing problems that have continued to plague those who have persisted in Marr's original mission of using Hebbian-like self-organizing learning mechanisms to develop internal representations in the brain. None of these Hebbian-based models have ever demonstrated the ability to solve truly challenging computational problems, whereas backpropagation has recently become re-appreciated for its considerable abilities to solve important problems such as object recognition better than any other competing algorithm (e.g., Ciresan, Meier, Gambardella, & Schmidhuber, 2010; Ciresan, Meier, & Schmidhuber, 2012; Krizhevsky, Sutskever, & Hinton, 2012; Bengio, Courville, & Vincent, 2013).

The standard objection to error backpropagation from a biological perspective is that it doesn't fit with how the brain actually works (Crick, 1989). However, we have developed increasingly biologically-based error-driven learning mechanisms, starting with the central insight that bidirectional excitatory connections can convey error gradients in a fully biologically realistic manner (O'Reilly, 1996). More recently, we have been able to derive a local learning rule from a highly detailed and strongly empirically constrained model of spike-timing dependent plasticity (STDP) (Urakubo, Honda, Froemke, & Kuroda, 2008), which directly implements the temporal contrast that is required to perform error-driven learning based on bidirectional excitatory connections (O'Reilly et al., 2012). Thus, we do not see biological plausibility as a valid reason for excluding the adoption of computationally powerful error-driven learning in the neocortex anymore.

Insert Figure 2 about here.

There are nevertheless some problems associated with pure error-driven learning – it has very weak overall learning biases, and can often suffer from overfitting and generally be underconstrained by any given learning problem. Indeed, the recent advances in improving the performance of backpropagation networks have largely come from throwing much larger data sets at it, and using a few other tricks to further constrain the learning, importantly including reducing the number of synaptic weights by re-using a shared set of them across redundant pools of units within a layer – a trick known as *convolution*.

Our own approach to these issues has been inspired by the same biological properties of neocortex that inspired Marr's work: extensive networks of inhibitory interneurons, and Hebbian-like learning principles. Specifically, we include strong lateral inhibition in our models that produces *sparse distributed representations* that approximate a k-winners-take-all (kWTA) function (O'Reilly, 1998). The benefits of sparse coding have long been appreciated in multiple domains (e.g., Kanerva, 1988; Földiák, 1990; Olshausen & Field, 1996, 1997). Indeed, the recent work on the dropout regularizer for backpropagation networks (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) can be seen as achieving a similar objective as sparse coding: minimizing the number of neurons that are involved in representing any one thing to break up large-scale fragile interdependencies among neurons, while also preserving some redundancy in the distributed code. In our models, we also include a weak level of Hebbian learning, based on the Bienenstock-Cooper-Munro (BCM; (Bienenstock, Cooper, & Munro, 1982)) mechanism that emerges from the STDP learning models (Blair, Intrator, Shouval, & Cooper, 1998; Shouval, Wang, & Wittenberg, 2010), which introduces beneficial self-organizing learning constraints that act as a regularizer on error-driven learning (O'Reilly, 1998, 2001; O'Reilly & Munakata, 2000; O'Reilly et al., 2012). At the lowest level, we use well-validated spiking or rate-code neuron activation dynamics based on the AdEx model of Brette and Gerstner (2005).

Figure 2 summarizes the full set of basic neocortical learning principles as embodied in the *Leabra* learning algorithm, which stands for *Local, Error-driven and Associative, Biologically Realistic Algorithm*, and is pronounced like "Libra", which provides metaphorical inspiration in terms of striving to strike an

appropriate balance between many different competing forces and considerations in the construction of a coherent framework for cognitive modeling. This Leabra model has been successfully applied to a wide range of cognitive tasks (O'Reilly & Munakata, 2000; O'Reilly et al., 2012), including challenging object recognition problems (O'Reilly, Wyatte, Herd, Mingus, & Jilk, 2013). Below, more recent developments to this core set of principles are elaborated, that incorporate more of the detailed circuitry of the neocortex, including the division between superficial and deep layers, and the interconnections with the thalamus.

Hippocampus

My work on the hippocampus carries forward many of the core ideas developed by Marr (1971), including the pattern-separation abilities of the *R-theta codon* (O'Reilly & McClelland, 1994), and the nature of the interplay between hippocampal and neocortical learning mechanisms, emphasizing the ability of information initially encoded in the hippocampus to be then integrated into the neocortical system (McClelland et al., 1995). In subsequent work with Jerry Rudy, we attempted to further elaborate the core principles that allow one to clearly predict when the hippocampus would be essential for a given task (O'Reilly & Rudy, 2001), and tested these ideas in a number of experiments with rats, largely using the contextual fear conditioning paradigm (e.g., Rudy & O'Reilly, 2001, 1999). Overall, we find the core idea articulated by Marr (1971), that the hippocampus is a fast learning system that uses extremely sparse coding to encode memories in as distinct and non-overlapping a manner as possible, to remain the essential unique feature of the hippocampal system.

Insert Figure 3 about here.

One major development relative to Marr's original ideas is in the way that the hippocampal memory system supports recall or re-activation of the memory back down in the neocortex. Marr envisioned this occurring directly on synapses projecting back into all areas of the neocortex from the hippocampus. Instead, we have emphasized that the role of the CA1 is to learn a *sparse invertible mapping* between the input projections in the entorhinal cortex (EC), so that the sparse, pattern-separated memory encoding in area CA3 can be directly *decoded* back into the language of the neocortex, before going back

down through existing, previously-learned bidirectional pathways from EC out into the broader neocortex (Ketz, Morkonda, & O'Reilly, 2013). This eliminates the need for massive synaptic plasticity out in the neocortex for every new memory encoded by the hippocampus, and really enables the system to live up to the function of rapid low-interference simple memory encoding. Interestingly, by appreciating the critical role of the CA1, it becomes clear why the synapses between the CA3 and CA1 are actually the most important locus of new memory formation, which has been demonstrated in recent experiments.

In the next major section, the interplay between hippocampus and neocortex, along with the prefrontal cortex / basal ganglia system, will be elaborated further.

Prefrontal Cortex and Basal Ganglia

Insert Figure 4 about here.

The prefrontal cortex (PFC) shares many anatomical properties in common with posterior neocortex, and in many ways we think it operates according to the same general kinds of computational and biological mechanisms as all of neocortex. However, it is now well established that neurons in the PFC are uniquely capable of sustained active firing, in the face of distraction and other intervening processing – commonly referred to as *working memory* (Fuster & Alexander, 1971; Goldman-Rakic, 1987; Baddeley, 1986; Miyake & Shah, 1999). We have developed the idea that a critical element of what allows the PFC to support this robust working memory ability is its intimate interconnection with the basal ganglia (BG), which supplies a critical *adaptive gating* function to determine when to update vs. maintain information in PFC (Frank et al., 2001; O'Reilly & Frank, 2006; O'Reilly, 2006). The BG in turn learns how to perform this adaptive gating function based on its dopaminergic modulation and associated learning mechanisms (Frank, 2005; O'Reilly & Frank, 2006). We refer to this overall framework as the *PBWM* (prefrontal cortex, basal ganglia working memory) model.

We have recently shown that single PBWM model can learn 9 different widely-studied executive function tasks that have been strongly associated with PFC / BG function empirically (Friedman, Herd, Hazy, Chatham, Kriete, Brant, & O'Reilly, submitted). Thus, we believe that this framework has the

capacity to explain how dynamic updating of PFC-mediated active maintenance can enable the system to juggle information in the service of performing complex cognitive tasks.

Interactions Among the Systems

Insert Figure 5 about here.

The above brain systems provide a minimal set of functionality necessary to account for a wide range of cognitive function. In broad form, the PFC/BG system can drive top-down attentional biasing (O'Reilly et al., 1999; Miller & Cohen, 2001) of processing in the posterior neocortex and hippocampus, to keep these areas focused on the task at hand. The powerful overlapping distributed representations in the posterior neocortex support complex inference abilities, including multiple constraint satisfaction mediated by bidirectional excitatory projections. The hippocampus in turn is constantly and rapidly encoding simple-memory snapshots of the state of the system, and in response to partial cues (provided in part from top-down PFC inputs), recalling relevant information to support the ongoing information processing flow.

One very important source of support that this overall dynamic among these three systems is indeed sufficient to account for a wide range of cognitive function comes from the ACT-R cognitive architecture (Anderson, Bothell, Byrne, Douglass, Lebiere, & Qin, 2004), which can be mapped directly onto our biologically-based cognitive architecture (Jilk et al., 2008) (Figure 5). ACT-R, which combines both symbolic and subsymbolic processing elements, centers around a production system that is mapped onto the basal ganglia and frontal cortex, interacting with a declarative memory system that combines the functions of the posterior neocortex and the hippocampus. The essential idea behind the production system is that the basal ganglia evaluates all of the possible courses of action available at the given moment of time (i.e., all the productions that have matched their conditions and could possibly fire), and selects the one with the best history of prior reward utility. When a production then fires, it updates the contents of a *buffer*, which we associate with the contents of working memory in different parts of the PFC. This buffer update then triggers further processing in other parts of the posterior neocortex (e.g., motor, visual, etc), which then triggers a new set of possible productions, and so on.

The ACT-R framework has been applied to a very large number of different cognitive phenomena `http://act-r.psy.cmu.edu/publication/`, and thus provides a very capable level of abstraction above the neural level of analysis we take with our models (Jilk et al., 2008). One of the most important differences between the two levels is that it is relatively straightforward to directly program an ACT-R model to accomplish a given task, whereas our models must be trained somehow from the ground up. Thus, we think of the ACT-R framework as a kind of rapid prototyping language for figuring out the scope and essential cognitive processes involved in a given cognitive task, which then makes it easier to develop a corresponding biologically-based model in the Leabra framework.

Initial Large Scale Models

Insert Figure 6 about here.

Over the past several years, we had the opportunity to finally put together all of the brain systems described above into single integrated large-scale models, in the context of the ICArUS project, which was focused on developing biologically-based cognitive models of the sensemaking process and the origin of various cognitive biases that arise in this process (Figure 6) (Herd et al., 2013; Ziegler et al., 2014). This project provided the necessary incentives to manage the considerable challenges in producing this large-scale integration for the first time. A large team of collaborators were successful in producing a series of models that truly demonstrated how the kinds of interactive dynamics described above can actually play out in a working model. The perceptual system of the model processed geospatial input displays, and the simulated parietal cortex extracted relevant spatial and numerical information from these displays. This information was then encoded into active PFC working memory representations, and also encoded into hippocampal snapshots, which then drove the subsequent processing and decision-making steps to more fully analyze and interpret the situation at hand. Throughout, parallel ACT-R models were developed to prototype and guide the construction of our large-scale biological models, as described above.

Aside from the specific details of our models and the various ways in which we were able to account for particular sensemaking phenomena and cognitive biases, the most important lessons derived

from this effort are driving our next generation of cognitive models, which we elaborate in the next section. For these first pass models, we resorted to a very pragmatic approach of training individual parts of the model separately, and then integrating them under the guidance of a set of programmed steps that moved the processing along from one step to the next. An unfortunate amount of the overall intelligence of the system ended up being carried by these "outer loop" guiding programs. Thus, our major goal for the next generation of such models is to develop robust neural mechanisms that can learn to control the overall flow of processing in the system, so that they can truly become autonomous information processing systems that are in control of their own cognitive functionality.

Nevertheless, there was still considerable room for emergent complexity in these models – the constant hippocampal encoding and recall dynamics for example produced an overall anchoring bias to focus on information presented earlier in time. And we developed improved versions of all of our component models (e.g., Ketz et al., 2013), including a vastly improved PBWM model that is just being finalized and will be written up soon.

Toward a Biologically-based Autonomous Cognitive System

Here is a list of the most important lessons we learned from our large-scale integrated models, that provide the action items for our new framework that is still under construction, as described in the next section. As you may notice, many of these elements of functionality represent features of more symbolic information processing architectures, including ACT-R. Indeed, one of our longstanding goals is to develop a Synthesis of ACT-R and Leabra (*SAL*) architecture that integrates the best of both architectures (Jilk et al., 2008). Thus, the overall challenge is to leverage the powerful distributed representations and learning mechanisms of the Leabra neural-network framework, while still achieving the critical high-level symbolic-like processing abilities that truly distinguish human level cognition.

- While our basic Leabra model of the posterior neocortex excels at constraint-satisfaction processing of individual input patterns through attractor-based bidirectional excitatory dynamics, it consequently lacks the ability to transition effectively between different, but related, cognitive states. We have in general studiously avoided this issue of temporal dynamics in the models, because such

dynamics can end up being complex, brittle and generally difficult. However, we recently realized that the venerable simple recurrent network (SRN) framework (Elman, 1990) provides a good computational model of temporal processing in the neocortex (O'Reilly, Wyatte, & Rohrlich, 2014c), and this allows us to apply powerful error-driven learning to shape the temporal dynamics of our models in an effective way, avoiding much of the parametric complexity of the dynamics problem. This is a similar transition to that between designing complex self-organizing biases into a model, versus using a powerful general-purpose learning mechanism, as discussed above.

- Complex cognitive processing is an *articulated* process that happens over time, and requires specific cognitive functions to be applied at specific points in time. In addition to temporal dynamics, achieving the necessary specificity of processing requires more capable *attentional* dynamics in the model. We need to be able to "shine a spotlight" on a particular cognitive function and have it then work on the relevant information coming from other parts of the cognitive system. In contrast, our existing models are excessively diffuse in their processing – activation spreads liberally throughout the network, typically engaging all parts of the model in every single step of cognitive processing. This results in high levels of interference and general confusion. A playful pejorative term for this problem is *connectoplasm* – undifferentiated blobs of oozing network goo, that lacks the rigid skeleton and infrastructure necessary to accomplish complex cognitive tasks.

- Another critical element of articulated cognitive processing is that some level of localized functionality needs to emerge within the system, so that attention can have something discrete and specific to focus on in the first place. In our existing models, information processing occurs in a highly diffuse, distributed manner within so-called *hidden* layers that compute transformations over input patterns to produce specific output patterns – it is very difficult for an external system (e.g., the PFC) to provide precise control over such a system, because everything is so diffuse and distributed. This problem is typical of most neural network systems. We certainly don't want to go to the other extreme of rigid, encapsulated modularity (Fodor & Pylyshyn, 1988), but rather seek a fertile middle ground (O'Reilly, Petrov, Cohen, Lebiere, Herd, & Kriete, 2014b).

- The dynamic binding of specific information for a particular functional role is another important challenge for all information processing systems, and particularly so for neural networks, with their slowly modifying synaptic connections making it difficult to rapidly update a representation to transiently encode a new binding. One might be tempted to consider rapidly adapting synaptic weights as a solution here, but there is a deeper problem: even if you can rapidly change synaptic weights, it is not clear that downstream neurons would be able to make any sense of the new information encoding (OReilly, 2010). This is because all neurons receive an extremely impoverished type of information – they cannot use any kind of referential language to communicate with each other – there are just spikes that just excite the receiving neuron to a graded extent. Thus, learning over time is essential for other neurons to be able to effectively use any information being sent by a given neuron. There is no escaping the need for accumulating sufficient statistics to interpret what a given neuron is encoding. The natural conclusion from this chain of reasoning is that the neocortex must somehow develop over time stable, structural mechanisms for solving the dynamic binding problem.

With these considerations in mind, we now elaborate the outlines of new biologically-based extensions to the existing Leabra framework that can potentially achieve these critical functional desiderata.

DeepLeabra: Deep Neocortical Layers and Thalamocortical Loops

Biologically, the key insight for how we need to extend the existing Leabra mechanisms, is to recognize that these existing mechanisms provide a reasonable model of the essential functionality of the superficial (supergranular) neocortical layers (layers 2/3 receiving input from layer 4), but they are entirely missing the extensive functions of the deep (infragranular) neocortical layers (layers 5 and 6), and the interconnectivity with the thalamus. The superficial pyramidal neurons are broadly, bidirectionally interconnected to other neurons within the same area, and in other areas, and seem to produce the same kind of more distributed processing and representations as seen in our existing models. These superficial networks can provide the genesis of new representations and support powerful error-driven learning, as captured in our existing Leabra models, but it is the deep layers and the thalamus that provide the metaphorical skeleton upon which all of this superficial connectoplasm rests.

Insert Figure 7 about here.

Thus, our new framework is called *DeepLeabra*, because it adds the functionality associated with these deep neocortical layers to the existing Leabra framework (Figure 7). In addition, some of the mechanisms facilitate learning in deep hierarchical networks. We hypothesize that these deep layer neurons support three interrelated but separable computational functions: temporal integration of information over time, attentional modulation of processing guided by both bottom-up and top-down signals, and an auto-encoder based learning mechanism that leverages predictive learning over time. Elements of our new framework are similar to those in other existing models, but the full package of mechanisms in the DeepLeabra framework is unique, and we argue provides a uniquely compelling account of a wide range of biological data. For example, in their extensive analysis of the feedforward (FF) and feedback (FB) circuits of the neocortex, Markov, Vezoli, Chameau, Falchier, Quilodran, Huissoud, Lamy, Misery, Giroud, Ullman, Barone, Dehay, Knoblauch, and Kennedy (2014) concluded that all the extant models failed to account for the available data. In contrast, we argue that our model does accord with this data, and we also elaborate a number of detailed, testable predictions that could be used to invalidate our model.

In overview form, the central claims of the DeepLeabra model are as follows:

- Deep pyramidal neurons, driven directly by the superficial neurons within the same cortical microcolumn and other inputs, produce an attentional modulation of the superficial network that is functionally similar to the abstract model of Reynolds and Heeger (2009), which can account for a wide range of relevant data. The critical elements include the computation of raw attentional modulatory signals, which we associate with deep layer 5b intrinsic bursting (IB) neurons, and a subsequent renormalization of these modulatory signals which requires considerable spatial integration across a given area, before these renormalized signals can then be applied in a multiplicative fashion to the inputs to the superficial layer neurons. The extensive lateral interconnectivity of the 5a and 6a corticocortical (CC) regular-spiking (RS) neurons can provide the necessary spatial integration, which then can drive the layer 6 corticothalamic (CT) neurons to effect the modulation. These 6CT neurons project simultaneously up to layer 4 and down to the thalamic

neurons that then project back up to layer 4 of this same area, and are thus in a position to apply the appropriate multiplicative modulation of both of these pathways for the input signals to the area (and the extensive collateral of the 6CT neurons into the thalamic reticular nucleus likely adds important additional center-surround attentional contrast effects). Consistent with this idea, there is good evidence that these layer 6 neurons provide a multiplicative gain modulation of other cortical layers (Olsen, Bortone, Adesnik, & Scanziani, 2012; Bortone, Olsen, & Scanziani, 2014).

- An important source of attentional control signals derive from feedback projections of higher layers down to lower layers in the cortical hierarchy, and the deep layer CC/RS neurons (predominantly in layer 6 and lower 5) provide just such a projection, which targets both the superficial and deep layer neurons in the lower areas, including their deep5b IB neurons (Thomson & Lamy, 2007; Markov et al., 2014). Thus, these top-down signals can modulate attentional dynamics in lower layers in part by tapping directly into the deep attentional network, producing a direct multiplicative modulation, in addition to driving some attentional modulation directly in the superficial layers. Consistent with other models (Grossberg, 1999; Raizada & Grossberg, 2003; Montijn, Klink, & Van Wezel, 2012), we hypothesize that this direct deep-to-deep modulation is particularly important for effective attentional focusing on task-relevant information, which as noted above has been sorely missing from our existing Leabra models. As emphasized in Markov et al. (2014), the longer-distance top-down projections (i.e., from very high layers to much lower layers in the hierarchy) are almost exclusively of this deep-to-deep form, providing very high-level top-down attentional control. Furthermore, our DeepLeabra model is unique in providing a compelling functional story for the existence of these two separable and yet interdependent FF and FB circuits between areas – the superficial-to-superficial network is essential for constraint satisfaction processing, while the deep-to-deep network is essential for effective top-down and bottom-up attentional modulation.

- The superficial and deep networks operate on different clocks, as evidenced by the ˜40Hz gamma-frequency oscillations of superficial neurons, compared to the ˜10Hz alpha-frequency oscillations of the deep networks (Lorincz, Kekesi, Juhasz, Crunelli, & Hughes, 2009; Franceschetti, Guatteo, Panzica, Sancini, Wanke, & Avanzini, 1995; Buffalo, Fries, Landman, Buschman, & Desimone, 2011; Luczak, Bartho, & Harris, 2013). We hypothesize that this acts like an

Expectation-Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977), where the slower, outer-loop expectation step sets the parameters for the attentional modulation provided by the deep networks, while the faster inner-loop maximization process performs constraint satisfaction maximization within these parameters. The EM algorithm is important for problems of this form where you have two interdependent sets of unknowns, and it is impossible to simultaneously optimize both of them. This represents an important novel computational motivation for these distinct brain rhythms. Biologically, the deep5b IB neurons have a preferred bursting frequency at the alpha rhythm, so we hypothesize that this alpha bursting is the key updating event associated with triggering the new set of attentional parameters, which are then integrated, normalized, and communicated continuously through the deep CC, RS neurons to impose an attentional filter for the remainder of the ˜100msec alpha cycle.

- A critical implication of the EM-like dynamics of the attentional network is that the attentional state lags processing in the superficial network as it moves forward through the subsequent alpha cycle. This lag is exactly what is needed to provide temporal context information to appropriately contextualize the integration of information over time. We have previously developed this specific idea in the form of the LeabraTI (temporal integration) model (O'Reilly et al., 2014c; Kachergis, Wyatte, O'Reilly, Kleijn, & Hommel, 2014; Sun, O'Reilly, Bhattacharyya, Smith, Liu, & Wang, 2015). In computational terms, this temporal context acts just like the context layer in a simple recurrent network (SRN; Elman, 1990), and this temporal context, properly integrated with learning mechanisms, enables a particularly powerful form of temporal integration learning in a variety of tasks. Biologically, a subset of the deep CC RS are the most likely source of this contextual information, which is then broadcast to the superficial layers and other parts of the network. Furthermore, a particular form of the contextual integration can support both this temporal integration function, and the spatial integration needed for the attentional renormalization computation per the Reynolds and Heeger (2009) model.

Insert Figure 8 about here.

- The alpha frequency frame-rate also plays an important role in framing the learning process, which in the Leabra model has been conceived of in terms of an expectation or minus-phase of activation followed immediately by an outcome or plus-phase state, that serves as the target state in the error-driven learning dynamic (O'Reilly, 1996, where the minus-plus terminology derives from the Boltzmann Machine). Although we have a reasonably solid synaptic-level mechanism for this form of error-driven learning, and we know that the bidirectional connectivity in the superficial layer neurons can appropriately communicate backpropagation-style error gradients throughout the network, we have never had a particularly compelling story for where exactly the plus-phase activation states come from, and how they can drive appropriate forms of learning. The alpha bursting of the deep5b IB neurons suggests a novel, powerful solution to this problem: each layer in the network receives a local error-driven learning signal via a *thalamic auto-encoder* operating at the alpha frequency. Specifically, the plus phase signal is the driving activation of the thalamic relay cell (TRC) neurons from lower-level deep5b IB neurons, which contrasts with the immediately prior top-down activation of these TRC neurons via projections from the area itself. When this circuit is appropriately diagrammed (Figure 8), it is clear that each layer acts as an auto-encoder for the information projected from the lower layer(s) in the hierarchy via their deep5b IB transthalamic projections. This kind of recursive auto-encoder architecture has been developed in a large number of different neural network models over the years (Pollack, 1990; Dayan, Hinton, Neal, & Zemel, 1995; Hinton & Salakhutdinov, 2006), and in other Bayesian and Bayesian-inspired generative model frameworks (e.g., Lee & Mumford, 2003; Friston, 2005). Three key differences from these other models are that:

  - The error signal difference between bottom-up ground truth and the generated expectation thereof exists as a temporal difference signal, not as an explicit difference computed between two distinct neural populations, as it is in some models.
  - The bottom-up ground truth is subject to the top-down attentional modulation parameters, and is furthermore transformed into the topographic space of the receiving layer – e.g., V2 does not actually learn to reconstruct all of V1, but rather only learns the attentionally-modulated transformation of V1 that arrives at the TRC neurons projecting to V2.

- The time lagged nature of the deep layer network that drives the TRC neurons ensures that the learning is *predictive* over time – the network is always trying to predict what the *next* deep5b signal will be, based in part on prior contextual information via the TI mechanism described above.

These differences enable the DeepLeabra model to fit better with the available data on neural dynamics associated with FF and FB projections (Markov et al., 2014), and computationally learn to abstract and transform the inputs in ways that would be difficult under the constraint of full reconstruction of the entire input pattern. Thus, consistent with the ART model of Grossberg (1999, 2013), the top-down attentional signals shape the learning process in important ways. However, our model does not require the discretization of a match vs. mismatch process as occurs in the ART model – everything just emerges in distributed representations under the forces of the constraint satisfaction process under the EM-like attentional constraints.

Overall, these mechanisms enable models to process information over time in a much more powerful and selective way than in the basic Leabra framework, while also giving it a much richer and yet more realistic source of learning signals, in the form of predictive auto-encoding error-driven learning. However, these mechanisms do not deal with two important additional considerations: the need for a stable structural organization of representations to deal with binding problems, and the importance of an agentic, goal-driven perspective for truly understanding how the system learns and actually decides what to do. These are discussed in the following two sections.

Grounded, Replicated, Indexed Data (GRID) Theory

In almost every neural model that deals with complex structured information, some kind of slot-like organization is inevitably used, where the different slots contain representations of different entities and their relationships. The advantage of these slots is that they allow the system to represent multiple different entities at the same time, such that other processing layers can then compute relevant relationships or other information integrating over these multiple entities. Furthermore, the role-filler binding for a slot-like structure is very explicit and clear: the slot defines a functional role, and the filler is

whatever pattern of activity happens to be inside the slot at the time. But these slot-based structures are often described apologetically, as some kind of place-holder until some better idea comes along. One important source of aversion to slots comes from the apparent need to replicate these things *ad naseum* as the number of different functional roles one might consider is expanded to include a wider swath of human cognition. Do we really have different slots for all the different things that have been put in slots in all these different models?

Our current thinking about this issue is informed by developments in the ACT-R architecture, which also shares this slot replication problem, in the form of the slots defined within the *chunk types* that hold structured information within the system. Recently Taatgen (2013) was able to create a version of ACT-R that only uses completely *generic* chunk types, with unnamed slots that are just used in an arbitrary convention established by the way the system uses the information. Furthermore, he developed similarly generic productions that handled a large number of common information processing patterns, such that any given task could be handled by the proper sequencing of these generic productions. In so doing, he was able to account for detailed patterns of transfer in human subjects among a number of different tasks, often in fairly non-obvious ways.

Insert Figure 9 about here.

Thus, one potential answer to the slot replication problem is to instead adopt a wide-spread system of generic slots, that are appropriately coordinated throughout the brain (Figure 9). We refer to this as a Grounded, Replicated, Indexed Data (GRID) system, where each slot is grounded in the semantics of a particular brain area, and cross-referenced with corresponding slots in other brain area that encode other aspects of the semantics of that same entity.

This overall architecture is consistent with the arguments in favor of *indexical* representations (Pylyshyn, 2000), which abstract over the detailed features of a given object, and instead encode "pointer like" indexes that refer to a given entity in the environment, without needing to fully describe it. He also refers to these as FINST's or *fingers of instantiation* – a term that has been adopted in the ACT-R framework as well. The virtue of indexical representations is that, like slots, there are small finite number

of such indexes (evidence suggests 3-4), and you can dynamically load these indexes to point to different locations or objects. Having done so, all of your pre-existing computational apparatus can be brought to bear on the contents of these fixed indexical slots. Thus, if you have a set of slots encoding the spatial locations of different objects, and you've learned to be able to fixate and reach to these location-slots, then all you need to do to drive these actions appropriately is update the appropriate slot with the relevant information, and the action from that point on is essentially automatic.

As we extend this notion into more domains, it provides a general recipe for thinking about how information processing can become appropriately articulated and localized – the slot-like structure provides a skeletal framework for organizing and controlling cognitive processing. The dynamic attentional mechanisms supported by the DeepLeabra model can then operate on top of this structural framework, to provide a more fully articulated model of cognition. Functionally, it ends up being very similar to the Taatgen (2013) model, but cast in neural terms. And the replication of slots problem is avoided by extensive re-use of the same slots (i.e., making the slots very generic, as in Taatgen's model), with appropriate further contextualization to determine what kinds of operations to perform on them.

We are currently exploring whether this kind of architecture naturally emerges in the context of naturalistic, grounded sensory-motor learning experiences, along with the DeepLeabra attentional mechanisms and some initial topographic connectivity structures.

Goal-Driven Processing

Finally, we have also been developing more elaborated versions of our PBWM framework to account for the contributions of the orbital frontal cortex (OFC) and anterior cingulate cortex (ACC), in guiding the selection of actions that the system choose to perform. In the course of this work, we began to appreciate the truly pervasive role for goal-driven cognitive processing in all aspects of human cognition (O'Reilly, Hazy, Mollick, Mackie, & Herd, 2014a). Specifically we can divide human mental life into two discrete stages: goal selection and goal engaged processing, and each of these stages has a distinctive value function that strongly suggests that the distinction is real and important. During the goal selection stage, all manner of potential costs and benefits are carefully weighed, and the system's value function is relatively conservative – it operates much like a rational economist might hypothesize. However, during the goal

engaged phase, after a particular goal has been selected and action is now proceeding to achieve that goal, the value function becomes dominated by progress toward that selected goal, and costs and other obstacles can be significantly downweighted.

Some real-life examples provide the best intuitive evidence for these distinct stages. For example, when you're procrastinating in starting some kind of large, relatively effortful tasks (doing taxes, writing a paper, paying the bills, packing for a trip, etc), this reflects the goal selection phase, where the likely true costs of this action plan are being rationally considered. It is actually entirely reasonable for your brain to decide to do shorter, more rewarding tasks, given the basic economics of reward and effort in the brain! Only when the costs of *not* doing these onerous tasks outweighs the costs of doing them do we finally get over threshold. However, once a goal has been engaged, one often finds that the task seems much easier than expected. Furthermore, the dominant currency becomes incremental steps of progress toward the goal. Video games are a prime example where this kind of progress tracking system has been completely hijacked, and people spend far too long doing things that they would never have rationally decided to do, when they were in the goal selection phase. Indeed, this is why goal selection must be so conservative: once the goal has been engaged, it really takes over, and it is difficult to interrupt and abandon that goal – one feels a strong sense of disappointment in doing so.

One can usefully take advantage of the features of these two systems to do the things you want to do, and not the things you don't. For example, when starting a big task, always set a very small, easily achieved initial goal (e.g., opening up the word processor and writing your name as author of the paper), so that then the progress motivation sets in, and you can then bootstrap that into the next subgoal (just write whatever is easiest to write first – don't start at the start!). Likewise, to avoid getting sucked into obsessive tasks or distractions for too long, always try to just take a break and interrupt the cycle of progress – even a few minutes can allow you to disengage and reevaluate your priorities.

Evidence of the central importance of goal-driven cognition is available to any parent: the onset of tantrums at around age 2 is evidence that the young cognitive system is taking control of its own goal selection process, and often this process can end in significant frustration, due to manifest lack of control over most aspects of life at that age. More generally, parents may recognize that when a child is actively engaged in some task, they tend to be much better behaved and happy. When however they are bored and

unable to settle upon something to do, behavior deteriorates rapidly. Our affective life is strongly tied to the goal engagement process, to the extent that depression can be characterized by a persistent inability to achieve goal engagement, while life satisfaction is strongly tied to overall goal achievement.

Our goal is to capture these kinds of core goal-driven dynamics within our computational models, such that the model is actively deciding what it wants to do given a range of options, and it experiences the same kinds of motivational and affective dynamics that sustain performance on engaged goals, while also leading to careful selection of new goals. Initial models have cached out the relevant functional roles of the OFC, ACC, PFC, ventral striatum (including the Nucleus Accumbens), and midbrain dopaminergic pathways involved in the goal selection and goal engaged stages (O'Reilly et al., 2014a). Ongoing work is elaborating these models to produce more complex, dynamic and emergent agentic behavior, and attempting to understand the impact of these goal-driven systems on how the rest of the cognitive architecture develops and functions.

Discussion and Conclusions

Hopefully this chapter has provided a useful sketch of some of the many fascinating issues that emerge in considering the large-scale functional cognitive architecture of the brain, building on the firm foundation that David Marr's pioneering work provided. The incredible complexity of these interacting brain systems can be daunting, and it is often tempting to retreat back into optimizing the independent functioning of separable brain models, but hopefully it is clear that such a narrow perspective cannot anticipate many of the critical issues that arise from considering the bigger picture.

There are a few final reflections on differences in overall approach that are worth discussing here. The approach I've taken is in many ways the inverse of what Marr was attempting. Marr's papers start with mathematical definitions, and proceed with mathematical precision in attempting to formalize the nature of information processing in different brain areas. In contrast, my approach is characterized by considerable exploratory, prototype-level investigations of various ideas, leveraging computational simulations as a fast and efficient way of determining whether an idea might work in practice. Eventually, once I narrow down and refine these rough drafts, it might be appropriate to submit some of the ideas to mathematical precision. On the other hand, analytical techniques place considerable limitations on the complexity of a

system that can be successfully treated. Thus, I have established a few general mathematical boundaries and approximations that guide the work (e.g., O'Reilly, 1996), but as yet have not found mathematically explicit approaches to provide a lot of direct guidance in developing my models. If Marr were around today, he would probably find my approach deeply unsatisfying, but perhaps he might nevertheless be excited by the kinds of problems we're confronting in taking these initial exploratory steps into the great unknown of the human mind.

References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychol. Rev.*, *111*(4), 1036–1060.

Baddeley, A. D. (1986). *Working memory*. Oxford University Press.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, *35*(8), 1798–1828.

Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *J. Neurosci.*, *2*(2), 32–48.

Blair, B. S., Intrator, N., Shouval, H., & Cooper, L. N. (1998). Receptive field formation in natural scene environments. comparison of single-cell learning rules. *Neural Comput.*, *10*, 1797–1813.

Bortone, D. S., Olsen, S. R., & Scanziani, M. (2014). Translaminar inhibitory cells recruited by layer 6 corticothalamic neurons suppress visual cortex. *Neuron*, *82*.

Brette, R., & Gerstner, W. (2005). Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *J. Neurophysiol.*, *94*(5), 3637–3642.

Buffalo, E. A., Fries, P., Landman, R., Buschman, T. J., & Desimone, R. (2011). Laminar differences in gamma and alpha coherence in the ventral stream. *Proc. Natl. Acad. Sci. U. S. A.*, *108*(27), 11262–11267.

Ciresan, D., Meier, U., & Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. *IEEE Conf Comput. Vis. Pattern Recognit. CVPR 2012*, 3642–3649.

Ciresan, D. C., Meier, U., Gambardella, L. M., & Schmidhuber, J. (2010). Deep, big, simple neural nets for handwritten digit recognition. *Neural Comput.*, *22*(12), 3207–3220.

Crick, F. (1989). The recent excitement about neural networks. *Nature*, *337*, 129–132.

Dayan, P., Hinton, G. E., Neal, R. N., & Zemel, R. S. (1995). The Helmholtz machine. *Neural Comput.*, *7*(5), 889–904.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38.

Elman, J. L. (1990). Finding structure in time. *Cogn. Sci.*, *14*(2), 179–211.

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

Földiák, P. (1990). Forming sparse representations by local anti-hebbian learning. *Biol Cybern*, *64*(2), 165–170.

Franceschetti, S., Guatteo, E., Panzica, F., Sancini, G., Wanke, E., & Avanzini, G. (1995). Ionic mechanisms underlying burst firing in pyramidal neurons: Intracellular study in rat sensorimotor cortex. *Brain Res.*, *696*(1–2), 127–139.

Frank, M. J. (2005). Dynamic dopamine modulation in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and non-medicated parkinsonism. *J. Cogn. Neurosci.*, *17*, 51–72.

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between the frontal cortex and basal ganglia in working memory: A computational model. *Cogn. Affect. Behav. Neurosci.*, *1*, 137–160.

Friedman, N., Herd, S. A., Hazy, T. E., Chatham, C. H., Kriete, T. E., Brant, A. M., & O'Reilly, R. C. (submitted). Neural network models of individual differences in executive functions.

Friston, K. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. B*, *360*(1456), 815–836.

Fuster, J. M., & Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science*, *173*, 652–654.

Geman, S., Bienenstock, E. L., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Comput.*, *4*, 1–58.

Goldman-Rakic, P. S. (1987). Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. *Handb. Physiol. — Nerv. Syst.*, *5*, 373–417.

Grossberg, S. (1999). How does the cerebral cortex work? learning, attention, and grouping by the laminar circuits of visual cortex. *Spat. Vis.*, *12*.

Grossberg, S. (2013). Adaptive resonance theory: How a brain learns to consciously attend, learn, and recognize a changing world. *Neural Networks*, *37*, 1–47.

Herd, S. A., Krueger, K. A., Kriete, T. E., Huang, T.-R., & O'Reilly, R. C. (2013). Strategic cognitive sequencing: A computational cognitive neuroscience approach. *Comput. Intell. Neurosci.*, *2013*, 149329.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

Jilk, D., Lebiere, C., O'Reilly, R., & Anderson, J. (2008). SAL: an explicitly pluralistic cognitive architecture. *J. Exp. Theor. Artif. Intell.*, *20*(3), 197–218.

Kachergis, G., Wyatte, D., O'Reilly, R. C., Kleijn, R. d., & Hommel, B. (2014). A continuous-time neural model for sequential action. *Phil. Trans. R. Soc. B*, *369*(1655), 20130623.

Kanerva, P. (1988). *Sparse distributed memory*. Boston: Bradford MIT.

Ketz, N., Morkonda, S. G., & O'Reilly, R. C. (2013). Theta coordinated error-driven learning in the hippocampus. *PLoS Comput. Biol.*, *9*, e1003067.

Kohonen, T. (1977). *Associative memory: A system theoretical approach*. Berlin: Springer-Verlag.

Kohonen, T. (1982). Clustering, taxonomy, and topological maps of patterns. *Proceedings of the 6th International Conference on Pattern Recognition* (pp. 114–128). Silver Spring, MD: IEEE Computer Society Press.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 25* (pp. 1097–1105). Curran Associates, Inc.

Lee, T. S., & Mumford, D. (2003). Hierarchical bayesian inference in the visual cortex. *J. Opt. Soc. Am.*, *20*(7), 1434–1448.

Lorincz, M. L., Kekesi, K. A., Juhasz, G., Crunelli, V., & Hughes, S. W. (2009). Temporal framing of thalamic relay-mode firing by phasic inhibition during the alpha rhythm. *Neuron*, *63*(5), 683–696.

Luczak, A., Bartho, P., & Harris, K. D. (2013). Gating of sensory input by spontaneous cortical activity. *J. Neurosci.*, *33*(4), 1684–1695.

Markov, N. T., Vezoli, J., Chameau, P., Falchier, A., Quilodran, R., Huissoud, C., Lamy, C., Misery, P., Giroud, P., Ullman, S., Barone, P., Dehay, C., Knoblauch, K., & Kennedy, H. (2014). Anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex: Cortical counterstreams. *J. Comp. Neurol.*, *522*(1), 225–259.

Marr, D. (1969). A theory of cerebellar cortex. *J. Physiol. Lond.*, *202*, 437–470.

Marr, D. (1970). A theory for cerebral neocortex. *Proc. R. Soc. Lond. B Biol. Sci.*, *176*(1043), 161–234.

Marr, D. (1971). Simple memory: A theory for archicortex. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, *262*(841), 23–81.

Marr, D. (1982). *Vision*. New York: Freeman.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.*, *102*(3), 419–457.

McClelland, J. L., Rumelhart, D. E., & the PDP Research Group (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*, Vol. 2: Psychological and Biological Models. MIT Press.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.*, *24*, 167–202.

Miyake, A., & Shah, P. (Eds.). (1999). *Models of working memory: Mechanisms of active maintenance and executive control.* New York: Cambridge University Press.

Montijn, J. S., Klink, P. C., & Van Wezel, R. J. A. (2012). Divisive normalization and neuronal oscillations in a single hierarchical framework of selective visual attention. *Front. Neural Circuits*, *6*, 22.

Olsen, S., Bortone, D., Adesnik, H., & Scanziani, M. (2012). Gain control by layer six in cortical circuits of vision. *Nature*, *483*(7387), 47–52.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*, 607.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Res.*, *37*(23), 3311–3325.

O'Reilly, R. (2006). Biologically based computational models of high-level cognition. *Science*, *314*(5796), 91–94.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Comput.*, *8*(5), 895–938.

O'Reilly, R. C. (1998). Six principles for biologically-based computational models of cortical cognition. *Trends Cogn. Sci.*, *2*(11), 455–462.

O'Reilly, R. C. (2001). Generalization in interactive networks: The benefits of inhibitory competition and hebbian learning. *Neural Comput.*, *13*(6), 1199–1242.

O'Reilly, R. C., Braver, T. S., & Cohen, J. D. (1999). A biologically based computational model of working memory. In A. Miyake, & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control.* (pp. 375–411). New York: Cambridge University Press.

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Comput.*, *18*(2), 283–328.

O'Reilly, R. C., Hazy, T. E., Mollick, J., Mackie, P., & Herd, S. (2014a). Goal-driven cognition in the brain: A computational framework. *ArXiv14047591 Q-Bio*.

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a tradeoff. *Hippocampus*, *4*(6), 661–682.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.

O'Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E., & Contributors (2012). *Computational cognitive neuroscience*. Wiki Book, 1st Edition, URL: http://ccnbook.colorado.edu.

O'Reilly, R. C., Petrov, A. A., Cohen, J. D., Lebiere, C. J., Herd, S. A., & Kriete, T. (2014b). How limited systematicity emerges: A computational cognitive neuroscience approach.

O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychol. Rev.*, *108*(2), 311–345.

O'Reilly, R. C., Wyatte, D., Herd, S., Mingus, B., & Jilk, D. J. (2013). Recurrent processing during object recognition. *Front. Psychol.*, *4*(124).

O'Reilly, R. C., Wyatte, D., & Rohrlich, J. (2014c). Learning through time in the thalamocortical loops. *ArXiv14073432 Q-Bio*.

OReilly, R. C. (2010). The What and How of prefrontal cortical organization. *Trends in Neurosciences*, *33*(8), 355–361.

OReilly, R. C., Hazy, T. E., & Herd, S. A. (2015). The Leabra cognitive architecture: How to play 20 principles with nature and win! In S. Chipman (Ed.), *Oxford handbook of cognitive science*. Oxford University Press.

Pollack, J. B. (1990). Recursive distributed representations. *Artif. Intell.*, *46*(1), 77–105.

Pylyshyn, Z. W. (2000). Situating vision in the world. *Trends Cogn. Sci.*, *4*, 197–207.

Raizada, R. D. S., & Grossberg, S. (2003). Towards a theory of the laminar architecture of cerebral cortex: computational clues from the visual system. *Cereb. Cortex*, *13*(1).

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron*, *61*(2), 168–185.

Rudy, J. W., & O'Reilly, R. C. (1999). Contextual fear conditioning, conjunctive representations, pattern completion, and the hippocampus. *Behav. Neurosci.*, *113*, 867–880.

Rudy, J. W., & O'Reilly, R. C. (2001). Conjunctive representations the hippocampus and contextual fear conditioning. *Cogn. Affect. Behav. Neurosci.*, *1*(1), 66–82.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*, Vol. 1: Foundations. Cambridge, MA: MIT Press.

Shouval, H. Z., Wang, S. S.-H., & Wittenberg, G. M. (2010). Spike timing dependent plasticity: A consequence of more fundamental learning rules. *Front. Comput. Neurosci.*, *4*(19).

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, *15*(1), 1929–1958.

Sun, Y., O'Reilly, R. C., Bhattacharyya, R., Smith, J. W., Liu, X., & Wang, H. (2015). Latent structure in random sequences drives neural learning toward a rational bias. *PNAS*, *112*(12), 3788–3792.

Taatgen, N. A. (2013). The nature and transfer of cognitive skills. *Psychol. Rev.*, *120*.

Thomson, A. M., & Lamy, C. (2007). Functional maps of neocortical local circuitry. *Front. Neurosci.*, *1*(1), 19–42.

Urakubo, H., Honda, M., Froemke, R. C., & Kuroda, S. (2008). Requirement of an allosteric kinetics of NMDA receptors for spike timing-dependent plasticity. *J. Neurosci.*, *28*(13), 3310–3323.

Ziegler, M. D., Chelian, S. E., Benvenuto, J., Krichmar, J. L., O'Reilly, R., & Bhattacharyya, R. (2014). A model of proactive and reactive cognitive control with anterior cingulate cortex and the neuromodulatory system. *Biologically Inspired Cognitive Architectures*, *10*, 61–67.

Figure Captions

Figure 1. The large-scale cognitive architecture of cognition involves three principle subsystems, each with different functional specializations as noted. The posterior neocortex and hippocampus function much as Marr envisioned, while the prefrontal cortex / basal ganglia system adds the ability to sustain robust active maintenance and select cognitive and motor actions according to history of reward.

Figure 2. The core microstructural properties of the Leabra architecture.

Figure 3. Overall structure of the hippocampal memory system. The perforant path projections from EC to DG and CA3 are largely as envisioned by Marr, but the role of the CA1 as a invertible decoder of hippocampal memories provides a better solution than direct association of memories back in the neocortex.

Figure 4. Overall architecture of the PBWM model of PFC-BG working memory function.

Figure 5. The ACT-R cognitive architecture and its mapping onto brain areas – note the similarity to the biologically-based cognitive architecture.

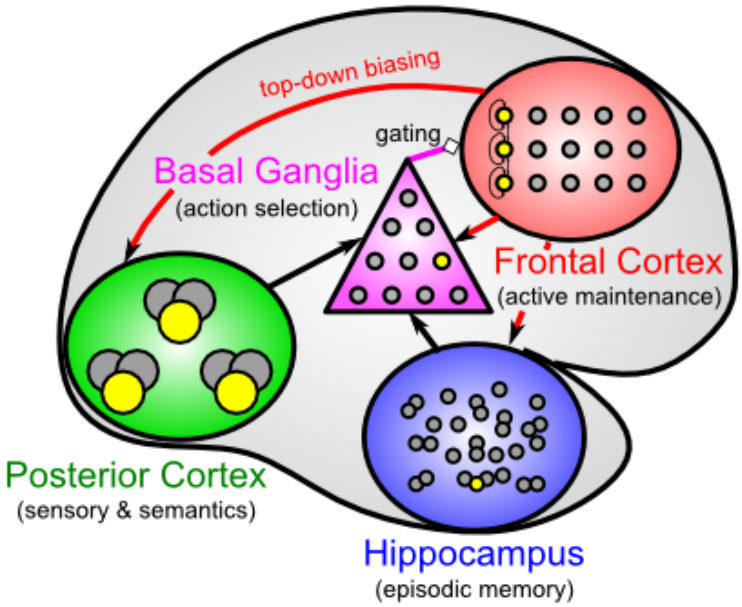
Figure 6. A large-scale integrated cognitive model developed for the ICArUS project to develop a model of the sensemaking process and the origin of cognitive biases.

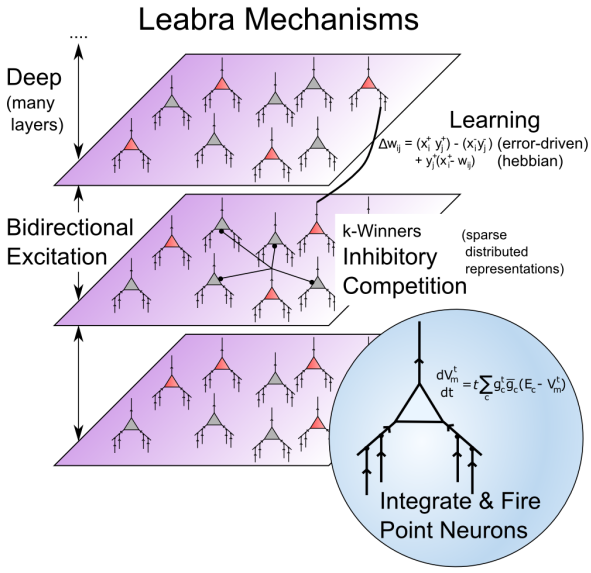
Figure 7. Detailed neocortical and thalamocortical connectivity, e.g., between V1 and V2, and the overall functionality they support. TRN = thalamic reticular nucleus inhibitory neuron, TRC = thalamic relay cell, ib = intrinsic bursting, rs = regular spiking, cc = cortico-cortical, ct = cortico-thalamic.

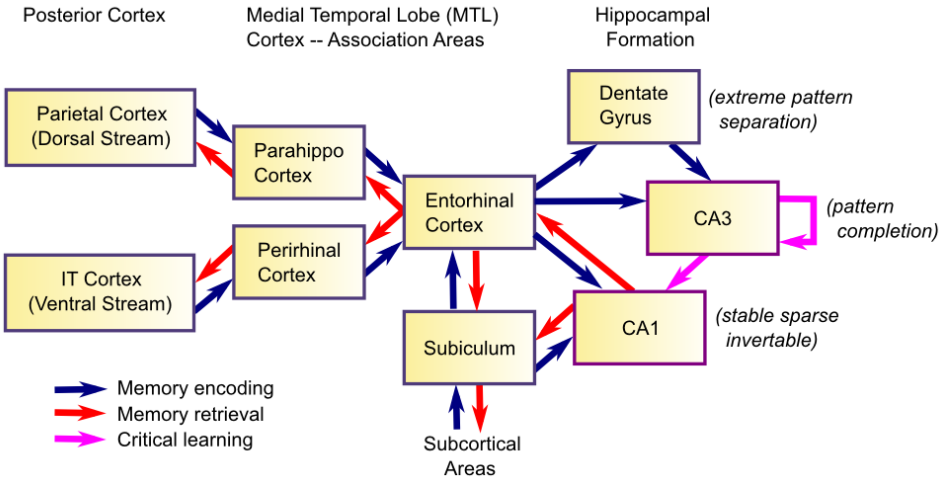
Figure 8. Structure of auto-encoder learning based on the thalamic relay cells (TRC's) acting as a target "output" layer when they are driven by strong feedforward driver inputs originating in the deep5b IB neurons from the

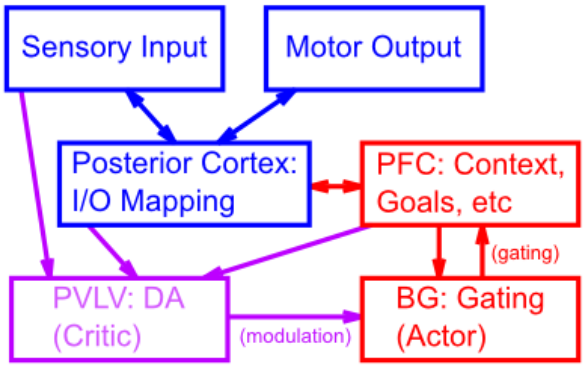
previous layer(s) in the network. The network learns to encode the input information to accurately drive the TRC's in the minus phase. This circuit and associated error-driven learning dynamic, operating locally in every neocortical area, provides a powerful learning signal, in addition to the bidirectional propagation of differences through longer superficial-layer interconnections.

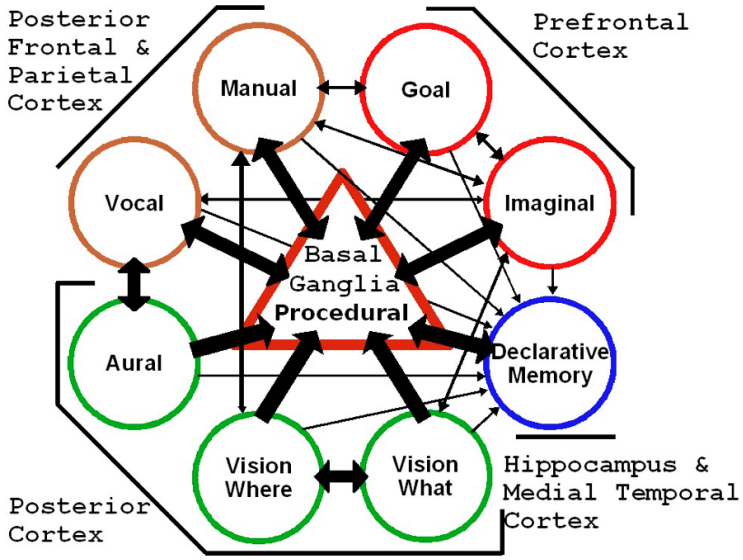
Figure 9. GRID representation for sentence / visual experience "Tom ate steak at the table" – Tom gets put in index 1 of the "thing" slots as the subject / agent, steak is object in slot 2, and table is slot 3 as a location modifier. The parietal representations contain the visual location (real or imagined) of each entity. The ventral visual representations encode the visual category of each entity – an abstracted version of what each looks like, which is also a point of entry into the vast semantic network (not shown, but theoretically populated with large numbers of different GRID slots). The name in higher-level auditory cortex represents the phonological representation of each item. And so on.. The overall idea, captured in the usual meaning of grid, is that these 3-4 slots are present in each domain-specific area, and coordinated so that the same thing is encoded in each, with each area encoding a different aspect of that thing. Other kinds of representations such as actions (e.g., "ate") and relationships (e.g., "at") can then operate over multiple "thing" slots.

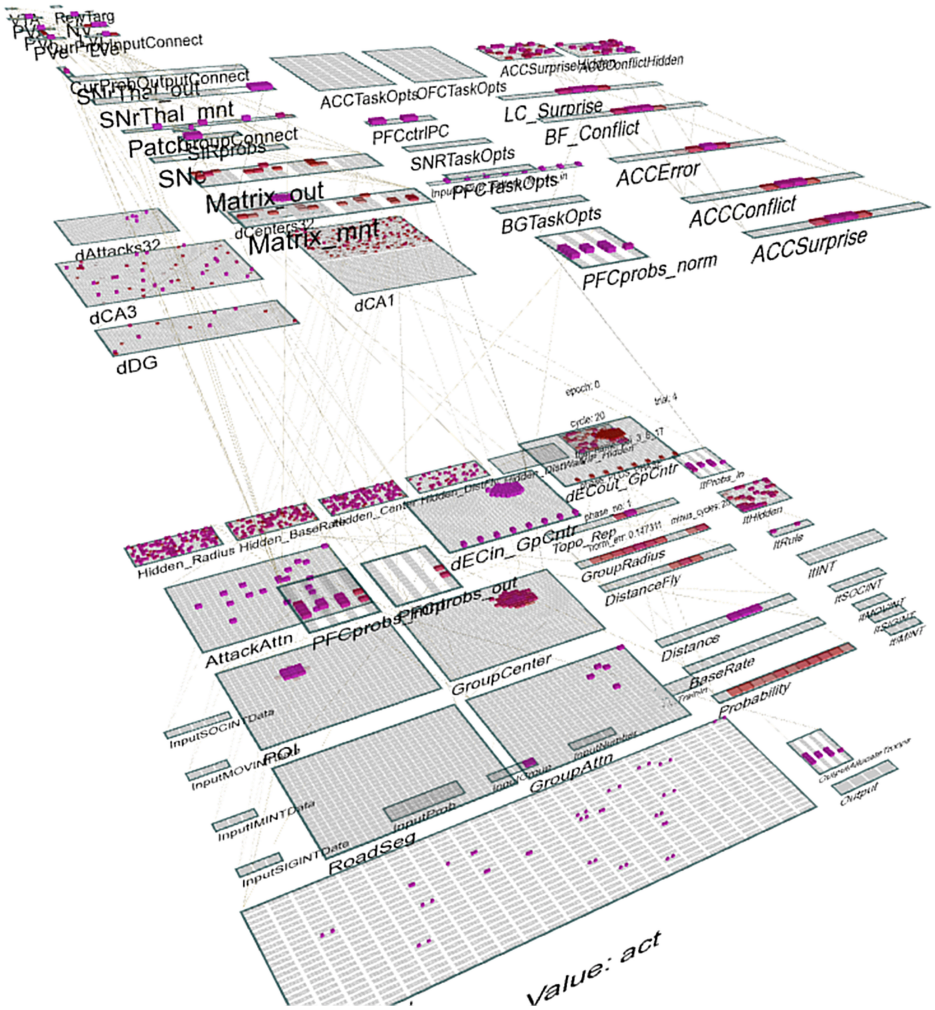

top-down biasing
gating
Basal Ganglia
(action selection)
Frontal Cortex
(active maintenance)
Posterior Cortex
(sensory & semantics)
Hippocampus
(episodic memory)

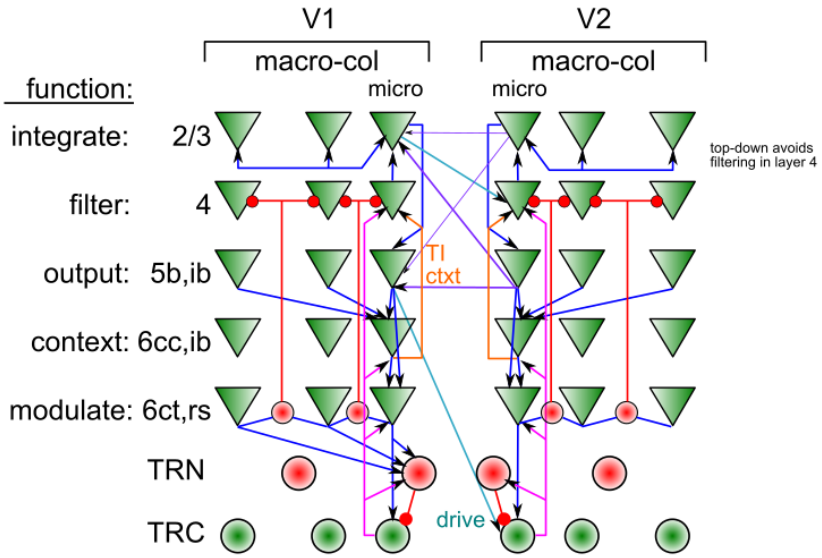

Leabra Mechanisms
Deep
(many
layers)
Learning
Δwij = (xi+ yj+) - (xi yj) (error-driven)
+ yj+(xi+ - wij) (hebbian)
Bidirectional
Excitation
k-Winners
Inhibitory
Competition
(sparse
distributed
representations)
dVmt/dt = τ Σc gct gc (Ec - Vmt)
Integrate & Fire
Point Neurons

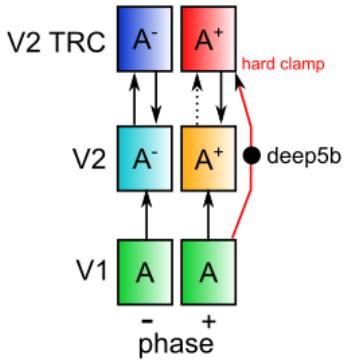

Posterior Cortex
Medial Temporal Lobe (MTL)
Cortex -- Association Areas
Hippocampal
Formation
Parietal Cortex
(Dorsal Stream)
IT Cortex
(Ventral Stream)
Parahippo
Cortex
Perirhinal
Cortex
Entorhinal
Cortex
Subiculum
Dentate
Gyrus
(extreme pattern
separation)
CA3
(pattern
completion)
CA1
(stable sparse
invertable)
Memory encoding
Memory retrieval
Critical learning
Subcortical
Areas


Sensory Input
Motor Output
Posterior Cortex:
I/O Mapping
PFC: Context,
Goals, etc
(gating)
PVLV: DA
(Critic)
(modulation)
BG: Gating
(Actor)


Posterior
Frontal &
Parietal
Cortex
Prefrontal
Cortex
Manual
Goal
Vocal
Imaginal
Basal
Ganglia
Procedural
Aural
Declarative
Memory
Vision
Where
Vision
What
Hippocampus &
Medial Temporal
Cortex
Posterior
Cortex


ACCTaskOpts
OFCTaskOpts
ACCSurpriseHidden
ACCConflictHidden
LC_Surprise
BF_Conflict
PFCtrlPC
SNRTaskOpts
PFCTaskOpts
ACCError
ACCConflict
ACCSurprise
BGTaskOpts
PFCprobs_norm
SNrThal_out
SNrThal_mnt
PatchGroupConnect
SNrprobs
SNc
Matrix_out
Matrix_mnt
dCenters_in
dAttacks32
dCA3
dCA1
dDG
dECout_GpCntr
dECin_GpCntr
Topo_Rep
GroupRadius
DistanceFly
Hidden_Radius
Hidden_BaseRate
AttackAttn
PFCprobs_out
GroupCenter
Distance
BaseRate
Probability
InputSOCINData
InputMOVINData
InputMINTData
InputSIGINTData
RoadSeg
GroupAttn
Value: act


V1
V2
macro-col
macro-col
micro
micro
function:
integrate:
2/3
filter:
4
output:
5b,ib
context:
6cc,ib
modulate:
6ct,rs
TRN
TRC
TI
ctxt
drive
top-down avoids
filtering in layer 4


V2 TRC
A-
A+
hard clamp
V2
A-
A+
deep5b
V1
A
A
-
+
phase

GRID rep for "Tom ate steak at the table"