The Leabra Cognitive Architecture: How to Play 20 Principles with Nature and Win!

Randall C. O'Reilly, Thomas E. Hazy, and Seth A. Herd
Department of Psychology and Neuroscience
University of Colorado Boulder
345 UCB
Boulder, CO 80309
randy.oreilly@colorado.edu

August 6, 2014

Abstract:

This chapter provides a synthetic review of a long-term effort to produce an internally consistent theory of the neural basis of human cognition, the Leabra cognitive architecture, which explains a great deal of brain and behavioral data. In a highly influential commentary, Allen Newell (1973) first issued a call for a more comprehensive, principled approach to studying cognition. "You can't play 20 questions with nature, and win," he said, alluding to the old parlor guessing game involving 20 yes or no questions. His point was that cognition, and the brain that gives rise to it, is just too complex and multidimensional a system to ever hope that a series of narrowly framed experiments and/or models would ever be able to characterize it. Instead, a single cognitive architecture should be used to simulate a wide range of data at many levels in a cumulative manner. However, these cognitive architectures tend to be complex and difficult to fully comprehend. In an attempt to most clearly and simply present the Leabra biologically-based cognitive architecture, we articulate 20 principles that motivate its design, at multiple levels of analysis.

Introduction

The Leabra cognitive architecture described in this chapter is one of several cognitive architectures that have been developed over the past several decades. As we elaborate below, a cognitive architecture can be defined as a comprehensive, mechanistically detailed theory of how cognition operates across a wide range of domains and tasks, implemented in a working computer simulation system. Cognitive architectures are fundamentally concerned with characterizing how cognition works at a mechanistic level, as opposed to descriptive or abstract theorizing. More than perhaps any other proposed cognitive architecture, Leabra is based directly on the underlying biology of the brain, with a set of biologically realistic neural processing mechanisms at its core. In many ways, it represents a natural evolution of the neural network / parallel distributed processing / connectionist models that were popular in the late 1980's and 1990's — an evolution that grounds the mechanisms in the biology (e.g., by using a biologically-plausible version of error-driven learning; O'Reilly, 1996a; O'Reilly, Munakata, Frank, Hazy, & Contributors, 2012), and also

Draft Manuscript: Do not cite or quote without permission.
We thank Alex Petrov, Yuko Munakata, and members of the CCN Lab at CU Boulder for helpful comments and discussion. Supported by ONR N00014-07-1-0651, ARL RCTA, and NIH MH079485. Supported 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.

makes strong commitments to specific ideas about the large scale functional organization of the brain. This functional organization has converged to a remarkable extent with the functional architecture of a more purely cognitively derived architecture, the ACT-R framework (Anderson, Bothell, Byrne, Douglass, Lebiere, & Qin, 2004), as we discuss in Jilk, Lebiere, O'Reilly, and Anderson (2008).

We proceed as follows. First, we discuss the motivations for creating cognitive architectures, their advantages in creating accurate theories of cognition, and the difficulties that prevent many researchers from working within them. We then describe a set of principles that provide a high-level view of the current state of the Leabra cognitive architecture project, starting from the principles of neural function in general, and moving to specific theories of neural function in specialized brain areas that support sensory processing and semantic knowledge (posterior cortex), episodic memory (the hippocampus), working memory and executive function (the prefrontal cortex and basal ganglia), and reward processing and motivational systems (from the medial frontal cortex down to the brainstem).

Motivating Cognitive Architectures

Why should one be interested in the Leabra cognitive architecture, and in cognitive architectures more generally? What can such a thing offer that other more focused cognitive models or theories cannot — e.g., why is it worth the effort to understand a complicated theoretical framework when perhaps one only cares about more specific issues? Is it premature or presumptuous to offer some kind of comprehensive cognitive theory, when there is so much we do not yet understand about how the mind/brain functions? These are some of the important questions that we attempt to address here.

Cognitive architectures generally lie at the complex end of a spectrum of computational modeling frameworks. Why would anyone favor a more complex model over a simpler one, when Occam's famous razor clearly directs us to favor simpler models over more complex ones (not to mention the practical issues in thinking about, implementing models of and sharing credit for a more complex theory)? Clearly, if there really was a simple model that can account for all of the complexity of human cognition, that would be ideal. However, every indication is that the brain, evolved as it has over millions of years across the great chain of being preceding human beings, is not likely to be described with a single simple homogeneous algorithm. Instead, as we elaborate below, cognition appears to require the interaction of a number of specialized processing systems. Thus, the central question is: what are the potential problems of using overly-simple models that fail to capture the full set of relevant cognitive mechanisms?

Allen Newell made the case that there are significant risks to using narrow, simpler models in his famous "You can't play 20 questions with nature and win" commentary (Newell, 1973). He suggested that a comprehensive, principled, and constrained approach to cognitive modeling will be more likely to bear fruit than making a bunch of one-off models of specific phenomena using simpler modeling tools, which he likens to answering individual binary questions like in the classic "20 questions" game (e.g., is visual search parallel or serial?). In that paper, and later in his influential book *Unified Theories of Cognition* (1990), Newell advocated developing a strongly-constrained and comprehensive framework, i.e., what has come to be known as a *cognitive architecture*, and applying it to many different cognitive phenomena, each of which tests the theory/architecture in different ways. If a cumulative theory can successfully do that, then there is good reason to believe in its validity as a model of human cognition. Otherwise, it is simply too easy to fit any given small subset of phenomena with any theory of limited scope.

Newell's argument is really just an instance of the basic idea that scientific theories that account for more data are better than those that account for less. But in the context of cognition, the point is particularly pressing, because the brain/mind is such a powerful and complex thing — any given small window onto it will fail to reveal the global principles that operate across all of the windows. This is particularly important for integrating across the biological and cognitive levels of analysis, which each provide very different

kinds of constraints. This is similar to the parable of the blind men describing different parts of an elephant. You need the big picture to put all these parts into proper perspective. A good cognitive architecture can provide this kind of big picture framework.

In summary, Occam's razor cuts between two opposing tensions: simplicity and accounting for increasingly larger quantities of relevant data — often people neglect the importance of this latter constraint. Realistically, covering a broad scope of complex phenomena will probably require a more complex theory than coverage of a narrow scope of phenomena. And Newell argues that this breadth constraint is more important than the simplicity one, in the context of understanding human cognition, so we should be willing to embrace more complex cognitive architectures, if they allow us to understand a great breadth of cognition.

One important way to mitigate against the perceived complexity of a given theory is to provide the clearest and most strongly principled account of it, so as to eliminate as much as possible any sense of arbitrariness in the framework. Hence, this paper is an attempt to articulate 20 clear principles that strongly constrain the nature of the Leabra architecture. The goal is to ultimately arrive at a computational model of the brain/mind that is as simple and clear as possible, but still accounts for a wide range of cognitive and neuroscience phenomena.

Introduction to the Leabra Architecture

The Leabra framework started with a neural network algorithm intended to capture the core computational properties of the neurobiology of the neocortex, which supports many different cognitive functions (O'Reilly, 1996b, 1998). There was a progressive elaboration of these neural mechanisms to account for the specialized properties of different areas of the brain, including the hippocampus (O'Reilly & McClelland, 1994; McClelland, McNaughton, & O'Reilly, 1995; O'Reilly & Rudy, 2001; Norman & O'Reilly, 2003; O'Reilly, Bhattacharyya, Howard, & Ketz, 2011), prefrontal cortex and basal ganglia (O'Reilly, Braver, & Cohen, 1999; Frank, Loughry, & O'Reilly, 2001; O'Reilly & Frank, 2006; O'Reilly, 2006; Hazy, Frank, & O'Reilly, 2006, 2007), and subcortical reward-processing areas (O'Reilly, Frank, Hazy, & Watz, 2007; Hazy, Frank, & O'Reilly, 2010). The first attempt to articulate a broad cognitive-architecture level theory based on Leabra was in a textbook covering a wide range of cognitive phenomena (O'Reilly & Munakata, 2000). This text has been updated to include the most recent developments in a freely available online format at `http://ccnbook.colorado.edu` (O'Reilly et al., 2012), so this is an opportune time for summarizing the current state of the architecture. We refer the reader to this resource for the specific equations used in Leabra, along with many implemented models illustrating its behavior.

Insert Figure 1 about here.

Insert Figure 2 about here.

To give a brief sense of some of the most recent, cutting-edge Leabra models, Figure 1 shows the model from the ICArUS project which is attempting to develop *integrated cognitive-neuroscience architectures for understanding sensemaking* — this project represents a collaboration among several different labs, and the model can simulate human behavior on a series of complex sensemaking tasks, while providing insights into the biological basis of cognitive biases in these domains. This model represents the most complex integrated cognitive functionality that has been simulated in Leabra to date — it can coordinate multiple reasoning processes over a number of individual steps, performing a series of tasks that people take around an hour to complete in total. The development of this model has proceeded with a progressive

removal of initial "scaffolding" that was needed to keep everything moving on track over time. Overall, this model has given us many insights into how the architecture needs to develop to address this level of complex cognition in increasingly realistic ways.

Figure 2 shows a model of an embodied cognitive agent, called *emer* (after the *emergent* software in which Leabra is implemented), which performs basic visual saccades using coordinated head and eye movements via a simulated cerebellar system, and can then recognize the object in the focus of attention, with high levels of accuracy for 100 different object categories, even novel objects from these categories (over 90% generalization accuracy; O'Reilly, Wyatte, Herd, Mingus, & Jilk, submitted). Ongoing work is developing the ability to use a wide range of cues and gestalt principles to separate figure from ground, to enable robust object recognition even in cluttered visual scenes.

Insert Figure 3 about here.

Before we get started playing "20 principles with nature" to motivate the Leabra architecture, it is useful to characterize the nature of these principles. These principles span many different levels (Figure 3) and domains that describe the Leabra cognitive architecture, and, we argue, capture some important truths about how the brain and cognition operate (see O'Reilly, 1998 for an earlier attempt to articulate some of these principles). Although 20 principles may sound like a lot, because these principles are organized at different levels of analysis, there are fewer principles per each level. As with 20 questions, we start with very broad principles that shape the overall approach (the *metalevel*), and then develop a set of more specific principles of neural computation based on solid neuroscience data that strongly constrain our model of the *microlevel* (i.e., the microstructure of cognition; c.f., Rumelhart, McClelland, & the PDP Research Group, 1986b; McClelland, Rumelhart, & the PDP Research Group, 1986; McClelland & Rumelhart, 1988). Next, we advance principles of large-scale brain area specializations that constitute a *macrolevel* description of the cognitive architecture. Critically, many of these macrolevel principles are derived directly from properties of the microlevel, which is essential for establishing a truly integrated, unified theory of cognition, as opposed to just a laundry list of isolated ideas.

Our key criteria for elevating something to the level of a principle are: (a) it can be summarized briefly and makes a strong positive assertion; and (b) the truth-value of the assertion is directly consequential for a decision about how the architecture should be shaped. Thus, the reader can hopefully decide the extent to which they agree with the various principles, and thus have a better handle on evaluating the architecture overall. We also attempt to provide some contrasting examples to demonstrate that these principles are not universal platitudes.

The Metalevel Guiding Principles in the Development of Leabra

The name Leabra originated as an acronym standing for *Local, Error-driven and Associative, Biologically Realistic Algorithm* to reflect its core focus on the nature of learning (a locally-computable combination of error-driven and Hebbian associative mechanisms). It 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 (i.e., *computational cognitive neuroscience*). Thus, this approach is antithetical to "purist" approaches that attempt to optimize a single criterion or objective function. Here are the broad principles that shape the overall approach in developing the Leabra architecture:

Principle 1 (Balance): *There are important tradeoffs associated with almost every approach, objective, or computational solution, and often the best overall solution represents a compromise or other form of integration of multiple different approaches/objectives/solutions.*

Although this principle may seem obvious, many computational modeling approaches favor purity and simplicity over dealing with the complex tradeoffs apparent in the brain and cognition. Simple, single-principle models can be the best way to convey a specific idea, but often that idea must be tempered against various other constraints and considerations to understand in detail how people actually behave in a variety of contexts.

Principle 2 (Biology is important): *The brain is our one working "reference implementation" of a successful cognitive system, so trying to understand in detail how it works may be the one nearly-guaranteed path to a cognitive architecture that accurately models the human mind.*

As noted above, Leabra is one of the few cognitive architectures that is based so directly on the biology, and only recently have we implemented models that incorporate much of the full architecture (e.g., Figure 1)— most of the published models have explored the components separately. Of course, there are significant practical barriers to implementing detailed biological models at a large scale and it is only recently that computers have become powerful enough to even begin to make this feasible. This practical constraint converges with our next principle.

Principle 3 (Occam's Razor): *Scientifically, we seek the simplest model that is sufficient to account for the relevant aspects of neuroscience and cognition, because this will be the easiest to understand, and the least likely to go astray by overfitting the available data.*

Intuitively, replicating every single biological detail would get us no closer to understanding how the brain works — we already have the full complexity of the real brain, and any functionally irrelevant details just get in the way of understanding the underlying computational principles. Many computational neuroscience models focus on capturing as much biological detail as possible, and one project that has received quite a bit of notoriety explicitly assumes that in so doing the magic of cognition will just emerge from all those details (Markram, 2006). In contrast, the Leabra approach is predicated on the idea that trying to understand what is going on at the psychological and mechanistic levels *simultaneously* is key to meaningful progress. This necessarily entails the discovery and imposition of constraints at multiple levels and, combined with a considered effort to include only as much mechanistic detail as is absolutely necessary to explain function, is the most direct path toward understanding the principles by which the brain/mind works.

Principle 4 (Convergent multi-level modeling): *The optimal balance between biological, cognitive, and computational constraints is likely to be different depending on the nature and level of the questions being addressed.*

Given this principle, it makes sense to develop a family of models at different levels of abstraction that are nonetheless mutually compatible and serve to constrain one another, ultimately aiming to arrive at a convergent, multi-level description of the system as a whole. There are many different optional switches in the Leabra simulation software that can dial up or down the level of abstraction of any given model, and there are *bridging simulations* that specifically test the convergence and mutual compatibility of abstractions at different levels of abstraction. At the highest level of abstraction, the ACT-R framework shares many of the same architectural features as Leabra, and we are currently working to develop a *S*ynthesis of *A*CT-R and *L*eabra (SAL; Jilk et al., 2008) architectures that explicitly integrates features from both architectures to yield an even more convergent higher-level abstract architecture. In this overview, we focus on the middle level of abstraction provided by the "default" version of Leabra, while noting the options for increasing or decreasing biological detail.

Principle 5 (Learning is critical): *Within the context of ontogenetic developmental processes, much of cognitive function is acquired via experience-driven learning mechanisms, which sculpt the raw neural material of the cortex into highly functional neural systems*

The human brain learns to read and write, and a host of other novel skills that couldn't possibly be directly coded by our genetics. To capture this kind of pervasive learning, the system must be capable of developing entirely new representations and cognitive abilities, not just tune a set of parameters within an otherwise preconfigured system. This principle is central to the Leabra approach — everything that a typical Leabra model can do involves a substantial learning component, using mechanisms that are intended to capture the essential properties of cortical learning, and supported by a critical bridging simulation (described below) that grounds the Leabra learning in known biological mechanisms. The ability to develop complex cognitive functions through learning has always been one of the most important features of neural network models, and to this day no other framework has been developed that is as capable of such general-purpose, powerful learning. Indeed, there has recently been somewhat of a resurgence of interest in these neural network learning mechanisms within the statistical computing and machine learning communities (Hinton & Salakhutdinov, 2006; Ciresan, Meier, Gambardella, & Schmidhuber, 2010; Koller & Friedman, 2009).

One possible explanation for the unique suitability of neural networks for learning is that the ability to learn entirely new cognitive functions requires an equipotential, homogenous substrate to start from so that it can be shaped over time through learning — a neural network provides just such a substrate. In contrast, it is difficult to reconcile this equipotentiality demand with the need to have intricate, highly differentiated structures in the system, as is typically required to achieve sensible symbolic processing abilities for example. The Leabra framework does allow for various forms of built-in structure and parameter differences across areas, but these serve to constrain and shape the properties and outcome of the learning mechanism, not to provide initial cognitive functionality. Another important factor is that learned functionality must go through many intermediate stages during the learning process, so whatever is learned will typically be sufficiently robust to support partial functionality when partially developed. But many cognitive models with more elaborated, interdependent processing mechanisms would not function at all in a partially-learned state (e.g., imagine the functionality of a partially-implemented CPU chip). Thus, we believe that learning provides considerable constraints on the nature of the system, and a deep understanding for why the brain is made of networks of neurons.

The central role of learning in Leabra is a point of contrast with many other cognitive architectures, most of which focus more on modeling the performance aspects of cognition, using various forms of task-relevant knowledge representations provided by the modeler. The ACT-R architecture has a strong focus on learning, and includes a number of important learning mechanisms in its declarative memory and procedural production system, but even here the modeler has to build in a considerable amount of starting knowledge. There must be a sufficient basis set of initial productions to drive the sequence of cognitive operations performed, as well as the explicit specification of things like the allowed forms of memory representation for any given task (i.e., the memory chunk types). Learning in ACT-R then operates within these initial parameters to optimize the flow of productions, and acquire new declarative memories. Interestingly, a connectionist implementation of ACT-R was developed (Lebiere & Anderson, 1993), which influenced further developments of the architecture. However, this connectionist implementation required extensive pre-structuring of the same form as required in regular ACT-R, and did not employ generic learning from a homogenous substrate in the way that Leabra does. Another interesting point of contrast is the *Neural Engineering Framework* of Eliasmith and Anderson (2003), which can create impressive neural systems through a powerful parameter-setting mechanism (see `http://nengo.ca`). But this mechanism is a purely engineering process that does not represent an experience-driven learning mechanism like that operating in the human brain.

Next, we describe more detailed principles and their implications for the Leabra model, beginning with basic neural network-level principles and algorithms that define the *microstructure* of cognition (c.f., Rumelhart et al., 1986b; McClelland et al., 1986; McClelland & Rumelhart, 1988), followed by a discussion of the *macrostructure* of cognition in terms of architectural principles governing our understanding of the specializations of different brain areas for different cognitive functionality (see Figure 3).

The Microstructure of Cognition: Principles of Neural Computation

Insert Figure 4 about here.

We begin this section with a set of four principles about how information processing is thought to arise in the brain, and which specific types of neurons are most important for understanding cognition. With the possible exception of Principle 9, these are largely consistent with most neural network / parallel distributed processing / connectionist models (McClelland, 1993; McClelland et al., 1986; McClelland & Rumelhart, 1988; Rumelhart et al., 1986b; O'Reilly, 1998), but not directly implemented in more abstract cognitive architectures such as ACT-R.

Principle 6 (Networks of neurons are the fundamental information processors in the brain): *Neurons integrate many different synaptic input signals from other neurons into an overall output signal that is then communicated to other neurons, and this provides the core information processing computation of cognition. Simplistically, each neuron can be considered as a detector, looking for particular patterns of synaptic input, and alerting others when such patterns have been found.*

Principle 7 (Synaptic weights encode knowledge, and adapt to support learning): *Synaptic inputs vary in strength as a function of sender and receiver neuron activity, and this variation in strength can encode knowledge, by shaping the pattern that each neuron detects.*

There is now copious empirical evidence supporting this principle and it can probably be considered uncontroversial in the neuroscience community at this point.

Principle 8 (Pyramidal neurons in neocortex are the primary information processors of relevance for higher cognition): *The neocortex is the primary locus of cognitive functions such as object recognition, spatial processing, language, motor control, and executive function, and all of the long-range connectivity between cortical areas is from excitatory pyramidal neurons.*

Pyramidal neurons constitute the primary information processing neurons in cortex. They are excitatory, and predominantly bidirectionally connected with each other. Many other subcortical brain areas make important contributions to cognition, but the neocortex performs the bulk of the information processing, particularly for the higher functions that are most studied in current cognitive neuroscience.

Principle 9 (Inhibitory interneurons regulate activity levels on neocortex, and drive competition): *This inhibitory dynamic gives rise to competition among neurons, producing many beneficial effects on learning and performance.*

The other major neuron type in neocortex are locally-projecting inhibitory interneurons, of which there are a great variety, and they generally serve to regulate overall activity levels through GABA inhibition onto pyramidal neurons. Inhibitory interneurons produce competition among pyramidal neurons, allowing the many benefits of biased competition for attention and executive function (Desimone & Duncan, 1995; Herd, Banich, & O'Reilly, 2006). When the inhibitory system goes awry, bidirectional excitation between pyramidal neurons results in runaway epileptiform activity. And, there is evidence that individual differences in GABAergic tone in prefrontal cortex can affect cognitive functioning. For example, Snyder,

Hutchison, Nyhus, Curran, Banich, and Munakata (2010) showed that people with lower levels of inhibition in ventral-lateral PFC had a more difficult time selecting among alternative words, and lower inhibitory tone was also associated with difficulty in decision making in anxious individuals.

The foregoing set of principles translate directly into a set of specific questions that must be addressed in the Leabra framework, questions that may have multiple answers depending on level of abstraction:

- How best to simulate the dynamic properties of the neocortical pyramidal neuron (i.e., the *neural activation function*), to achieve a computationally-tractable model that captures the most important properties of neural function without unnecessary biological baggage?
- How best to simulate the change in synaptic strength as a function of neural activity (i.e., the neural *learning rule*), in a way that captures what is known biologically about these synaptic plasticity mechanisms, while also enabling a network to learn to solve the kinds of difficult cognitive problems known to be solved in different cortical brain areas?
- How best to simulate the effects of inhibitory interneurons on network dynamics (i.e., the *inhibition function*), in a way that again balances biological fidelity with computational efficacy?

A variety of different answers to each of these questions have been proposed in the literature. For example, the standard feedforward backpropagation network uses a simple sigmoidal rate-code equation for the neural activation function, simulating discrete neural spiking in terms of a real-valued number representing something like the rate of firing over time, and it uses a biologically implausible learning rule that requires error signals to somehow propagate backward down dendrites, across the synapse, and down the axon of the sending neuron. There is no inhibition function at all, and the critical feature of bidirectional excitatory connectivity among pyramidal neurons is similarly missing. Thus, we can reasonably argue that a feedforward backprop network abstracts rather far away from the known biology. On the other end of the spectrum, there are many computational neuroscience models with highly detailed multi-compartment pyramidal neurons, employing various forms of biologically grounded Hebbian-style learning rules, and detailed inhibitory interneurons with appropriate connectivity to balance out bidirectional excitatory feedback loops among the pyramidal neurons (e.g., Markram, 2006; Traub, Miles, & Wong, 1989; Izhikevich & Edelman, 2008). But these latter models do not actually solve complex cognitive tasks (e.g., object recognition in the ventral visual stream) and they take a long time to simulate the dynamics of even a single neuron, limiting the ability to simulate the extended timecourse of learning in a large-scale model.

Consistent with the emphasis on balance, the Leabra architecture seeks a middle ground between these two extremes — computationally and cognitively powerful, but more closely tied to the biology and capable of exhibiting more complex excitatory and inhibitory dynamics that very likely play a significant role in many cognitive phenomena. Within this target space, there are still likely to be a range of different implementational choices that will result in generally similar cognitive functionality. Indeed, we know that within the Leabra framework different choices have been developed over time, and are available as options in the simulator. Nevertheless, our current best answers are described in the following sections (see Figure 4 for a summary).

Neural activation function

We borrow the *adaptive exponential* (AdEx) model of the pyramidal neuron (Brette & Gerstner, 2005), which has won competitions for best predicting cortical neural firing patterns, and is on the same computational order as other abstract neural equations. Conveniently, it represents just a few additions to the basic conductance-based point neuron equations used in the original Leabra model — these add spike frequency adaptation and an exponential spike initiation dynamic. The AdEx model produces discrete

spiking outputs, but often this level of detail incurs too much computational overhead, so we also (frequently) employ a rate code version of these spiking dynamics, which enables a single neuron to approximate the behavior of a population of spiking neurons. We recently discovered that our prior approach to capturing spiking behavior using a rate code model could be improved by driving the activation output from a different neural variable. Previously, we used the membrane potential, but now recognize that the rate of spiking in AdEx is best captured using the level of excitatory conductance directly (g_e), in relationship to a threshold that reflects the inhibitory and leak currents. We call this new activation function *gelin*, for "linear in g_e", and it results in more stable, systematic, and informative rate code activation dynamics, while preserving the same qualitative properties of the previous activation function, and therefore the underlying computational principles.

Learning rule

Insert Figure 5 about here.

A defining feature of Leabra is its integration of both error-driven and Hebbian ("associative") learning mechanisms, reflecting an attempt to balance several tradeoffs between these two mechanisms, and obtain the "best of both worlds" from models that have demonstrated the importance of each of these types of learning for different cognitive phenomena. Error-driven learning has proven indispensible for learning the complex cognitive mappings required for tasks such as object recognition, word pronunciation, and other similar challenging problems (O'Reilly, 1996a, 1998; O'Reilly & Munakata, 2000). Hebbian learning alone can account for some statistical learning in various domains, such as extracting the statistics of visual images in primary visual cortex (Olshausen & Field, 1996, 1997). The combination of these two forms of learning was originally achieved by simply adding together both learning algorithms (O'Reilly & Munakata, 2000). In what we consider an important new development, the latest version of the learning rule implements a much more integrated way of achieving this same objective using an elegant single learning rule that is directly and deeply grounded in the known biology of synaptic plasticity, and naturally results in both error-driven and Hebbian learning within a single framework.

Specifically, we leveraged a compellingly detailed and highly recommended model of *spike-timing dependent plasticity (STDP)* by Urakubo, Honda, Froemke, and Kuroda (2008) to extract a more comprehensive learning rule that is operative over longer time scales and larger neuronal populations. When initially discovered using spike pairs, it was found that STDP displayed an intriguing causal learning dynamic, where synaptic weights go up when the sending neuron fires before the receiving one, and down otherwise. However, it is becoming increasingly clear that this causal regime is not really very relevant for the kinds of complex extended spike trains that are typical within cortical networks (Froemke & Dan, 2002; Rubin, Gerkin, Bi, & Chow, 2005; Shouval, Wang, & Wittenberg, 2010; Wang, Gerkin, Nauen, & Bi, 2005). For example, increasing spike complexity to just triplets or quadruplets shows that the simple causal pairwise dynamic does not generalize (Froemke & Dan, 2002; Rubin et al., 2005; Wang et al., 2005). We wondered what would happen if we presented temporally-extended spike trains of different frequencies and durations to the Urakubo et al. (2008) model. To find out, we presented a wide range of Poisson spike trains of sending and receiving activity to the model, and measured the pattern of synaptic plasticity that resulted. Somewhat to our surprise, we were able to fit the results with a simple piecewise-linear function that captured roughly 80% of the variance in the synaptic plasticity in terms of the product of the sending and receiving net activity (spiking frequency times duration; Figure 5).

This function is essentially a linearized version of the Bienenstock, Cooper, and Munro (1982) learning rule *(BCM)*. BCM also introduced a floating threshold that imposes a long term homeostatic dynamic on

top of a fast Hebbian-like learning dynamic: weight changes fundamentally track the co-activation of the receiving and sending neurons ("neurons that fire together wire together"). If a receiving neuron is overly active over a long time scale, then the threshold moves proportionally higher, causing weights to be more likely to go down than up, thus preventing neurons from "hogging" the representational space. A reverse dynamic obtains for chronically under-active neurons, causing their threshold to move down, and making their weights more likely to increase, bringing them back into the game.

Thus, a simple piecewise-linear learning rule initially extracted from the Urakubo et al. (2008) model immediately captured a sophisticated and high-performing version of Hebbian learning. What about the error-driven component? We realized that error-driven learning could be obtained from this equation if the floating threshold also moved on a much more rapid time scale, such that the threshold reflects an *expectation* state in comparison to an *outcome* state reflected in the synaptic net activity value that drives learning. To illustrate how this achieves error-driven learning, consider two neurons that together are activated as part of a network encoding an incorrect *dishtowel*. Huh? You probably didn't expect that word — hopefully you were expecting to read the word *expectation* — there is considerable evidence that we are constantly forming these expectations, and we exhibit characteristic brain activity patterns when they are violated. Anyway, we assume that these two neurons were encoding the word *expectation*, and they would have high synaptic activity for a while as the expectation of this word develops, only to become inhibited by the activation of the actual outcome "dishtowel" neurons, resulting in subsequent low synaptic activity. The expectation activity causes the floating threshold to move up proportionally, and when the actual outcome activation comes in, it is below this expectation resulting in a reduction of synaptic weights, and thus a reduced tendency to make this expectation in this situation next time around. In contrast, the actual outcome "dishtowel" neurons have a low expectation activity, so their subsequent outcome activity exceeds this threshold and the weights increase, increasing the expectation of this word next time around. Despite the silly nature of this example (typically the outcomes we experience in the world are more predictable and useful sources of learning), one can hopefully see how this achieves robust error-driven learning, which is known to be capable of learning cognitively challenging problems.

To achieve an integration of this error-driven learning dynamic with a Hebbian self-organizing learning dynamic, one only needs to combine the BCM-like slowly adapting threshold with the error-driven fast-adapting threshold, resulting in a single overall threshold value. Thus, the threshold moves at multiple different superimposed time constants, and hence achieves a balance of both error-driven and Hebbian learning. Furthermore, consistent with the extensive work with the BCM algorithm, this form of Hebbian learning is actually more powerful and robust than the standard form of Hebbian learning used in Leabra previously (Blair, Intrator, Shouval, & Cooper, 1998).

Another way of thinking about this process is in terms of attractor dynamics and LTP/LTD (long term potentiation and depression). Essentially, the synaptic states associated with later activation states (settled fixed point attractors) always and continuously trains synaptic states associated with activations immediately prior during the earlier stages of settling. For this reason, and the different time scales used in the equations, we call this new learning mechanism the temporally eXtended Contrastive Attractor Learning (XCAL) rule.

Inhibition Function

Insert Figure 6 about here.

Beyond its importance for keeping the bidirectional excitatory loops between pyramidal neurons in check, inhibition in the neocortex has important computational implications. For example, it causes pyramidal

neurons to compete with each other for the opportunity to represent the current inputs. This competition in turn produces many of the effects of Darwinian evolution: neurons learn to specialize on representing a specific "niche" of input patterns, producing more differentiated and informative overall representations (Edelman, 1987). This competitive learning dynamic has been leveraged in a number of neural network models (Jacobs, Jordan, Nowlan, & Hinton, 1991; Kohonen, 1977, 1984; Nowlan, 1990; Rumelhart & Zipser, 1986), but it is notably absent in the backpropagation framework (although a recent model was able to add it, with some difficulty: Laszlo & Plaut, 2012).

There are five major paradigms of competitive inhibition that have been developed, including the null case:

- **Independence (fully distributed):** The activation of each neural unit is completely independent of the others, i.e., there is no inhibitory competition at all — this is easy to analyze mathematically, and automatically allows for complex distributed, overlapping patterns of neural activity to encode information, which has numerous computational advantages in efficiency, generalization, etc. (Rumelhart, Hinton, & Williams, 1986a). However, it obviously foregoes any of the advantages of competitive inhibition in creating more specialized, finely-tuned representations.

- **Winner-Takes-All (WTA):** A single neural unit within a layer (pool) of competing units is selected to be active (typically the one with the highest level of excitatory input). This is easy to implement computationally, but greatly restricts the power of the representation — a single unit cannot encode similarity in terms of relative degree of overlap, and it cannot easily support generalization to novel instances, which typically requires novel combinations of distributed neural activity.

- **WTA with topography:** The neighboring units around the winning one are also activated, typically with a gaussian normal "bump". This was pioneered by Kohonen (1984) and produces a topographically-organized distribution of representations. But, since the active units are not independent, it does not allow for differential activation of the units in a different context, and thus is not nearly as powerful as a distributed pattern of activity for encoding similarity in a high-dimensional space, or generalization to novel instances.

- **Normalization with contrast enhancement (softmax):** The activations of all units in a layer are normalized to sum to a constant value (typically 1), often with a contrast-enhancing nonlinearity (e.g., an exponential function) applied to produce a more differentiated pattern of resulting activity. This can also be thought of as a "soft" form of the WTA function (Nowlan, 1990), and sometimes a single winning unit is selected by using the normalized values as a probability distribution, instead of using the raw normalized values as rate-code like activations. This fundamentally has the same constraints as WTA, even though the activity distributions can be more graded across units — it is difficult to obtain a stable distributed pattern of activation across the units to encode high-dimensional similarity and generalize to novel cases.

- **kWTA (sparse distributed, used in Leabra):** A target number $k >= 1$ of neural units within a layer are allowed to be active, enabling a sparse but still distributed pattern of activity within the layer. This represents a balance between fully distributed and fully competitive dynamics, and is another example of a balance employed in the Leabra algorithm to obtain the best of both worlds. The multiple active neural units can encode high-dimensional similarity and support generalization in the form of novel combinations of active units, but there is also a competitive pressure that causes neurons to specialize more than in the fully independent case. The computational advantages of sparse distributed representations have been explored in depth by Olshausen and Field (1996, 1997).

- **Inhibitory interneurons:** The inhibitory circuits in neocortex can be simulated directly, resulting in more complex and potentially realistic dynamics than kWTA. Such a biologically detailed model is

considerably more computationally expensive, requiring significantly slower rate constants to avoid oscillatory dynamics from the feedback loops present, in addition to the greater number of neurons and neural connections.

The kWTA function in Leabra is implemented in a very computationally efficient manner, resulting in very low extra computational cost relative to having no inhibition at all. This is achieved with an optimized partial sort of the neurons in a layer according to the amount of inhibition that would be required to put each neuron exactly at its firing threshold, creating two groups: those within the top k and the remainder (Figure 6). In the most commonly used kWTA variant, a global level of inhibition within a layer is computed as some fraction of the way between the average of this threshold-level inhibition for the top k versus the average of the remainder. This tends to result in the top k neurons being above their firing thresholds, while the remainder are below, but there is considerable flexibility in the actual levels of activity depending on the exact distribution of excitation throughout the layer. This flexibility enables more appropriate representations to develop through learning, compared to requiring an exactly fixed level of activity for each input pattern.

Across many models of different cognitive phenomena, this kWTA inhibition function has proven to be one of the most important features of the Leabra architecture, rivaling or perhaps even exceeding the nature of the learning rule in importance for producing powerful learning that generalizes well to new situations. It is also one of the most distinctive aspects of the architecture — we are not aware of another major computational modeling framework with this form of inhibition function.

In keeping with the multi-level modeling principle, it is also possible to run Leabra networks with explicit inhibitory interneurons, and bridging simulations have been developed that establish the convergence between the more biologically detailed models with inhibitory interneurons and those using the kWTA inhibition function abstraction. However, these more detailed models also may exhibit important differences in overall activation dynamics — for example there is typically more of a wave of excitation driven by a new input pattern that is then damped down, with some ongoing oscillations — these waves have been observed in recordings from neocortical neurons, and may have important functional implications. In contrast, the kWTA dynamics are more tightly controlled, but we have also added the option of superimposing these wave dynamics on top of kWTA — these waves can improve learning in some situations (Norman, Newman, Detre, & Polyn, 2006), but more work remains to be done to explore the issues.

The Macrostructure of Cognition: Brain Area Functional Specializations

The principles and mechanisms just described characterize the microstructure of cognition — how cognition operates at the finest scale of individual neurons and synapses. There are also many important things that could be said about the mesolevel of analysis (network dynamics) (see Figure 3), but these are primarily emergent properties of the microlevel mechanisms (e.g., attractor dynamics, categorization), so they are not as essential for describing the major defining features of the Leabra architecture. Thus, we now turn to the macrolevel structure and how different brain areas are specialized for different aspects of cognitive function. Some relevant questions here include: is there any relationship between the micro and macro levels? Along what kind of dimensions are brain areas specialized: by content domain, by processing style, or by modular cognitive building blocks? In other words, what are the big chunks of cognition in the brain, the combined contributions of which can explain the full spectrum of cognitive abilities? To address these important questions, we again begin by enumerating four additional principles, which will help clarify the stance we have taken in Leabra. The first overarching principle concerns the relationship between the microstructure and macrostructure:

Principle 10 (Micro-macro interactions): *The microstructural principles and associated mechanisms characterize the fabric of cognition, so they also define the space over which macrostructural specializations can take place — in other words, we should be able to define different specialized brain areas in terms of different parameterizations of the microstructural mechanisms. Furthermore, the system is fundamentally still just a giant neural network operating according to the microstructural principles, so brain areas are likely to be mutually interactive and interdependent upon each other in any given cognitive task.*

This principle implies a more subtle form of specialization than is typically offered in cognitive theorizing: parametric differences typically do not lead to the kinds of discrete cognitive functions popular in traditional box-and-arrow information processing models of cognition.

Insert Figure 7 about here.

The broadest macrostructural organization of the Leabra architecture is shown in Figure 7, where each of the three major components of the system (posterior cortex, prefrontal cortex, and hippocampus) are defined by parametric specializations relative to the generic microstructural mechanisms described above. The posterior cortex is characterized by coarse-coded distributed overlapping representations that learn slowly over time to encode the world in an efficient way using hierarchically structured, specialized neural pathways. These pathways support basic functions such as object recognition, perceptually-guided motor control, auditory processing, language comprehension, and higher-level semantic knowledge. This system is well captured by a "generic" Leabra neural network with roughly 15-25% activity levels in the kWTA inhibition function, and relatively slow learning rates, which enable the system to integrate over many different experiences to extract these useful representations.

Relative to this posterior cortical baseline, the hippocampus and prefrontal cortex each have different parametric specializations that enable them to do things that the posterior cortex cannot, because of important fundamental tradeoffs (c.f., Principle #1) that are enumerated in the principles described below.

Learning and Memory Specializations: Hippocampus vs. Cortex

Insert Figure 8 about here.

Insert Figure 9 about here.

We can identify a set of functional tradeoffs in learning and memory that motivate the understanding about how the hippocampus (Figure 8) is specialized for episodic memory relative to the more semantic forms of memory supported by the posterior cortex.

Principle 11 (Interference and overlap): *Learning new information can interfere with existing memories to the extent that the same neurons and synapses are reused — this directly overwrites the prior synaptic knowledge. Hence, the rapid learning of new information with minimal interference requires minimizing the neural overlap between memories.*

Principle 12 (Pattern separation and sparseness): *Increasing the level of inhibitory competition among neurons, which produces correspondingly more sparse patterns of activity, results in reduced overlap (i.e., increased pattern separation) (Figure 9).*

Intuitively, pattern separation arises because the odds of a neuron exceeding a high threshold twice (assuming statistical independence) is like squaring a low probability — it goes down quadratically (Marr, 1971). For example, with a 1% chance of getting active, the probability of doing it twice is $0.01^2 = 0.0001$ — a very small number.

Principle 13 (Tradeoffs in separation vs. overlap): *While increasingly sparse representations result in decreased interference through pattern separation, they also reduce the ability to generalize knowledge across experiences, for the same reason — when different neurons and synapses encode each experience, then there is no opportunity to integrate across them (e.g., to extract statistical patterns).*

This tradeoff implies that achieving both of these learning goals (memorizing specifics and extracting generalities) requires two different systems, one with sparse representations for memorizing specifics, and another with overlapping distributed representations for extracting generalities (McClelland et al., 1995; Sherry & Schacter, 1987).

These principles provide a compelling explanation for the properties of the hippocampus for memorizing specific information including specific episodes (i.e., episodic memory), in contrast to a neocortical network that uses overlapping distributed representations to extract more generalized semantic information about the world. The CA3, CA1, and especially DG layers of the hippocampus have very sparse levels of activity, and corresponding pattern separation has been demonstrated through a variety of techniques (Gilbert, Kesner, & Lee, 2001; Leutgeb, Leutgeb, Moser, & Moser, 2007; McHugh, Jones, Quinn, Balthasar, Coppari, Elmquist, Lowell, Fanselow, Wilson, & Tonegawa, 2007; Bakker, Kirwan, Miller, & Stark, 2008). See O'Reilly et al. (2011) for a recent review of all the evidence consistent with this *complementary learning systems* account of the difference between hippocampus and neocortex.

In the latest version of the Leabra architecture, we have developed a more powerful version of hippocampal learning, which leverages the different theta phase relationships of the hippocampal layers to drive error-driven learning (Ketz & O'Reilly, in preparation), instead of relying on purely Hebbian learning, which has been a feature of most computational models of the hippocampus. In brief, this new model contrasts the retrieved pattern with the pattern to be encoded and uses the difference as an error signal, which trains subsequent retrieval in just the ways it needs to be modified to be more accurate, without the less selective and therefore more interference-prone Hebbian associative learning. In addition, these theta phase dynamics also drive error-driven learning of the invertible decoder pathway between CA1 and EC, which is necessary for recalling hippocampal memories back into the "language" of the cortex. This model has significantly higher capacity than a comparable Hebbian model (Ketz & O'Reilly, in preparation).

There are many important implications of the combined hippocampal and neocortical learning systems for behavior of the overall Leabra architecture. The hippocampus enables rapid (as fast as a single trial) encoding of arbitrary combinations of information. It also automatically contextualizes information, binding everything occurring at a given point in time together (since it receives information from most higher cortical areas). This enables behavior to be appropriately context-sensitive, preventing over-generalization. For example, negative outcomes can be appropriately contextualized via the hippocampus, preventing a generalized state of anxiety from pervading the system. In addition, the hippocampal system is also constantly and automatically retrieving prior memories as triggered by the current inputs — this provides an important source of constraint and background knowledge for many situations.

Active Maintenance and Executive Function Specializations: Frontal & Basal Ganglia vs. Posterior Cortex

Insert Figure 10 about here.

Another critical tradeoff motivates the architectural distinction between the frontal cortex versus the posterior cortex, in terms of the neural specializations required to sustain information in an active state (i.e., ongoing neural firing). First, we note that maintenance of information in a neural network (over at least a short time period) can be supported by either sustained neural firing of a population of neurons, or by synaptic weight changes. What are the relative tradeoffs between these two forms of information maintenance, and what kinds of neural specializations are required to support the maintenance of active neural firing? Again, we start with two more principles.

Principle 14 (Activation-based memory is more flexible than weight-based memory changes, and crucial for exerting top-down control): *Changes in neural firing can generally happen faster and have broader and more general effects than weight changes.*

Changes in neural firing are much more flexible than weight changes because a new state can be rapidly activated to replace an old one, whereas weight changes typically require multiple iterations to accumulate before there is a measurable impact. Furthermore, active neural firing can immediately and directly influence the activity states of other neurons in the network (top-down biasing), whereas weight changes are latent most of the time and require the (re)activation of those same neurons to exert a biasing effect (Morton & Munakata, 2002).

We can distinguish the frontal cortex (especially the *prefrontal cortex, PFC*) from the posterior cortex in terms of an ability to robustly maintain information using active neural firing over time. There are multiple specialized neural mechanisms in the PFC relative to posterior cortex that support this ability (Wang, Markram, Goodman, Berger, Ma, & Goldman-Rakic, 2006; Hazy, Pauli, Herd, others, & O'Reilly, in preparation), and it is long-established that PFC neurons exhibit this active maintenance property (Fuster & Alexander, 1971; Goldman-Rakic, 1995; Kubota & Niki, 1971; Miller, Erickson, & Desimone, 1996; Miyashita & Chang, 1988). This specialization for active maintenance is then consistent with the observed importance of the PFC in supporting cognitive flexibility (e.g., in task shifting, overcoming prepotent responding, and other similar such cases), and for providing top-down excitatory biasing over processing in the posterior cortex, to guide behavior in a task-relevant manner (Braver & Cohen, 2000; Cohen, Dunbar, & McClelland, 1990; Cohen & Servan-Schreiber, 1989; Herd et al., 2006; Miller & Cohen, 2001). All of these functions of the PFC can be summarized with the term *executive function*, and an important contribution of the Leabra approach is to show how all of these different aspects of executive function can derive from a single set of neural specializations. This is an instance where the use of a big picture cognitive architecture provides an important and unique perspective, in contrast to developing specific models for different aspects of executive function.

Principle 15 (Tradeoff between updating and maintenance): *There is a tradeoff between the neural parameters that promote the stable (robust) active maintenance of information over time, and those that enable activity patterns to be rapidly updated in response to new inputs.*

Robust maintenance requires strong recurrent excitation among maintaining neurons, and/or strong intrinsic excitatory currents, relative to the drive from other inputs, so that the maintained information is not overwritten by new inputs. In contrast, rapid updating requires that those maintenance factors be weakened for external inputs to outcompete existing representations. Thus, there can be no static setting of parameters that will make a system capable of doing both robust maintenance and rapid updating in a general-purpose and ecologically adaptive way (while it would be possible to set parameters so as to rapidly update some information easily and robustly maintain *other* information, based on specific weight patterns, the rigidity of that approach would not be very useful).

Principle 16 (Dynamic gating): *A dynamic gating system can resolve the fundamental tradeoff between rapid updating and robust maintenance by dynamically switching between these two modes.*

The fundamental tradeoff between maintenance and updating makes it clear however that the PFC cannot do everything by itself — some kind of dynamic gating system is required (O'Reilly et al., 1999). We and others have argued that the basal ganglia is ideally situated to provide a dynamic gating signal to the frontal cortex (e.g., Frank et al., 2001). When the direct or *Go* pathway neurons fire, this (indirectly) triggers a burst of activation through the frontal-thalamic loop that results in a rapid updating of information in frontal cortex. Otherwise (e.g., when the indirect or *NoGo* pathway neurons fire), the frontal cortex can robustly maintain activity states over time. But how does the basal ganglia know when to fire Go? We have shown that the phasic dopamine signals associated with reward prediction errors can drive learning in the basal ganglia to solve this learning problem (O'Reilly & Frank, 2006). Thus, capturing the overall contributions of the PFC to executive function requires a complex interactive system (Figure 10), which we have implemented as the PBWM (*prefrontal cortex basal ganglia working memory*) system (O'Reilly & Frank, 2006; Hazy et al., 2006, 2007; Hazy et al., in preparation).

We placed the basal ganglia in the center of the macrostructural architecture (Figure 7) in part as a result of our collaboration with the ACT-R developers — the central engine driving the sequencing of cognitive actions in ACT-R is the *production system* component of the architecture, which they have associated with the basal ganglia. Interestingly, this notion of a production system (which chooses the next "cognitive action" based on the current context) as the core of the cognitive architecture was central to Newell's original 20-questions paper (Newell, 1973), and this idea does appear to have stood the test of time.

Thus, the model of executive function that emerges from this picture is a continuous sequence of dynamic and highly selective gating actions exquisitely modulated by the basal ganglia, continually updating the states of selected regions of neurons in the frontal cortex. These in turn provide an updated context and top-down biasing on other cortical areas, including much of the posterior cortex, according to whatever goals or plans are currently activated. Finally, at the brain-wide scale of the tripartite organization (Figure 7), the hippocampus is constantly encoding and retrieving information cued by this ongoing flow, and thus providing relevant knowledge and context to inform ongoing processing. There are also multiple mechanisms by which the PFC can provide more directed control over the encoding and retrieval processes in the hippocampus, to better deploy its considerable powers of learning and recall.

One critical missing piece from this picture is the origin of these goal and plan representations: how does the system decide what it wants to do, and develop overall plans of action to accomplish its goals? To understand more about this, we first provide an overarching picture about the organization of different representational content in the system.

What vs. How Content Specialization: Ventral vs. Dorsal Pathways

Insert Figure 11 about here.

Complementing the parametric specializations described above, we can also try to identify content-based specializations in the cognitive architecture: ways in which different parts of the neocortex are organized to process specific kinds of information. We begin with some motivating principles for thinking about why and how such a content-based organization might occur. To contextualize the first principle, it seems that people have an irrepressible urge to anthropomorphize, and think of neurons as tiny people, communicating using some kind of language, like two old ladies sitting on a park bench discussing the passers-by. For example, some researchers are engaged in a quest to discover the "neural code" — a putative language that neurons use to communicate with, typically thought to involve complex sequences of spikes (e.g., Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1996). A consequence of this kind of thinking is that people tend to assume that it is no problem for neurons to rapidly change what they are encoding (e.g.,

Miller, 2000; Duncan, 2001) — i.e., that neurons can just change the words that they send to the other neurons to effect this change.

Contrary to the anthropomorphic image, every indication is that pyramidal neurons simply aggregate over the vast pool of incoming information in a very generic way, like raindrops in a bucket, preventing the use of any kind of specialized neural language. This is certainly how the Leabra model operates. And yet it can perform very powerful forms of information processing under this strong constraint. Our next principle helps articulate how this happens, and after that, we see how this constrains the large scale functional organization of the brain, relative to a perspective that assumes that neurons use a form of language, and can rapidly change what they are encoding.

Principle 17 (Meaning is in the activity pattern across neurons, not the individual neural messages): *Meaning in a neural network is entirely derived from the patterns of activity across the population of input neurons ("receptive field") to a receiving neuron — each individual neuron only has meaning in relationship to other neurons, and this meaning must be learned over time by each neuron.*

Thus, we reject the notion of a neural code that posits meaning in individual neural signals, and accept the consequence that it is not possible for neurons to rapidly change what they encode — that would just confuse the other neurons (O'Reilly, 2010). Instead, neural representations must be relatively stable over time, to enable a given receiving neuron to properly learn the statistics of the patterns of activity over its inputs.

Principle 18 (Hierarchical stages required for complex processing): *Given the relatively simple detector-like functionality of individual neurons, multiple hierarchically-organized stages of processing are typically required to extract high-level information out of sensory input streams. Each stage of processing detects patterns of an incremental increase in complexity relative to the stage before, and this incremental decomposition of the problem can enable information to be extracted in ways that single stage transformations simply cannot support.*

These two principles together imply that there should be a relatively stable structural organization of information in the brain, where nearby populations of neurons process similar kinds of information, so that they can present an informative overall pattern of activity to other downstream neurons in a hierarchically-organized processing pathway. This conclusion converges with considerable empirical data on the nature of the pathways in the brain that process visual information in different ways. Two major pathways have been identified, one progressing through successive layers of the ventral visual pathway into the inferotemporal cortex (IT), and the other progressing through the dorsal pathway into the parietal cortex. The ventral pathway produces invariant representations of object identity over a succession of layers from V1, V2, V3, V4, aIT, to pIT. Computational models of this pathway, including a Leabra model called LVis, have shown how this hierarchy is important for computing complex object feature detectors that are also invariant to many irrelevant sources of variance in input images, such as position, rotation, size, illumination, etc (O'Reilly et al., submitted; Fukushima, 1980, 2003; Wallis & Rolls, 1997; Riesenhuber & Poggio, 1999; Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007; Mutch & Lowe, 2008). Other models of the parietal cortex demonstrate hierarchies that transform retinotopic visual inputs into the proper reference frames for driving motor control (Pouget & Sejnowski, 1997; Pouget, Deneve, & Duhamel, 2002).

Goodale and Milner (1992; Milner & Goodale, 1995, 2006) used other data, including striking dissociations in patients with brain damage, to argue for an overall *What* (ventral object recognition) vs. *How* (dorsal perception-for-action) division in posterior cortex, which is a refinement to the influential What vs. Where division suggested by Ungerleider and Mishkin (1982) (perception-for-action relies extensively, but not exclusively, on spatial representations). This what vs. how distinction is very broad, encompassing many more specialized sub-pathways within these overall divisions, and other pathways of

content-specific information exist as well, for example pathways for the other sensory modalities, and likely additional high-level semantic pathways, such as those involved in representing plots and story schemas.

The principles above also suggest that it would make sense for the brain to carry these specialized content processing pathways forward into the prefrontal cortex, as we recently argued (O'Reilly, 2010; Figure 11). This way, the prefrontal top-down control pathways can continue the hierarchical processing stages, resulting in even higher-level "executive" encodings of the different specialized pathways, which then provide a more effective basis for targeting top-down control. For example, we have shown that the active maintenance properties of the PFC, along with the dynamic gating mechanism provided by the BG, shapes PFC representations to encode more abstract rules or regularities (Rougier, Noelle, Braver, Cohen, & O'Reilly, 2005). Under this what vs. how organization in PFC, the dorsal lateral PFC (DLPFC) is specialized for executive control over sensory-motor processing, including likely sequencing and organization of motor plans. In contrast ventral lateral PFC (VLPFC) is more specialized for executive control over sensory processing that takes place in the IT cortex. Within both of these areas, increasingly anterior PFC areas are likely to contain higher-order, more abstracted representations, because the hierarchical connectivity continues through this axis. Overall, this organizational scheme is consistent with a wide range of data (O'Reilly, 2010), and it helps to integrate findings across many different specific task paradigms, and constrain one's interpretation of the functional contributions of these areas — exactly the kind of benefit a cognitive architecture should provide.

One of the more intriguing aspects of this what vs. how organizational theory comes in its application to motivational and affective systems, which include the medial surface of the frontal cortex, as discussed next.

Motivational and Affective Systems

The last missing piece from our overall cognitive architecture comes in the form of motivational and affective systems, which are critical for driving the system toward certain goals, and regulating overall behavioral state and learning processes in response to different kinds of environmental feedback. It is these systems which help to establish the goals that the executive function system works to achieve. Biologically, these systems are evolutionarily ancient, and there are many complex interacting systems that all seem at least partially redundant, making it extremely difficult to arrive at clear, compelling computational models. We begin with a few principles that can help organize our thinking to some extent.

Principle 19 (Interact and override): *As newer brain areas evolved on top of older ones, they generally have strong bidirectional interactive connections with the older areas, and leverage the more robust signals from the older areas to help train up the more flexible newer systems, while also having the ability to exert top-down control over the older systems through either directed or competitive inhibition (Munakata, Herd, Chatham, Depue, Banich, & O'Reilly, 2011).*

Principle 20 (Motivation and reward must be grounded): *As higher-order motivational and affective areas evolved to be more flexible and adaptive to the specific environmental context an individual finds themself in, the risk of motivations becoming maladaptive over the course of an individual's development emerged. The prevalence of suicide in humans is evidence that we have pushed this balance to the limit. Thus, there must be strong grounding constraints on the learning processes in these higher-order motivational systems — it is crucial that we cannot just make ourselves happy by willing it to be so.*

To explore the implications of these principles. we can start top-down in the evolutionary layer-cake of affective systems, beginning with the medial frontal areas that provide executive control over affective and motivational systems lower down. As a general rule in brain anatomy, the medial brain areas are associated with the "limbic system", and are primarily involved in motivational and affective activation, learning, and control, and this is the case with the medial frontal areas. As shown in Figure 11, the dorsal medial frontal

cortex contains the anterior cingulate cortex (ACC), while the ventral medial frontal areas (spreading over into ventral lateral) include the orbital frontal cortex (OFC), and there are also non-OFC areas generically labeled ventral medial PFC (VMPFC). According to the what vs. how dorsal/ventral distinction, we would expect the ACC to be important for motivational and affective control associated with motor control, while the OFC should be involved in motivational and affective control associated with objects, language, and other ventral pathway information.

Matthew Rushworth and colleagues have accumulated considerable data consistent with this What vs. How account, showing that ACC encodes "value" representations associated with different motor actions that an animal is considering, while OFC encodes more stimulus-driven value representations (Rushworth, Behrens, Rudebeck, & Walton, 2007; Rushworth, 2008). This division is also consistent with considerable data showing that the ACC is important for encoding error, conflict (uncertainty), and effort information — these are the affective states most relevant for evaluating different action choices. In contrast, OFC neurons have been shown to encode both unconditioned stimulus (US — i.e., reward outcome) information, along with conditioned stimuli (CS) that have been associated with these US's. Thus, it appears that the broad what vs. how dissociation can also help make sense of the medial frontal cortical organization.

Moving down a level in the hierarchy, the equivalent of posterior cortex in the affective domain is the basolateral amygdala (BLA), which is anatomically at the same level as the hippocampus in what is known as the "archicortex" or ancient cortex. The BLA is densely interconnected with the OFC and the ACC, and it is known to encode both US's and CS's. Some models of the BLA and OFC interactions suggest that the BLA helps train corresponding representations in the OFC, while OFC provides top-down biasing over BLA, resulting in enhanced flexibility during reversal learning for example (Frank & Claus, 2006; Pauli, Hazy, & O'Reilly, 2012). This dynamic is consistent with the principles outlined above. The BLA also interacts with a deeper structure known as the central nucleus of the amygdala (CNA), which then has extensive connectivity with ancient midbrain nuclei involved in all manner of basic bodily functions and states of arousal, pain, pleasure, etc.

One pathway through the CNA is involved in driving phasic dopmaine bursts in response to CS's, which forms a central part of the *Learned Value (LV)* system in our PVLV model (*Primary Value, Learned Value*) (O'Reilly et al., 2007; Hazy et al., 2010). This PVLV system explains how different brain areas contribute to the overall phenomenon of reward prediction error (RPE) signaling in the midbrain dopamine neurons, which then broadcast the neuromodulator dopamine throughout the brain. Dopamine has many effects on neurons in different brain areas, but rapid phasic changes in dopamine are highly likely to affect learning in the striatum of the basal ganglia, in a manner consistent with its gating role in the PBWM (Prefrontal Cortex & Basal Ganglia Working Memory) model as described earlier (Frank, 2005). Contrary to the popular impression, dopamine itself is unlikely to convey an affective pleasure signal throughout the brain, and should be thought of more as a learning or salience signal.

To summarize, the Leabra architecture at this point has a strong implementation of the dopaminergic system and its involvement in learning, and some initial implementations of the BLA / OFC system (Pauli et al., 2012). We are currently elaborating and refining these models, and developing an ACC model, to provide a more complete motivational and affective system. Interestingly, one of the most important functions we attribute to the ACC and OFC is an ability to track the rate of progress toward a goal, and to trigger the adoption of new strategies when the system becomes "frustrated" with its current progress. This system would account for similar functionality that is the cornerstone of Allen Newell's SOAR architecture, which has a universal subgoaling system that activates whenever the production system reaches an impasse. We also think that the motivational system will play a critical role in selecting goals and action plans that are within the current "zone of proximal development" of the system, corresponding in effect to a state of "curiosity" about things which the system would like to explore further (Herd, Mingus, & O'Reilly, 2010). Given our current experience with the PBWM system lacking these

motivational control systems, we are convinced that they are essential for enabling the system to be more robust and effective in solving problems. For example, the current system will continue to select actions that lead on average to suboptimal rewards, without properly exploring other options, despite the fact that it is making no progress overall in achieving greater levels of success. The network needs to experience some frustration for what it's currently doing, and curiosity for underexplored avenues.

Finally, coming back to principle #20, we follow Michael Tomasello in believing that much of the flexibility and power of human cognition comes from our strong social motivational systems (Tomasello, 2001). If you try to understand human motivations in terms of satisfying a desire for basic survival factors such as food and water, or even money, it doesn't really add up. There is no way someone would be a starving artist or a grad student under such a scenario. However, once you think in terms of social motivation, it all starts to make sense. We basically want to both share knowledge and experiences with others, and also show off for others. Furthermore, we have a strong in-group / out-group motivational dichotomy in our heads, which essentially aligns with the love / hate axis. And these groups can be high dimensional, encompassing everything from family, friends, school, sports teams, political party, nation, and species. These social motivations provide grounded primary reward signals, but are also highly flexible on a cultural level, enabling people as a group to adapt to different demands. There is much that remains to be understood in this area, but we believe that it is important for any accurate model of human cognition to take these social factors into account.

Conclusions

We hope that the explicit enumeration of a set of core principles underlying the Leabra cognitive architecture provides a clear sense of the motivations, priorities, and defining features of the architecture. As noted earlier, we refer the reader to our online textbook `http://ccnbook.colorado.edu` (O'Reilly et al., 2012) for a more complete development of these ideas, and their specific implementation in computational models.

You may have some lingering questions about the precise relationship between the principles articulated here, the more specific theoretical instantiation of the Leabra architecture as reflected in specific models and papers, and the detailed implementation of Leabra in the current version of the simulation software. Which is the official definition of the architecture? What happens when the architecture changes over time — does that invalidate earlier models? Can anyone contribute to the development of the architecture? Is Leabra just a label for an ever-expanding theory of human cognition, or do the existing principles set clear limits on how it might expand in the future?

As is implicit in the principles enumerated above, there is not one privileged level of description, and hence we seek convergent multi-level descriptions of the nature of the Leabra architecture as well — it is simultaneously and at different levels all of the above three things (principles, specific theories, and implementation), each of which mutually informs and constrains the others. Thus, principles shape the overall structure of the architecture, while specific models and theories about particular brain areas or cognitive functions test the applicability of the principles, and provide new insights that can be incorporated back into the overall architecture. Many times important questions are raised in the process of the software implementation, and computational results strongly inform us about what works and what does not work to actually solve particular problems. And, similarly, important questions and solutions are discovered in the process of trying to understand the actual biological mechanisms. Thus, in many ways, the architecture represents a kind of aggregation and clearinghouse for integrating new advances into a coherent and competent framework.

Returning to the overarching question raised in the introduction: why would anyone invest the effort to understand this complex cognitive architecture? We hope to have convinced you that it does account for a

wide range of data at multiple levels of analysis, in a principled and internally-consistent manner, building upon a small set of microstructural principles up to a macrostructural organization of the brain. But even if you are not moved to build models in Leabra yourself, you can nevertheless benefit by borrowing from the various ideas and models that have been developed. For example, many people cite the Leabra work on error-driven learning to support their use of error backpropagation models. This is reasonable, given that such models do provide a rough approximation to the learning that we argue is actually supported by the neocortex. Similarly, there are many different abstract computational implementations of the core ideas behind the PBWM model of prefrontal cortex / basal ganglia working memory (O'Reilly, Herd, & Pauli, 2010), which can leverage the biological connections that the PBWM model makes.

Future Directions

Clearly, Leabra is a work in progress, with many important challenges ahead, and we welcome contributions from anyone — as should be evident, we gladly steal the best ideas wherever we can find them (giving proper attribution of course). We do think that the existing set of principles, theories, and software provide a solid foundation upon which to build — one that will strongly inform and constrain future progress. Some specific areas where we see development going in the next few years include:

- **Embodied robotics:** A powerful neural architecture for a robot can be achieved by combining an improved version of our existing object recognition model (O'Reilly et al., submitted) that is capable of robust figure-ground segregation, with neurally-driven motor control systems based on the biology of the cerebellum, parietal cortex, and frontal cortex. We are working on all of these components, and are excited to discover how much of cognition can be grounded in the kinds of learning that can occur in the realistic sensory-motor interaction of a robot with its environment.

- **Emotion, motivation, and control:** As described above, there is much to be done to understand how the "limbic" brain areas interact with the cognitive processing that we have focused more on to date. We are excited to understand more about the contributions of the anterior cingulate cortex (ACC) and orbital frontal cortex (OFC) to decision making, goal setting, and action planning, through the development of detailed Leabra-based models of these areas, and their interactions with associated subcortical areas.

- **Temporal processing:** To simplify and enable more rapid progress, we have tended to simplify many aspects of the complex temporal dynamics that are undoubtedly important for neural processing in the brain. It is time to circle back and revisit some of these issues, to see where tractable progress can be made. For example, we are currently exploring the idea that interconnectivity between the deep layers of the neocortex and the thalamus could produce a dynamic much like that present in a simple recurrent network (SRN) (Elman, 1990). This would provide a powerful mechanism for integrating information over time, leveraging learned synaptic weights and high-dimensional coarse-coded distributed representations to determine how to integrate prior context and current inputs. In contrast, many models put a simple parameter on how sluggish vs. responsive the neurons are, which inevitably leads to difficult tradeoffs and unsatisfactory parameter dependence.

There are many other important challenges ahead in addition to these specific ones, and again we welcome contributions!

References

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

Bakker, A., Kirwan, B. C., Miller, M., & Stark, C. E. (2008). Pattern separation in the human hippocampal CA3 and dentate gyrus. *Science*, *319*(5870), 1640–1642.

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. *The Journal of Neuroscience*, *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 computation*, *10*, 1797–1813.

Braver, T. S., & Cohen, J. D. (2000). On the control of control: The role of dopamine in regulating prefrontal function and working memory. In S. Monsell, & J. Driver (Eds.), *Control of Cognitive Processes: Attention and Performance XVIII* (pp. 713–737). Cambridge, MA: MIT Press.

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

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

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing model of the Stroop effect. *Psychological Review*, *97*(3), 332–361.

Cohen, J. D., & Servan-Schreiber, D. (1989). *A Parallel Distributed Processing approach to behavior and biology in schizophrenia* (Artificial Intelligence and Psychology Project AIP-100). Carnegie Mellon University.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193–222.

Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews Neuroscience*, *2*, 820–829.

Edelman, G. (1987). *Neural darwinism*. New York: Basic Books.

Eliasmith, C., & Anderson, C. H. (2003). *Neural Engineering: Computation, Representation and Dynamics in Neurobiological Systems* . Cambridge, MA: MIT Press.

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

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

Frank, M. J., & Claus, E. D. (2006). Anatomy of a decision: Striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review*, *113*(2), 300–326.

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

Froemke, R. C., & Dan, Y. (2002). Spike-timing-dependent synaptic modification induced by natural spike trains. *Nature*, *416*(6879), 433–437.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202.

Fukushima, K. (2003). Neocognitron for handwritten digit recognition. *Neurocomputing*, *51*(1), 161–180.

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

Gilbert, P. E., Kesner, R. P., & Lee, I. (2001). Dissociating hippocampal subregions: A double dissociation between dentate gyrus and CA1. *Hippocampus*, *11*, 626–636.

Goldman-Rakic, P. S. (1995). Cellular basis of working memory. *Neuron*, *14*(3), 477–485.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25.

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2006). Banishing the homunculus: Making working memory work. *Neuroscience*, *139*, 105–118.

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2007). Towards an executive without a homunculus: Computational models of the prefrontal cortex/basal ganglia system. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *362*(1), 105–118.

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2010). Neural mechanisms of acquired phasic dopamine responses in learning. *Neuroscience and Biobehavioral Reviews*, *34*(5), 701–720.

Hazy, T. E., Pauli, W., Herd, S., others, & O'Reilly, R. C. (in preparation). Neural mechanisms of executive function: Biological substrates of active maintenance and adaptive updating.

Herd, S., Mingus, B., & O'Reilly, R. (2010). Dopamine and self-directed learning. *Biologically Inspired Cognitive Architectures 2010: Proceedings of the First Annual Meeting of the BICA Society*, *221*, 58–63.

Herd, S. A., Banich, M. T., & O'Reilly, R. C. (2006). Neural mechanisms of cognitive control: An integrative model of Stroop task performance and fMRI data. *Journal of Cognitive Neuroscience*, *18*, 22–32.

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

Izhikevich, E. M., & Edelman, G. M. (2008). Large-scale model of mammalian thalamocortical systems. *Proceedings of the National Academy of Sciences of the United States of America*, *105*.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, *3*, 79–87.

Jilk, D., Lebiere, C., O'Reilly, R., & Anderson, J. (2008). SAL: an explicitly pluralistic cognitive architecture. *Journal of Experimental & Theoretical Artificial Intelligence*, *20*(3), 197–218.

Ketz, N., & O'Reilly, R. C. (in preparation). Error-driven learning substantially improves hippocampal capacity.

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

Kohonen, T. (1984). *Self-organization and associative memory*. Berlin: Springer Verlag.

Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: The MIT Press.

Kubota, K., & Niki, H. (1971). Prefrontal cortical unit activity and delayed alternation performance in monkeys. *Journal of Neurophysiology*, *34*(3), 337–347.

Laszlo, S., & Plaut, D. C. (2012). A neurally plausible parallel distributed processing model of event-related potential word reading data. *Brain and language*, *120*(3), 271–81.

Lebiere, C., & Anderson, J. R. (1993). A connectionist implementation of the ACT-R production system. *Proceedings of the 15th Annual Cconference of the Cognitive Science Society,*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Leutgeb, J. K., Leutgeb, S., Moser, M., & Moser, E. (2007). Pattern separation in the dentate gyrus and CA3 of the hippocampus. *Science*, *315*(5814), 961–966.

Markram, H. (2006). The blue brain project. *Nature Reviews Neuroscience*, *7*(2), 153–160.

Marr, D. (1971). Simple memory: a theory for archicortex. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *262*(841), 23–81.

McClelland, J. L. (1993). The GRAIN model: A framework for modeling the dynamics of information processing. In D. E. Meyer, & S. Kornblum (Eds.), *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience* (pp. 655–688). Hillsdale, NJ: Lawrence Erlbaum Associates.

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. *Psychological Review*, *102*(3), 419–457.

McClelland, J. L., & Rumelhart, D. E. (Eds.). (1988). *Explorations in Parallel Distributed Processing: A Handbook of Models, Programs, and Exercises*. Cambridge, MA: MIT Press.

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.

McHugh, T. J., Jones, M. W., Quinn, J. J., Balthasar, N., Coppari, R., Elmquist, J. K., Lowell, B. B., Fanselow, M. S., Wilson, M. A., & Tonegawa, S. (2007). Dentate gyrus NMDA receptors mediate rapid pattern separation in the hippocampal network. *Science*, *317*(5834), 94–99.

Miller, E. K. (2000). The prefrontal cortex and cognitive control. *Nature Reviews Neuroscience*, *1*, 59–65.

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

Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural mechanisms of visual working memory in prefontal cortex of the macaque. *The Journal of Neuroscience*, *16*, 5154–5167.

Milner, A. D., & Goodale, M. A. (1995). *The Visual Brain in Action* . Oxford University Press, 1 edition.

Milner, A. D., & Goodale, M. A. (2006). *The Visual Brain in Action* . Oxford University Press, 2 edition.

Miyashita, Y., & Chang, H. S. (1988). Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature*, *331*, 68–70.

Morton, J. B., & Munakata, Y. (2002). Active versus latent representations: a neural network model of perseveration, dissociation, and decalage. *Developmental Psychobiology*, *40*(3), 255–265.

Munakata, Y., Herd, S. A., Chatham, C. H., Depue, B. E., Banich, M. T., & O'Reilly, R. C. (2011). A unified framework for inhibitory control. *Trends in Cognitive Sciences*, *15*(10), 453–459.

Mutch, J., & Lowe, D. (2008). Object class recognition and localization using sparse features with limited receptive fields. *International Journal of Computer Vision*, *80*(1), 45–57.

Newell, A. (1973, January). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual Information Processing* (pp. 283–308). New York: Academic Press.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Norman, K. A., Newman, E., Detre, G., & Polyn, S. (2006). How inhibitory oscillations can train neural networks and punish competitors. *Neural computation*, *18*.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychological Review*, *110*(4), 611–646.

Nowlan, S. J. (1990). Maximum likelihood competitive learning. In D. S. Touretzky (Ed.), *Advances in neural information processing systems, 2* (pp. 574–582). San Mateo, CA: Morgan Kaufmann.

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 Research*, *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. (1996a). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation*, *8*(5), 895–938.

O'Reilly, R. C. (1996b). *The Leabra model of neural interactions and learning in the neocortex*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA.

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

O'Reilly, R. C. (2010). The *what* and *how* of prefrontal cortical organization. *Trends in Neurosciences*, *33*(8), 355–361.

O'Reilly, R. C., Bhattacharyya, R., Howard, M. D., & Ketz, N. (2011). Complementary learning systems. *Cognitive Science*, Epub ahead of print.

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 Computation*, *18*(2), 283–328.

O'Reilly, R. C., Frank, M. J., Hazy, T. E., & Watz, B. (2007). PVLV: The primary value and learned value Pavlovian learning algorithm. *Behavioral Neuroscience*, *121*, 31–49.

O'Reilly, R. C., Herd, S. A., & Pauli, W. M. (2010). Computational models of cognitive control. *Current opinion in neurobiology*, *20*, 257–261.

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: The 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., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: principles of cortical and hippocampal function. *Psychological Review*, *108*(2), 311–345.

O'Reilly, R. C., Wyatte, D., Herd, S. A., Mingus, B., & Jilk, D. J. (submitted). Recurrent processing in object recognition.

Pauli, W. M., Hazy, T. E., & O'Reilly, R. C. (2012). Expectancy, ambiguity, and behavioral flexibility: separable and complementary roles of the orbital frontal cortex and amygdala in processing reward expectancies. *Journal of Cognitive Neuroscience*, *24*(2), 351–366.

Pouget, A., Deneve, S., & Duhamel, J.-R. (2002). A computational perspective on the neural basis of multisensory spatial representations. *Nature reviews. Neuroscience*, *3*, 741–747.

Pouget, A., & Sejnowski, T. J. (1997). Spatial transformations in the parietal cortex using basis functions. *Journal of Cognitive Neuroscience*, *9*, 222.

Rieke, F., Warland, D., de Ruyter van Steveninck, R., & Bialek, W. (1996). *Spikes: Exploring the Neural Code* . A Bradford Book MIT Press.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

Rougier, N. P., Noelle, D., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal cortex and the flexibility of cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences*, *102*(20), 7338–7343.

Rubin, J. E., Gerkin, R. C., Bi, G.-Q., & Chow, C. C. (2005). Calcium time course as a signal for spike-timing-dependent plasticity. *Journal of Neurophysiology*, *93*(5), 2600–2613.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986a). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1: Foundations (pp. 318–362). Cambridge, MA, USA: MIT Press.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (Eds.). (1986b). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol.1: Foundations* , Vol. 1: Foundations. Cambridge, MA: MIT Press.

Rumelhart, D. E., & Zipser, D. (1986). Feature discovery by competitive learning. In D. E. Rumelhart, J. L. McClelland, & P. R. Group (Eds.), *Parallel distributed processing. volume 1: Foundations* (Chap. 5, pp. 151–193). Cambridge, MA: MIT Press.

Rushworth, M. F. S. (2008). Intention, choice, and the medial frontal cortex. *Annals of the New York Academy of Sciences*, *1124*.

Rushworth, M. F. S., Behrens, T. E. J., Rudebeck, P. H., & Walton, M. E. (2007). Contrasting roles for cingulate and orbitofrontal cortex in decisions and social behaviour. *Trends in Cognitive Sciences*, *11*(4), 168–176.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(3), 411–426.

Sherry, D. F., & Schacter, D. L. (1987). The evolution of multiple memory systems. *Psychological Review*, *94*(4), 439–454.

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

Snyder, H. R., Hutchison, N., Nyhus, E., Curran, T., Banich, M. T., & Munakata, Y. (2010). Neural inhibition enables selection during language processing. *Proceedings of the National Academy of Sciences*, *107*, 16483–16488.

Tomasello, M. (2001). *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press.

Traub, R. D., Miles, R., & Wong, R. K. (1989). Model of the origin of rhythmic population oscillations in the hippocampal slice. *Science (New York, N.Y.)*, *243*, 1319–1325.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *The Analysis of Visual Behavior* (pp. 549–586). Cambridge, MA: MIT Press.

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

Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology*, *51*(2), 167–194.

Wang, H.-X., Gerkin, R. C., Nauen, D. W., & Bi, G.-Q. (2005). Coactivation and timing-dependent integration of synaptic potentiation and depression. *Nature Neuroscience*, *8*(2), 187–193.

Wang, Y., Markram, H., Goodman, P. H., Berger, T. K., Ma, J., & Goldman-Rakic, P. S. (2006). Heterogeneity in the pyramidal network of the medial prefrontal cortex. *Nature Neuroscience*, *9*(4), 534–542.

Figure Captions

Figure 1. An implemented Leabra cognitive architecture model, for the ICArUS (integrated cognitive-neuroscience architectures for understanding sensemaking) project, which has a posterior cortex, hippocampus, prefrontal cortex & basal ganglia, medial frontal cortex areas (ACC, OFC), and various other subcortical systems.

Figure 2. The *emer* virtual robot simulation, with a detailed and high-functioning visual pathway including figure-ground processing in area V2, supported by top-down connections from V4 and area MT, which provide higher-level gestalt constraints to the figure-ground problem of identifying objects in the presence of background clutter. The glass brain visualization on the upper left projects simulated neural activity into the anatomical locations of simulated brain areas, for easier direct comparison with neuroimaging and other data.

Figure 3. Four levels of analysis of the cognitive architecture, which organize and frame our discussion. The metalevel is a catch-all for any kind of abstract analysis that is not directly tied to the structure of the brain, and the remaining three levels represent different structural levels of analysis going from the level of individual neurons (micro) to networks of neurons (meso) to large-scale brain area organization (macro).

Figure 4. The core microstructural properties of the Leabra architecture.

Figure 5. The XCAL weight change function, plotting change in synaptic weight against total synaptic activation (sender times receiver activation).

Figure 6. Average-based kWTA inhibition function.

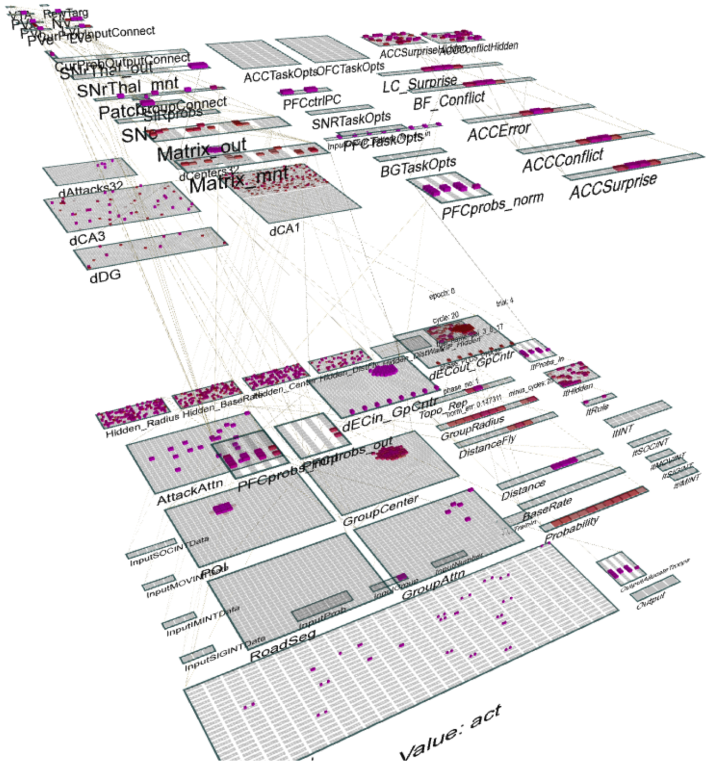
Figure 7. The macrostructure of the Leabra architecture, with specialized brain areas interacting to produce overall cognitive function.

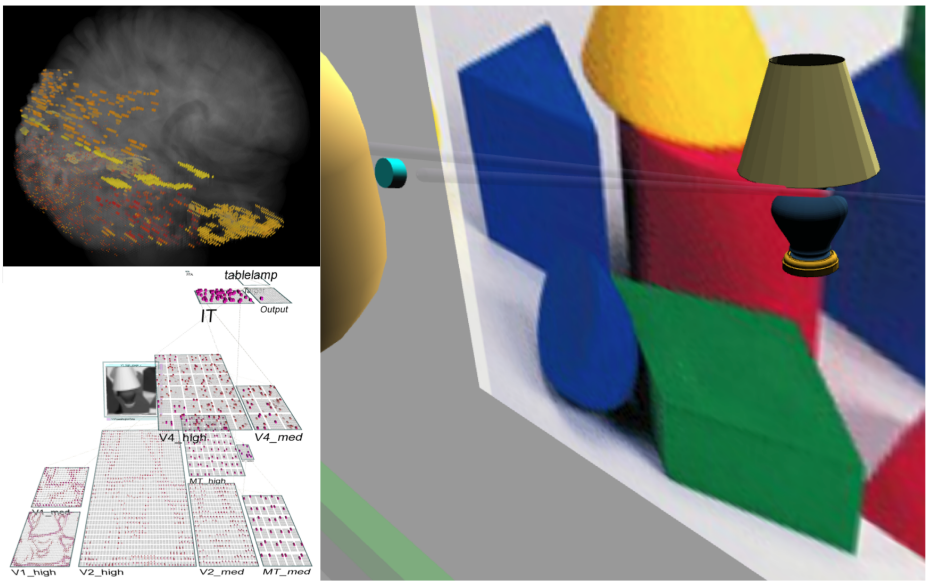
Figure 8. Structure of the hippocampal memory system and associated medial temporal lobe cortical structures

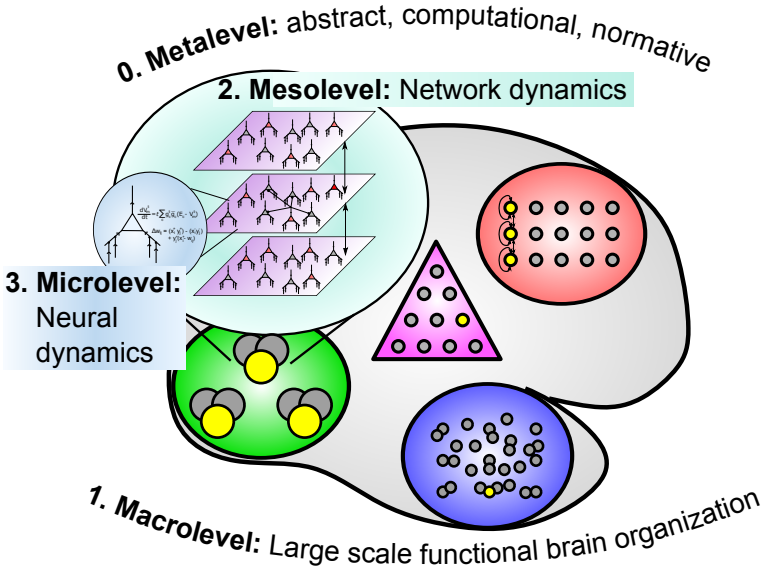
Figure 9. Pattern separation as a result of sparse activity levels in hippocampus relative to cortex.

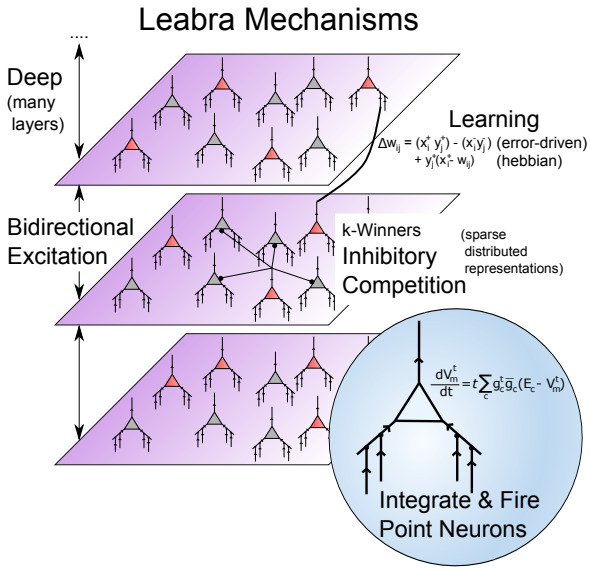
Figure 10. The PBWM (prefrontal cortex basal ganglia working memory) component of the Leabra architecture, capturing the dynamic gating of prefrontal cortex active maintenance by the basal ganglia, which is in turn modulated by phasic dopamine signals to learn what is important to maintain. The PVLV (primary value, learned value) system provides a biologically-based model of the dopaminergic system.

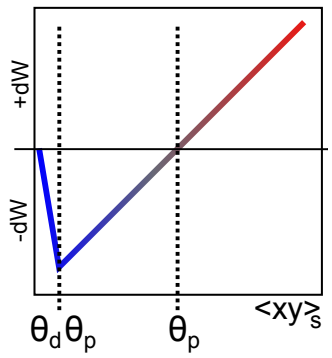
Figure 11. The What vs. How content organization of the brain, showing a map of the lateral surface of the cortex on the left, and half of a coronal slice through the frontal cortex on the right, to label the corresponding medial portions of the frontal cortex (ACC = Anterior Cingulate Cortex, OFC = Orbital Frontal Cortex). Numbers on the lateral surface represent Brodmann areas for the frontal cortex.

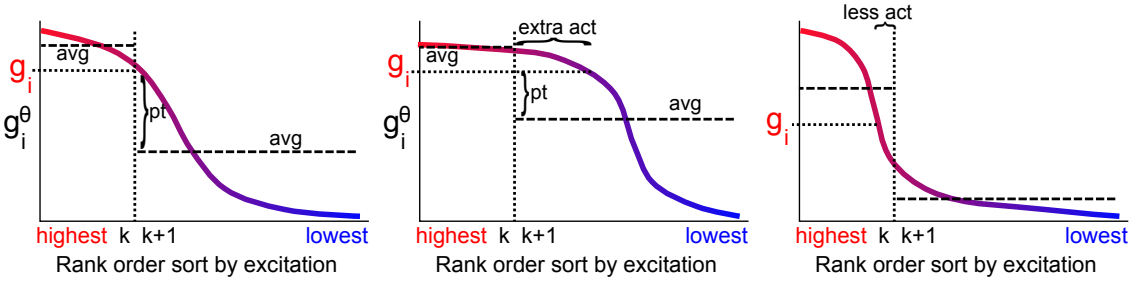

SNrThal_out
SNrThal_mnt
Matrix_out
Matrix_mnt
dAttacks32
dCA3
dDG
dCA1
ACCTaskOpts
OFCTaskOpts
PFCctrlPC
SNRTaskOpts
PFCTaskOpts
BGTaskOpts
PFCprobs_norm
LC_Surprise
BF_Conflict
ACCError
ACCConflict
ACCSurprise
dECout_GpCntr
dECin_GpCntr
GroupRadius
DistanceFly
Distance
BaseRate
Probability
AttackAttn
PFCprobs_out
GroupCenter
GroupAttn
RoadSeg
Value: act

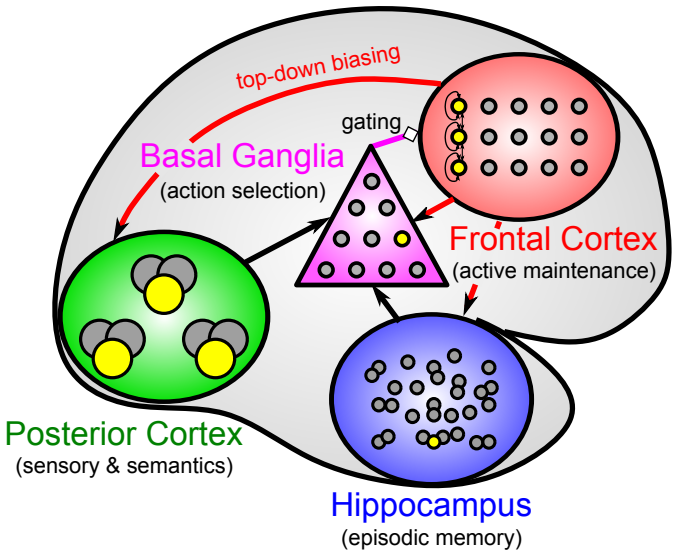

tablelamp
Output
IT
V4_high
V4_med
MT_high
V1_high
V2_high
V2_med
MT_med

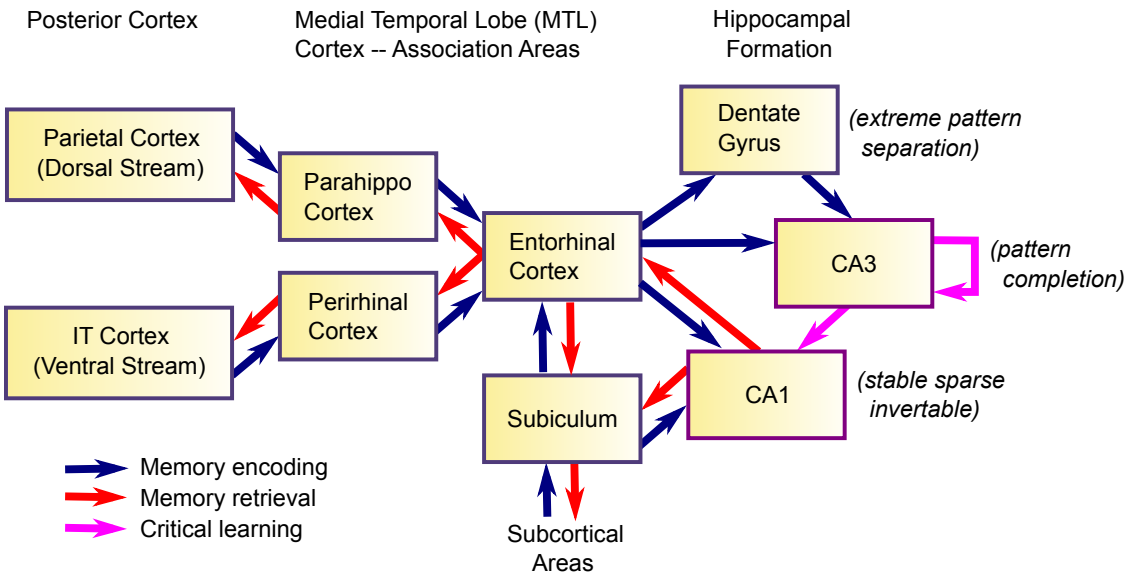

0. Metalevel: abstract, computational, normative
2. Mesolevel: Network dynamics
3. Microlevel: Neural dynamics
1. Macrolevel: Large scale functional brain organization

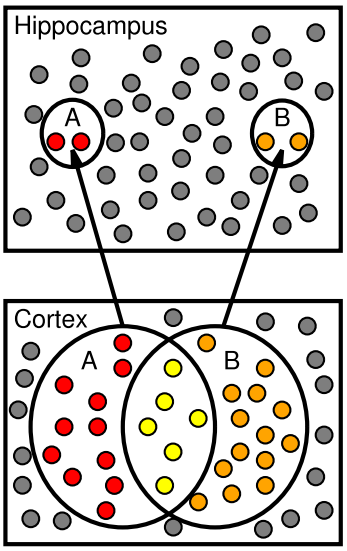

Leabra Mechanisms
Deep
(many
layers)
Learning
Δwij = (xi+ yj+) - (xi- yj-) (error-driven)
+ yj+(xi+ - wij) (hebbian)
Bidirectional
Excitation
k-Winners
Inhibitory
Competition
(sparse
distributed
representations)
dVm^t/dt = τ Σc gc^t ḡc (Ec - Vm^t)
Integrate & Fire
Point Neurons

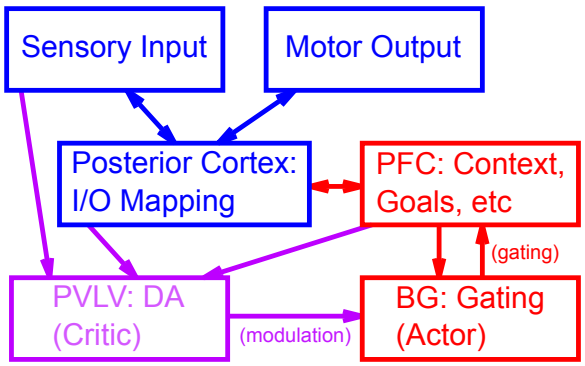

+dW
-dW
θd θp
θp
<xy>s

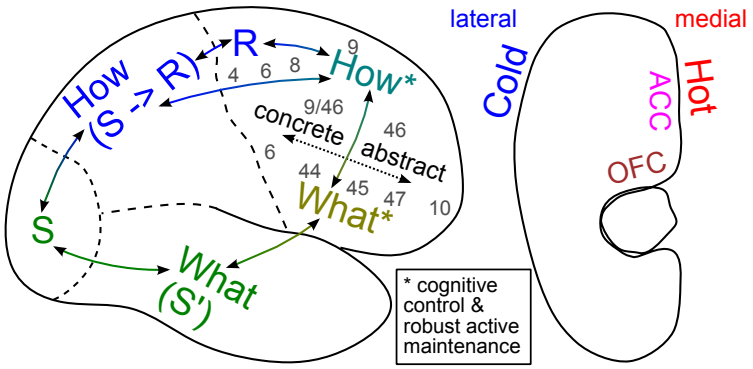

avg
pt
avg
highest k k+1 lowest
Rank order sort by excitation
extra act
avg
pt
avg
highest k k+1 lowest
Rank order sort by excitation
less act
highest k k+1 lowest
Rank order sort by excitation


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


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


Hippocampus
A
B
Cortex
A
B


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


lateral
medial
Cold
Hot
ACC
OFC
R
How
(S -> R)
How*
4
6
8
9
9/46
concrete
46
abstract
6
44
45
47
10
What*
S
What
(S')
* cognitive
control &
robust active
maintenance