Journal of Integrative Neuroscience, Vol. 11, No. 4 (2012) 453–475

DOI: 10.1142/S021963521250029X

Imperial College Press
www.icpress.co.uk

On the neural basis of rule-guided behavior

David C. Noelle

University of California, Merced
5200 North Lake Road
Merced, CA 95343, USA
dnoelle@ucmerced.edu

[Received 26 July 2012; Accepted 14 November 2012; Published 24 December 2012]

Human behavior emerges from a complex dynamic interaction between graded and context-sensitive neural processes, the biomechanics of our bodies, and the vicissitudes of our environments. These coupled processes bear little resemblance to the iterated application of simple symbolic rules. Still, there are circumstances under which our behavior appears to be guided by explicit mental rules. A prototypical case is when succinct verbal instructions are communicated and are promptly followed by another. How does the brain support such rule-guided behavior? How are explicit rules represented in the brain? How are rule representations shaped by experience? What neural processes form the foundation of our ability to systematically represent and apply rules from the vast range of possible rules? This article reviews a line of research that has sought a computational cognitive neuroscience account of rule-guided behavior in terms of the functioning of the prefrontal cortex, the basal ganglia, and related brain systems.

Keywords: Rule-guided behavior; computational cognitive neuroscience; prefrontal cortex; basal ganglia; dopamine; cognitive control.

1. On the Neural Basis of Rule-Guided Behavior

Is human behavior guided by mental rules? If so, how are these rules instantiated in the brain? As the cognitive sciences advance, a growing array of regularities are discovered concerning human behavior, and, to the degree that these regularities appear rule-like, it is natural to hypothesize that they arise from some form of internal representation of such rules, along with a neural rule-following mechanism. On some occasions, behavioral rules seem to become observable, as when verbal instructions are provided and those instructions are promptly followed. On other occasions, the explicit rule-governed strategies that we employ may arise from within, but can be verbally described.

This article provides a brief overview of a specific line of research aimed at providing a computational account of how *rule-guided behavior* is supported by the neural circuits of the human brain. This line of research is ongoing, so the proposed computational account is still incomplete, but much progress has been made in using formal models of neural mechanisms to relate anatomical and physiological data to relevant cognitive and behavioral phenomena. Under this developing theoretical

account, the bulk of human behavior is *not* rule-guided in the current sense. It is asserted that the function of most neural networks is poorly described as implementing mental rules, for a variety of reasons. In terms of representation, many neural systems appear to manipulate information at a level below that of psychological concepts — at a sub-conceptual or sub-symbolic level (Smolensky, 1988). These circuits are often highly context sensitive, in a manner that defies description by compact mental rules (Elman, 1990), and they exhibit a continuity in their dynamics that does not appear in the sequential application of rules (Spivey, 2008). Still, there are some situations in which humans do appear to exhibit behavior driven by explicit rules, and it is the neural mechanisms employed in these situations of true rule-guided behavior that have been the target of the investigations described here.

In order to understand the nature of rule-guided behavior, it is important to clearly distinguish it from *rule-described behavior*, in the following sense. In some situations, a behavior might be *described* by a simple rule, but it does not necessarily follow that the simple rule is instantiated in the brain and is driving the behavior. For a behavior to be rule-guided, the rule in question must be represented or encoded in neural tissue, and that physical representation must have appropriate causal force in the production of the behavior. This distinction is well illustrated by an analogy to a simple physical process. Consider a thrown ball that is allowed to follow its natural parabolic ballistic trajectory. There is a simple mathematical "rule" that accurately describes this trajectory: the equation for a parabola. This quadratic equation does an excellent job of *describing* the "behavior" of the moving ball. This behavior is in no way *guided* by the quadratic "rule", however. There is no physical mechanism that explicitly computes the location of the ball at each moment using this equation. Instead, the motion of the ball is driven by inertia and gravity, with the parabolic description emerging from the interaction of these physical processes. Similarly, a behavior may be well described by a simple rule without that rule actually being instantiated in the brain. This is the distinction between *rule-guided behavior* and *rule-described behavior*, with the former generally being a proper subset of the latter.

This raises the question of how we can recognize when a rule is explicitly represented in the brain. Some have suggested that the only practical criterion is one of the rule being *verbalizeable* — represented by a linguistic code and readily available for verbal report (Ahsby *et al.*, 1998). Others have risked confusing rule-described behavior and actual rule-guided behavior by suggesting a looser criterion that includes implicit patterns of responding such as the "unspoken rules for social interaction" (Bunge & Wallis, 2008). The line of research reviewed in this article has largely remained agnostic concerning the role of language in rule representation, but, like verbalizeable sentences, sees explicit rules as constituents of working memory when being actively applied. Thus, for an explicit rule to guide behavior, it must have a corresponding neural representation that can become "active" in working memory. Grounded in work on the neural basis of working memory (Goldman-Rakic, 1987; Funahashi, Bruce & Golman-Rakic, 1989), these active representations are seen as being encoded in the distributed pattern of sustained neural firing across a delineated set of neurons, most

noteably in portions of the prefrontal cortex (PFC). Under this view, a behavior is rule-guided when it is at least partially caused by a pattern of sustained neural firing in the neural circuits supporting working memory, representing the given rule.

This account asserts that rule-guided behavior is exceptional, with many cognitive processes operating fluently and efficiently without the activation of explicit rule representations in working memory. One might be tempted to contrast this theoretical stance with that proposed by symbolic cognitive architectures, such as Soar (Laird, 2012) and ACT-R (Anderson & Lebiere, 1998), in which almost all cognitive capabilities are seen as arising from the application of symbolic production rules (Klahr, Langley & Neches, 1987). To highlight this contrast would be to equivocate on the "rule" label, however, as neither Soar nor ACT-R view their production rules as the kind of explicitly represented rules that we have associated with rule-guided behavior. A Soar or ACT-R theory of rule-guided behavior would need to explain how generic production rules could be used to interpret more explicit working memory "chunks" that encode the explicit rule to be followed (Taatgen & Lee, 2003). In contrast, the line of research described in this article grounds rule-guided behavior in the neural processes of the prefrontal cortex and related brain areas, rather than in symbolic production rules that are abstracted from the underlying biology. The goal is to produce a formal theory of the neural basis of rule-guided behavior, including the development and learning of rule-following capabilities.

Just as the explicit rules being discussed in this report should not be confused with symbolic production rules, they are distinct from the rules that capture the grammar of natural language — grammatical rules that some have argued to be implemented by a dedicated rule-based neural module. There has been a long running debate concerning the degree to which observed systematic and generative language use in humans entails the existence of a specialized (and, perhaps, innate) neural system for applying symbolic rules of language structure and language use (Pinker, 1999). Extensive computational modeling work has shown that human linguistic performance is better captured by adaptive, context-sensitive, dynamic neural circuits than by separate modules for rules and exceptions (Plaut *et al.*, 1996). While the proposal of a separate rule module seems to have arisen from a kind of confusion between rule-described behavior and rule-guided behavior, it is important to note that the grammatical rules discussed in this literature are distinct from the explicit rules of interest, here. Specifically, fundamental grammatical rules are generally not thought to demand working memory resources to be applied. While working memory may play a role in maintaining and integrating the information sequentially presented in an utterance, it is not thought to actively contain the grammatical rules, themselves.

In summary, the question at hand is how the brain represents and applies fairly arbitrary explicit rules, rather than the kinds of implicit rules hypothesized by some theories of language use or the kinds of production rules that appear in prominent symbolic cognitive architectures. A prototypical case of such rule-guided behavior is when we promptly and appropriately respond to simple verbal instructions (e.g.,

"When the red light comes on, press this button".). In order to address this question, this article reviews a historical sequence of computational cognitive neuroscience models of the prefrontal cortex and related brain areas.

Computational cognitive neuroscience models focus on explaining cognitive and behavioral phenomena in terms of the information processing capabilities of biological neural circuits. All of the computational cognitive neuroscience models discussed in this review have been implemented (or, in the case of earlier work, re-implemented) in the Leabra modeling framework (O'Reilly, 1996). Leabra is a set of mathematical formalisms for modeling neural circuits in a manner appropriate for capturing complex cognitive phenomena. The framework includes both spiking and firing-rate neural models using a point-neuron approximation, bidirectional excitation between brain areas, efficiently computed feedforward and feedback inhibition, a biologically grounded synaptic plasticity model that incorporates both correlational and error-driven learning, and mechanisms for dopamine-modulated reinforcement learning. A detailed understanding of the Leabra framework is not needed to grasp the theoretical contributions of the models being discussed, here, but such an understanding might be had by perusing O'Reilly & Munakata (2000) or O'Reilly *et al.* (2012).

2. Rules as Neural Activity

Early brain-based computational accounts of cognition followed the slogan that "the knowledge is in the weights", indicating that the knowledge needed to enact a cognitive capability was implicit in the distributed collection of synaptic strengths that played a central role in determining the behavior of a modeled neural circuit. With this perspective in mind, it is natural to imagine that any explicit rules that guide behavior might also be encoded in some pattern of synaptic weights. If this is the case, then the synaptic strengths in question must be of a kind that can change by substantial amounts in a very short amount of time, as explicit rules can be adopted and profoundly influence behavior immediately upon their receipt in the form of verbal instructions. This is consistent with the kind of rapid synaptic change thought to arise in some memory systems, such as those hypothesized to exist in the hippocampus, but it does not match the kind of integrative synaptic change hypothesized to take place in much of neocortex (McClelland, McNaughton & O'Reilly, 1995). Furthermore, the idea that explicit rules are those that become represented in working memory at the time of their application suggests that the activated representation of a rule needs to be of a form that changes very rapidly — as quickly as working memory contents can be modified. Both electrophysiological and brain imaging data have suggested that active information in working memory is encoded in neural firing patterns in regions of the prefrontal cortex (PFC), rather than patterns of synaptic strengths (Goldman-Rakic, 1987; D'Esposito, Postle & Rypma, 2000). Neural firing in PFC has been found to encode a wide variety of working memory content, including spatial locations (Funahashi, Bruce & Golman-Rakic, 1989), other stimulus features (Cohen *et al.*, 1994), and even abstract behavioral rules (Wallis, Anderson & Miller, 2001).

Based on this evidence, explicit rules are seen as being encoded as patterns of neural firing in PFC when they are actively available for application.

The neural circuits of PFC are well configured to support the active maintenance of a pattern of firing that might encode an explicit rule. This region of cortex contains many small patches or "stripes" of tissue within which dense recurrent excitatory connections are found (Levitt *et al.*, 1993; Pucak *et al.*, 1996). These dense patterns of recurrent connectivity suggest that neural activity in these PFC stripes should exhibit attractor dynamics (Port & van Gelder, 1995), supporting sustained neural firing (i.e., active maintenance of information in working memory) even after driving input activity has been removed. Such attractor dynamics in densely recurrent neural networks has been extensively modeled (Hopfield, 1982; Amit, 1989).

There are extensive projections both to and from PFC with most more posterior cortical areas (i.e., areas posterior to PFC) (Roberts, Robbins & Weiskrantz, 1998). These broad connections offer a pathway through which an explicit rule, represented as an actively maintained pattern of neural firing in PFC, could influence a broad range of neural circuits — not only prompting motor responses but also guiding attention, influencing memory encoding and retrieval, and biasing any other neural processes receiving input from PFC. This observation suggests an alternative to the idea that rule-guided behavior is produced by separate cortical pathways than those that produce more automatic forms of behavior. The broad projections from PFC to more posterior brain areas suggests that more automatic processing may arise from the spontaneous dynamics of these posterior circuits, and more *controlled* and rule-guided processing may stem from the modulation of these posterior circuits by neural activity ariving from PFC (Miller & Cohen, 2001). In this way, an explicit rule is actively represented by a pattern of neural firing in PFC, and the rule produces appropriate behavior by biasing neural circuits throughout cortex based on this PFC activity.

An early demonstration of the power of this approach to account for both neuroscientific and behavioral data appeared in the form of a computational model of *cognitive control* in the Stroop task. In the classic Stroop task, participants are presented with words written in colored typefaces, and they are instructed to either name the printed word or name the color of the characters. While error rates tend to be low on this task, responding is slowed when an ink color is requested and the presented word form is, itself, the name of a color (Stroop, 1935). Performance on this task was modeled by a simulated neural circuit that maps from input stimulus features to the output names of those features, with synaptic strengths configured so that word forms would be produced at the output in preference to ink colors. In the model, this circuit received additional input from a model of PFC function, which contained simulations of frontal pyramidal cells whose firing rates encoded either the "name the word form" rule or the "name the ink color" rule. By activating the PFC activation pattern for "name the ink color", the model could overcome its prepotent tendency to name words, but not without a slowdown in the case of conflict stimuli. (See Fig. 1(a) for a schematic of the network architecture that was used.) This model accounted for healthy Stoop performance data, as well as the performance of patient

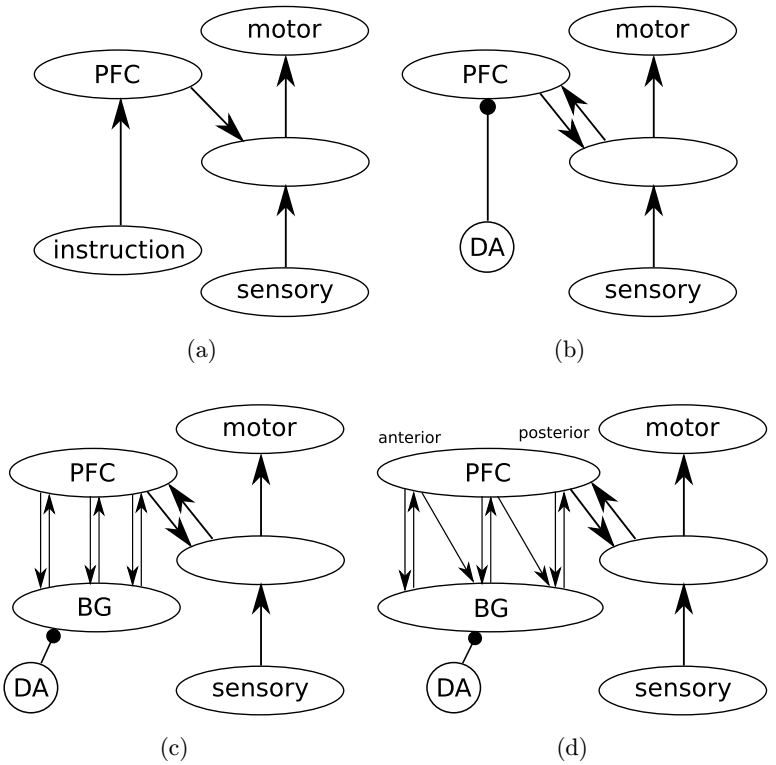

Fig. 1. Schematic diagrams of model architectures: Ellipses represent collections of modeled neurons, and arrows represent projections between these collections. Tasks typically involve producing an appropriate *motor* response for each *sensory* input, with this mapping mediated by association cortices (unlabeled ellipse). Explicit rules are encoded in patterns of neural firing in the prefrontal cortex (*PFC*), which modulates more posterior circuits. (a) The network configuration for the Stoop model and the instructed category learning model, with rules communicated to PFC by direct inputs. (b) The network configuration for the AX-CPT model, the IDED model, and the XT model, with gating of PFC contents controlled by a midbrain dopamine (*DA*) signal. Not shown are projections from all other modeled brain regions to the midbrain, allowing the gating signal to be shaped in a context-senstivity manner by a reinforcement learning process. (c) The network configuration for the PBWM model, with PFC gating controlled by the basal ganglia (*BG*) through topographically organized thalamocortical loops. The DA-based reinforcement learning process shapes BG performance in this architecture. (d) The network configuration for the indirection model, with both the PFC and the BG topographically arrayed along an anterior to posterior axis. More anterior PFC neurons may influence the gating of more posterior PFC neurons through projections to more posterior areas of the BG.

populations, including those with frontal damage and those with schizophrenia (Cohen, Dunbar & McClelland, 1990; Cohen & Servan-Schreiber, 1992).

It is worth noting that Stroop performance deficits in schizophrenia were captured by this model by weakening the sustained firing pattern in PFC encoding the instructed task rule. This manipulation was justified by the reduction in tonic dopamine (DA) levels in PFC commonly observed in patients with schizophrenia, and an account of DA modulation of frontal pyramidal cells in which DA influences the mathematical *gain* of each neuron's mapping from its inputs to its firing rate. Under this account, reduced tonic DA results in PFC neurons that are less sensitive to the

excitation that they are receiving from their neighbors, reducing the stability of the attractor states that encode the instructed task rule, allowing the rule representation to "fade" (Cohen & Servan-Schreiber, 1992). This connection between DA and the active maintenance of an explicit rule in PFC will be revisited in the next section of this review.

Similar computational models have been applied to other tasks involving the following of explicit instructions (Noelle & Cottrell, 1995). These neural network models learned, through a process of experience-based modification of synaptic strengths, to translate sequences of input tokens in a simple instructional language into an appropriate pattern of activity across simulated neurons in a PFC-like working memory circuit. This actively maintained rule representation (an *internal representation*, learned from experience) then modulated a network representing more posterior brain areas, mapping sensory inputs to motor outputs. Models of this kind were applied to an instructed category learning task which had been empirically investigated. In such an instructed category learning task, the learner is asked to categorize simple geometric stimuli into one of two categories, with the categorization rule provided through direct verbal instruction (Noelle & Cottrell, 1996b). One interesting aspect of these models involves interactions between rule-following performance and synaptic change, driven by experience categorizing stimuli. Extensive practice on the categorization task, using only a limited subset of the possible geometric stimuli, can cause learners to deviate from perfect rule-following behavior, producing a pattern of performance that reflects an interaction between explicit rule-following and incrementally learned associations between stimulus features and category labels (Noelle & Cottrell, 2000). The proposed neural basis of rule-guided behavior captures this interaction by incorporating more than a "mere implementation" of a simple symbolic rule-interpretation system, with the activation-based representation of the rule interacting in complex ways with the processes of synaptic plasticity, adjusting synapic strengths, that are driven by extensive practice on the task.

In summary, the proposed computational account of rule-guided behavior views explicit rules, when ready to be applied, as patterns of neural activity in PFC, actively maintained by dense recurrent excitation in this part of cortex. Broad projections from PFC to more posterior brain areas allow this rule representation to modify the activation dynamics of the sensory-motor circuits that produce overt behavior, producing performance aligned with the actively maintained explicit rule.

3. Adaptive Rule Selection

The attractor network account of the active maintenance of explict rules works well when the rule to be applied in the current situation is clearly specified through sensory input, as it is in the case of verbal instruction. Models of instruction following, in which the explicit rule that is to guide behavior is communicated through the senses (e.g., verbally), could simply update the pattern of firing rates in the simulated PFC when instructions were presented as input and actively maintain the

corresponding rule representation thereafter. Everyday occurrences of rule-guided behavior are not so straightforward. In general, the world does not provide clear and reliable signals concerning which explicit rule should be applied in each situation. Instead, some brain system must determine when the currently active rule is to be maintained and when it should be abandoned, perhaps in favor of another explicit rule. Even under direct verbal instruction, we have the option to obey the communicated rule or ignore it.

The dynamics of attractor networks are not sufficient to account for the need to intelligently toggle between states of *active maintenance* of the current rule and *rapid updating* of PFC. In broad strokes, if recurrent connections are stronger than input connections, an attractor network will persistently maintain its current state, ignoring its inputs. If the input connections are stronger, such a network will rapidly update its state with every new input that is presented, regardless of its semantic content. Thus, in order to adaptively toggle between these two states, some additional mechanism is needed. In the line of research under consideration, inspired by recurrent artificial neural networks making use of multiplicative synapses (Hochreiter & Schmidhuber, 1997), the need to intelligently control the state of PFC has been met by an *adaptive gating mechanism*. The metaphor employed, here, involves seeing PFC as a fenced enclosure with a gate. When the gate is closed, the contents of the enclosure cannot escape, and the corresponding PFC contents are actively maintained. When the gate is opened, the contents leave the enclosure, and a new pattern of neural firing may be instantiated by the inputs to PFC.

Importantly, just as explicit rules can be learned through direct instruction or through experience, the situations in which such rules are useful can also be learned. Thus, no simple fixed gating mechanism will address the problem of rule selection. Instead, any such mechanism must be *adaptive*, learning from experience when to maintain the current rule and when to abandon it.

This focus on learning led researchers to further consider interactions between PFC and the midbrain DA system. A growing body of research has related phasic DA bursts from the substantia nigra pars compacta (SNC) and the from the ventral tegmental area (VTA) with changes in expected future reward (Schultz, Dayan & Montague, 1997). Such a reward prediction signal is central to temporal difference accounts of reinforcement learning (Sutton, 1988), leading to neural models of learning from reward, grounded in the DA system (Montague, Dayan & Sejnowski, 1996). These computational neural models provide accounts of how the brain can learn to produce rewarding sequences of overt actions, matching both neuroscientific and behavioral data. Given this mechanism for the learning of overt actions, Braver & Cohen (2000) suggested that the same DA signal could be used to learn the covert choice of opening or closing the metaphorical gate on PFC working memory contents. If the active maintenance of an explicit rule in a given situation leads to reward, its maintenance can be encouraged in similar future situations. If the active maintenance of an explicit rule fails to produce expected reward, the rule can be abandoned in similar future situations. In this way, an adaptive gating mechanism for PFC could

be grounded in the same DA-based reinforcement learning mechanisms previously explored in the context of motor sequence learning.

This hypothesis concerning adaptive gating of PFC contents is also consistent with known widespread DA projections from the VTA to PFC, as well as observed properties of DA receptors in PFC pyramidal cells (Wang, Vijayraghavan & Goldman-Rakic, 2004), suggesting that a phasic DA signal from the VTA could contribute to active maintenance through the triggering of specific maintenance currents in these cells (Camperi & Wang, 1998). This has led to computational models in which a neural reinforcement learning process learns to predict future reward based on both the sensory state of the system and the maintained contents of PFC, with changes in expected future reward encoded in a VTA DA signal, and adaptive gating of PFC contents directly driven by this DA signal (Braver & Cohen, 2000).

An early model of this type captured human performance on the AX Continuous Performance Task (AX-CPT). In this task, participants are presented with a sequential stream of stimuli, such as individual letters. Each stimulus in the sequence is to be quickly labeled as target or non-target (e.g., by a button press), with targets being a specific stimulus ("X") but only when it is immediately preceded by another specific stimulus ("A"). Thus, a target response needs to be given for an "X", but only when it immediately follows an "A". By manipulating the frequencies of specific stimulus pairs, in sequence, different patterns of errors can be invoked in human participants. These patterns are well captured by a computational model in which explicit rules like "produce the target response if the next stimulus is an X" (i.e., an "A" stimulus was just presented) may be gated into PFC, with gating determined by a DA-based reinforcement learning process (Braver & Cohen, 2000). (See Fig. 1(b) for a schematic characterization of this network model.)

Perhaps more relevant to the study of rule-guided behavior is a computational model of this kind that was applied to the Intra-Dimensional/Extra-Dimensional (IDED) categorization task. In this task, learners are presented with a pair of stimuli and required to choose one of the two. If the correct stimulus is chosen, a reward is given. Each stimulus is composed of two features: one feature from each of two stimulus *dimensions*. For example, each stimulus might involve the superposition of a foreground black line shape on a background filled blue shape. For each stimulus pair, one specific stimulus feature predicts reward. For example, the presence of a blue filled triangle in the background may identify that stimulus as the rewarding stimulus. Learners must discover the rewarding stimulus feature by trial and error. Critically, once the learner's behavior indicates that the rewarding feature has been discovered, the rewarding feature is changed without notifying the learner. The rapidity with which the learner adapts to this change is used as a measure of *flexibility* of explicit rule selection, making this task much like the more common Wisconsin Card Sorting Test (WCST) (Berg, 1948). The IDED task gets its name from the different kinds of changes that are explored (Dias, Robbins & Roberts, 1997). For example, an "intra-dimensional reversal" involves a case in which a specific feature in one stimulus dimension (e.g., a blue background triangle) is initially rewarding, and

another feature in that same dimension is initially *not* rewarding (e.g., a blue background circle), and the change involves *reversing* these contingencies, so the rewarding feature becomes non-rewarding, and vice versa (e.g., the blue background circle is now rewarding, and the blue background triangle, when presented, is not). This can be contrasted with an "extra-dimensional shift", in which a novel stimulus feature in the opposite stimulus dimension becomes rewarding (e.g., while a blue background triangle was initially rewarding, the presence of a novel foreground black line cross becomes rewarding, with the blue triangle no longer appearing as a background in any stimulus).

A computational model of DA-based PFC updating applied to the IDED task was able to capture the performance of marmosets, including both healthy animals and those with experimental lesions to various regions of PFC (O'Reilly *et al.*, 2002). In this model, the explicit rules involved perceptual attention, indicating which feature(s) of the currently perceived stimulus pair should be attended, with attention implemented as an increase in the neural activation of the perceptual cells encoding the feature(s) being attended, driven by top-down projections from PFC (Desimone & Duncan, 1995). The attentional highlighting of rewarding stimulus features, in this way, supported the association of those features with the selection of stimuli that contained them, with that association emerging in learned synaptic strengths. Importantly, different parts of marmoset PFC were assumed to encode the attentional rules at different levels of abstraction, with more dorsal areas encoding more general rules (e.g., "attend to the background blue shapes") and more ventral areas encoding more specific rules (e.g., "attend to the background blue triangle"). This gradient of abstractness in the neural rule representaions allowed the model to capture differences in flexible rule switching between animals with dorsal PFC lesions versus animals with ventral PFC lesions. Of most relevance, here, is that fact that the DA-based PFC updating mechanism worked well at capturing flexible rule switching in this task.

This IDED model also offers a reminder that the account of rule-guided behavior being developed, here, does not involve a neural implementation of a general rule-interpretation mechanism that operates independently of other neural processes. In the IDED model, associations between stimulus features and stimulus selection actions were captured in changing synaptic strengths, with synaptic plasticity profoundly affected by top-down biasing of neural activity from PFC. In this way, the explicit attentional rule interacted with non-rule-guided neural processes in order to produce rewarded performance. Similarly, when the ventral PFC of the IDED model was lesioned, the model was still able to learn the initial rewarding stimulus feature, as were the lesioned animals, but profound deficits were observed upon rule switching. In the model, this is captured by an initial learning process that involves no detailed explicit attentional rule, with all learning occurring as synaptic changes. These synaptics strengths are then much slower to reverse than when PFC is intact, allowing for the rapid updating of PFC to contain a new rule calling attention to a different stimulus feature. Thus, this model requires both its dynamically gated PFC

and its synaptic plasticity mechanisms for adjusting synaptic strengths to capture the full range of experimental observations. Indeed, since the rules maintained in PFC in this model are attentional in nature, rather than specifying specific responses, synaptic strength modifications are needed to allow even the non-lesioned model to learn the task, at all. In short, while rules are explicitly represented as neural firing patterns in PFC, under this account, the application of those rules involves complex interactions with the activation and synaptic dynamics of non-rule-guided neural processes.

In summary, neural reinforcement learning mechanisms based on the midbrain DA system can provide a means for learning, from experience, when to actively maintain the current explict rule and when to abandon it for another.

4. Learning Rule Representations

Our conceptual vocabulary is shaped by our experience. To the degree that the explict rules that guide our behavior leverage our full conceptual vocabulary, the representational scheme used to encode those rules in PFC must also be shaped by our experience. A frontal neural representation of a very simple rule-like instruction, like "press the space bar", must interact with learned neural circuits in more posterior brain areas for, say, visually recognizing a "space bar" and implementing a motor "press" action. While all of the models discussed in this review, so far, involved PFC representations that were hand-designed by the model builder, we know that these representations must actually arise in the human brain through a process of learning and development.

If the range of explicit rules that could guide our behavior was small, allowing us many opportunities to practice each possible rule over the course of development, standard formal accounts of synaptic plasticity would easily explain the learning of rule representations. More-or-less arbitrary patterns of neural firing in PFC, used to encode a given rule, could come to provide appropriate top-down biasing of posterior circuits through standard synaptic learning mechanisms. This is not the case, however.

Humans are remarkable in their ability to rapidly apply novel rules in a highly systematic and generative manner. Consider only simple verbal instructions of the form, "When you see X, do Y". A new rule of this kind exists for every possible *pair* of visually identifiable objects and performable actions. The space of possible rules is truly *huge*, arising largely due to the ability to *combine* fairly independent *components* of a rule in a *combinatoric* number of ways. While the ability to internally represent and successfully apply virtually any rule from this huge space is a hallmark human capability, this poses a serious problem for standard neural learning mechanisms, because these statistical learning methods generally fail to exhibit the kind of combinatoric generalization inherent in human instruction-following abilities (Hadley, 1999).

When rule representations are learned through a general statistical learning process, the problem of combinatoric generalization is one of spurious correlations,

including spurious anti-correlations. Most mechanisms for neural learning are exquisitely sensitive to the statistical structure of the experiences presented to the system. Thus, if a perfect anti-correlation incidentally appears in my experience (e.g., while I have been asked to push buttons, and I have seen dinosaurs, I have *never* been asked to press a button when I see a dinosaur), a neural learning mechanism will implicitly assume that the anti-correlation is real — that the anti-correlated components can *never* appear together. Similarly, if an accidental perfect correlation occurs in my experience (e.g., all of the buttons I have ever pushed were square), standard neural learning mechanisms will make it impossible to represent items that violate that accidental correlation (e.g., I won't be able to represent pushing a circular button). Because of this problem, learning PFC representations that support combinatoric generalization across the full space of explicit rules is a substantial challenge (Noelle & Cottrell, 1996a; Noelle & Zimdars, 1999).

The Cross-Task Generalization Model (XT Model) was produced in order to investigate the possibility that the DA-based adaptive gating mechanism might interact in a favorable way with standard methods of synaptic plasticity so as to produce more *componential* rule representations that supported some measure of combinatoric generalization (Rougier *et al.*, 2005). This neural network model was presented with a simple neural encoding of multi-dimensional stimulus objects, much like the cards used in the WCST (Berg, 1948). Each stimulus item varied along five different stimulus dimensions (e.g., size, color, shape,...), with each dimension having four discrete levels (e.g., tiny, small, large, huge). Also input into the model was a pattern of activity that encoded the task to be performed (e.g., name the color of the stimulus, indicate if two stimuli are of the same shape, specify which of two stimuli is larger,...). The network learned to perform these tasks through a process of experience-based synaptic plasticity (the standard synaptic strength change model used in the Leabra framework), with the response properties of neurons in the simulated PFC also shaped by experience. Importantly, as the model learned to perform these tasks, some features were never used with some tasks. For example, the model might be asked to name the color of a blue stimulus, but it might never be asked if two blue stimuli are of the same color. The ability of the network to perform its tasks on stimuli that were novel to the task at hand was used as a measure of generalization performance. The XT Model was directly compared with alternative PFC models, including one that possessed all of the properties of the XT Model except for the DA-based gating mechanism. Simulation studies found that good generalization to novel situations could be had only when both (1) the DA-based gating mechanism was in place, and (2) the model experienced training across many different tasks, rather than only across a randomly sampled pair of tasks. Importantly, good generalization was strongly correlated with the learning of *isolated* and *dimensional* representations across the simulated PFC neurons, with each PFC cell encoding a full stimulus dimension (e.g., color or size) to be attended. In other words, when the network model exhibited good generalization, few or no PFC cells contributed to the representation of multiple dimensions (e.g., firing when shape is relevant *and* when color is

relevant), and few or no cells encoded more arbitrary collections of stimulus features (e.g., attend to red squares). In this way, the XT Model learned the relevant independent components of the various tasks — namely, the stimulus dimensions — and encoded explicit rules to attend to the different dimensions across disjoint subsets of the PFC neurons.

Because of the nature of the tasks learned by the XT Model, it was able to perform both the Stroop task and the Wisconsin Card Sorting Test (WCST) without substantial modification. The learned isolated and dimension PFC representations allowed the XT Model to provide a good fit to human performance on Stroop and WCST, both in the case of healthy individuals and in the case of frontally damaged patients. Removing simulated PFC neurons from the model resulted in performance that matched that of people with frontal lesions (Rougier *et al.*, 2005). Also, in later work, selective damage to the DA-based gating mechanism in the XT Model was found to produce a pattern of Stroop and WCST performance that matched that seen in people with autism spectrum disorders (Kriete & Noelle, 2005).

Work with the XT Model demonstrated that PFC rule representations could be learned from experience over a developmental time scale, with the resulting representations having properties that support combinatoric generalization to novel situations. The full challenge presented by Hadley (1999), involving combinatoric generalization over the space of verbalizeable rules, was not met by this model, however. The model developed its own internal representations for rules like "attend to the stimulus color", and it learned to apply those rules in novel situations, but it did not learn to represent and apply a rule that it had never previously practiced. Addressing the challenge of combinatoric generalization to novel rules will require some additional neural mechanisms.

5. Gating Rule Components

All of the computational models discussed to this point have actively maintained a single explicit rule in PFC at any one time. The DA-based adaptive gating mechanism, thus, produced a global gating signal that called for either active maintenance or rapid updating for all of the contents of PFC, as a whole. This is problematic if we view humans as being capable of holding in mind multiple explicit rules at the same time, with each rule independently maintained or abandoned, or even if we suggest that components of rules might be independently gated into PFC. While working memory capacity is seen as extremely limited, experimental evidence suggests that more than a single "chunk" can be actively maintained (Cowan, 2001). This suggests a need for independent adaptive gating mechanisms for different "stripes" of PFC tissue (Levitt *et al.*, 1993; Pucak *et al.*, 1996).

It is not immediately clear how reinforcement learning using the standard method of temporal differences could support the adaptive gating of multiple PFC stripes. Certainly, a unitary global DA-signal would be insufficient to provide such refined control. Interestingly, this computational concern is mirrored by issues raised by

neuroscientific data. There is reason to doubt that direct phasic DA delivery to PFC happens quickly enough to support the needs of PFC gating. Indeed, recent physiological data suggests that the bistability of PFC cells, toggling between active maintenance and rapid updating, is more readily controlled by glutamatergic projections from the thalamus, through their effects on NMDA and metabotropic glutamate receptors.

This has led to a much more anatomically detailed computational model of thalamocortical loops, passing through the basal ganglia (BG) (Frank, 2005). The basic idea is that medium spiny neurons in the matrix areas (matrisomes) of the caudate, in the striatum, determine when a given stripe of PFC tissue is allowed to rapidly update. Some of these matrix neurons, called "Go" cells, disinhibit the thalamus (particularly the medial dorsal and ventral anterior nuclei) through the substantia nigra pars reticulata. Other matrix neurons, called "No Go" cells, strengthen the tonic inhibition of the thamalus from the substantia nigra pars reticulata by disinhibiting the substantia nigra cells through the external segment of the globus pallidus. In short, the firing of "Go" cells tends to allow the thalamus to excite PFC, while the firing of "No Go" cells inhibits this thalamic signal to cortex. Thus, when "Go" activity dominates over "No Go" activity, the corresponding PFC stripe undergoes rapid updating. (There is also a more global inhibitory signal through the subthalamic nucleus, but this does not play a central role in the models that are discussed here (Frank, 2006).) In this detailed model of the BG, the reinforcement learning mechanisms of the DA system modulate synaptic plasticity in the striatum, where the matrix cells receive projections both from PFC and from more posterior brain areas. In this way, the DA system allows adaptive gating to be learned, as in previous models, but the gating signal sent to PFC is now mediated through the BG and the thalamus. Also, the matrix cells associated with different PFC stripes may acquire different synaptic strengths, allowing different regions of PFC to rapidly update under different conditions.

The resulting model, called the PFC-BG Working Memory (PBWM) Model, can learn to selectively update multiple rule components, stored in different PFC stripes (O'Reilly & Frank, 2006). While early versions of this model reverted to modeler-designed representations across simulated PFC neurons, later versions allowed these representations to be learned, as they were in the XT Model (Hazy, Frank & O'Reilly, 2006). (See Fig. 1(c) for a schematic characterization of this network architecture.)

It was expected that the inherently componential structure of a collection of independently gated PFC stripes would further support combinatoric generalization in these models. An obstacle to such generalization remained, however. As in previous models, the explicit rules actively maintained in PFC influenced behavior by biasing the activity of neurons in more posterior brain areas. If these posterior neural circuits were shaped by experience-based synaptic plasticity, then they could suffer from the same sort of spurious correlation problems that were previously discussed in the context of learning PFC representations. Even if the rule representation in PFC perfectly segregated the representation of independent rule components between

separate pools of PFC cells, the posterior neurons receiving this full pattern of neural firing from PFC could come to depend on accidental correlations that had appeared during development. For example, even if the novel rule "when you see a dinosaur, press the space bar" was cleanly encoded across multiple PFC stripes (perhaps a stripe encoding the "seen" thing and a stripe encoding the action to take), more posterior neural circuits may have difficulty generalizing to this rule, especially if it is highly dissimilar to any explicit rule encountered in the past.

The proposed solution to this problem has been the introduction of an *output gating* mechanism. The idea is to use DA-based reinforcement learning to not only learn when a given PFC stripe, containing a rule component, should be rapidly updated with new contents, but to also use this learning method to learn when the sustained pattern of neural activity stored in a given PFC stripe should be released to more posterior brain areas. Introducing such additional control on the top-down biasing of posterior areas by PFC allows the posterior circuits to be sequentially exposed to individual rule components, reducing the generalization demands placed on posterior circuits. For example, if the novel rule "when you see a dinosaur, press the space bar" is broken up into sequential components, "when you see a dinosaur, do something" and "when it's time to do something, press the space bar", there is little opportunity to learn spurious correlations. Once the posterior circuits had learned to recognize a dinosaur, given a rule to do so, and had learned to press the space bar on demand, the novel combination, presented sequentially, would follow naturally. This strategy can only be learned, however, if there is a mechanism for exposing posterior circuits to individual rule components at any one time. An output gating mechanism provides this.

One hypothesized implementation of such an output gating mechanism involves segregating those PFC pyramidal cells that undergo sustained activation when representing a rule from those PFC cells that project to more posterior brain areas, with local connections in place from the PFC cells exhibiting sustained firing to those sending outputs. This segregation of PFC neurons might involve cells at different cortical layers, or it might involve different PFC stripes. With such a PFC architecture in place, different striatal cells would control "input gating" and "output gating" in much the same way. Striatal cells that controlled thalamic input to sustained activity cells in PFC would provide an "input gate", determining when rapid updating allowed these PFC cells to encode a new active rule component. Striatal cells that controlled thalamic input to PFC output cells would provide an "output gate", rapidly updating the output cells to match their corresponding sustained activation cells, thereby making the content of the corresponding PFC stripe visible to posterior areas (Hazy *et al.*, 2006). Such an output gating mechanism has been shown to greatly improve generalization performance (Kriete & Noelle, 2011).

The PBWM Model, with learned PFC representations within stripes and output gating, demonstrates that a DA-based reinforcement learning mechanism can learn to independently update individual rules or rule components in PFC, and it does so using neural circuits more carefully aligned with known anatomy than previous

models. This model also comes closer to achieving combinatoric generalization over the space of rules by directly supporting a compositional encoding of rules across multiple PFC stripes.

6. Indirection in Rule Representations

The PBWM Model provides an account of rule-guided behavior that allows for the active application of multiple rules and/or the representation and application of rules with multiple components. The componential nature of isolated PFC stripes, in the model, greatly supports generalization to novel rules. Human generalization abilities are still much greater, however, indicating a profound shortcoming of the model.

The isolated PFC stripes in the PBWM Model produce a kind of "role-filler" representation scheme. In broad strokes, the collection of PFC neurons in a given stripe come to be associated with a particular role, and the pattern of sustained firing exhibited by those neurons encodes the current filler for that role. For example, the PFC may learn, from experience, to encode rules of the kind, "When you see X, do Y". This might be done by associating one collection of PFC neurons with the "X" role (i.e., the thing that is to be seen) and associating another collection of PFC cells with the "Y" role (i.e., the action to take). Appropriate neural activation representations of any such rule will be formed in the PBWM Model, even for completely novel rules, as long as the system experienced every possible "X" filler in some rule encountered during development and experienced every possible "Y" filler in some rule, as well. While the PBWM Model will generalize to novel pairs of "X" and "Y" fillers, it cannot generalize to the case in which a filler had not been experienced in the context of this rule template. For example, consider the novel rule, "When you see a dinosaur, do a foot tap". If you know how to tap your foot, perhaps having done so in response to an auditory cue, but you have never followed a rule involving performing a foot tap in response to a visual stimulus, then PBWM would fail to account for your ability to apply this rule. If "foot tap" had never been encoded over the "Y" pool of PFC neurons during development in PBWM, then these cells will likely fail to properly represent this action as part of this novel rule. Thus, PBWM can generalize to novel combinations of fillers, but it cannot generalize to the case of assigning a filler to a role for which that filler had never previously been assigned.

Current work, involving extending the PBWM Model to account for this weakness in combinatoric generalization, has focused on the computer science concept of indirection — of "pointers" (Kriete *et al.*, 2012). The idea is that the neural firing pattern appearing in a PFC stripe need not encode a rule component directly, but may, instead, encode a reference to a different PFC stripe where the rule component is currently being actively maintained. For example, the "X" neurons and the "Y" neurons in a representation of "When you see X, do Y", do not encode things you see and actions, respectively, but both the "X" cells and the "Y" cells encode pointers to other PFC stripes, where the constituent components of the rule are being actively maintained. Over developmental learning, the "X" neurons learn to encode for each

of the various PFC stripes where a visual object representation had been stored, and the "Y" neurons learn to encode for each of the various PFC stripes where an action representation had been stored. In this case, if I have experience encoding rules involving the "foot tap" action in PFC, perhaps in response to an auditory stimulus, there will be one or more PFC stripes that have learned to encode the "foot tap" action. When presented with the novel rule, "When you see a dinosaur, do a foot tap", the "Y" neurons simply need to fire in a manner that references one of the action PFC stripes that has learned to represent "foot tap". In this way, the indirection version of PBWM can generalize to novel combinations of rule components, even when one (or more) of those components had never been experienced in its newly specified role.

In this augmented model, a reference to another PFC stripe is implemented by a projection from the "role" PFC stripe to the output gating striatal "Go" units for the referenced PFC stripe. With this pattern of connectivity in place, the network can learn, from experience, to release the contents of the referenced PFC stripe to posterior brain areas whenever the referencing "role" stripe receives its own output gating signal. Thus, following the previous example, when the "Y" neurons receive an output gating signal, the pattern of neural firing over the "Y" units is released to the output gating striatal cells for the PFC stripe containing the "foot tap" action representation. This results in the "foot tap" representation being sent to more posterior brain areas. Just as in the original PBWM model, sending an output gating signal to a "role" PFC stripe results in an interatal representation of the associated "filler" to be sent to posterior brain circuits. In the indirection model, this process simply involves an intermediate step. The indirection model provides support for full combinatoric generalization, allowing for the internal representation of novel combinations of rule components, even when a given component had never been experienced in a particular role before. The theoretical claims implicit in the indirection model have yet to be empirically tested. One possibility is an anterior-to-posterior gradient across PFC, with more anterior PFC stripes encoding for specific more posterior PFC stripes. (See Fig. 1(d) for a schematic diagram of this modified PBWM network architecture.)

7. Future Directions

Work with the indirection model is in its early stages, but it offers promise as a means to produce the kind of combinatoric generalization seen in human rule-followers. This generalization problem is but one of many issues that have yet to be addressed by this line of research.

For example, in the models discussed in this review, rule selection has involved choosing between maintaining a current rule (or rule component) and abandoning it. These models lack any intelligence in how the space of possible rules is searched, seeking out a rule that results in rewarding behavior. When the adaptive gating mechanism indicates that it is time for rapid updating of PFC contents, the newly

generated PFC activation pattern is generally sampled randomly, though some models guide this sampling process through attention to sensory inputs. (For example, in the IDED model, if one of the current stimuli contains a background blue circle feature, attention to that feature becomes one of the candidate rules that may be gated into PFC. Gating in that rule is much more likely than gating in a rule that makes no contact with the stimulus features that are present in the current stimulus pair.) Future models will need to incorporate more sophisticated neural mechanisms for searching the space of explicit rules in a systematic (or, at least, not overly redundant) manner.

All of the models reviewed in this report have focused on explicit rules that are actively maintained in a PFC-based working memory, ready to be applied. Some of these models focused on scenarios in which the rules were seen as arriving in working memory through a recent language understanding process, operating on direct verbal instructions. More often, explicit rules are retrieved from memory, prompted by situational cues. Fully understanding this latter phenomenon will require computational models that integrate prefrontal function with that of the hippocampus and cortical declarative memory systems, allowing these longer-term memory circuits to provide appropriate inputs to PFC. Good computational models of hippocampal function exist, but detailed integration with PFC has been elusive (Norman & O'Reilly, 2003).

The rule selection process might be guided by more than a prediction of expected reward. In particular, there is some evidence to suggest that the behavioral control provided by the PFC is modulated by perceived task difficulty. One such difficulty measure is that of *cognitive conflict*, with brain imaging and event related potential (ERP) data suggesting a special role for the anterior cingulate cortex in *conflict monitoring* (Botvinick *et al.*, 2001). Future computational models of rule-guided behavior should further explore the possibility that situations exhibiting persistently high conflict, as monitored by the ACC, may contribute to the abandonment of a rule.

Finally, it is worth noting that this particular line of research has consistently been built upon a foundation of neural network modeling methods that focus on rate-coding in cortical pyramidal cells and learning through the experience-based modification of synaptic strengths (Parks, Levine & Long, 1998). This history need not constrain future models of rule-guided behavior. Future research efforts should explore the role that spike-timing might play in the encoding of explicit rules, perhaps by encoding working memory contents as reverberating polychronous neuronal groups (Szatmáy & Izhikevich, 2010). While the models that have been discussed, here, have incorporated molecular level effects in their mechanisms for synaptic plasticity (O'Reilly & Munakata, 2000) and the influence of modulatory neurotransmitters (Cohen & Servan-Schreiber, 1992), there is much room for further investigation into the role that molecular biological mechanisms, including physical processes such as broad electrical fields, might play in rule-guided behavior (Aur, Jog & Poznanski, 2011). Alternative modeling frameworks that eschew a focus on synaptic strengths in favor of a detailed account of the effects of individual action

potentials (Poznanski, 2002a), as well as those that focus on dynamic and integrative approaches (Poznanski, 2002b; Cacha & Poznanski, 2011), could provide novel insights into the neural basis of explicit rules.

8. Summary and Conclusion

The article has briefly reviewed a specific line of research aimed at producing a detailed computational account of the neural basis of rule-guided behavior. Together, the computational models that have been developed as part of this effort offer a short list of hypotheses concerning the use of explicit rules:

- Explicit rules, when ready to be applied, are encoded as actively maintained neural firing patterns over the pyramidal cells of the prefrontal cortex. In all of the models discussed here, these encodings are seen as distributed patterns of firing rates over PFC neurons.
- In order to intelligently determine when a given explicit rule should be applied and when it should be abandoned, the contents of prefrontal cortex working memory circuits must be controlled by an adaptive gating mechanism. Such an adaptive gating mechanism can learn to identify when given rules are useful, using a neural reinforcement learning process grounded in the midbrain dopamine system.
- The rule representations in prefrontal cortex must be learned from experience. A dopamine-based gating mechanism interacts with standard models of synaptic plasticity to support the development of appropriately isolated and dimensional prefrontal representations, giving rise to improved generalization to novel situations when adequately diverse training experiences are provided.
- An anatomically detailed account of interactions between the prefrontal cortex and the basal ganglia can explain how individual rules or rule components can be independently manipulated in prefrontal cortex, supporting further componential generalization to novel rules and novel situations, particularly if an output gating mechanism is assumed to be present.
- Some regions of prefrontal cortex, perhaps those closer to the frontal pole, may encode references or "pointers" to other prefrontal areas. Such a representational scheme, utilizing indirection, allows for essentially full combinatoric generalization to novel rules.

These conjectures form the foundation of an ongoing effort to understand, in formal computational terms, how the human brain supports flexible rule-guided behavior.

It is important to note that, while these computational models have progressively improved in their ability to demonstrate combinatoric generalization, they do not embody "mere implementations" of symbolic rule-interpretation mechanisms. Complex interactions between the rule representations actively maintained in prefrontal cortex and the dynamic processes of more posterior neural circuits give rise to graded and context-sensitive patterns of performance that escape description by a purely symbolic rule account. Also, statistical regularities in the experiences present

during the development of prefrontal cortex can profoundly shape the kinds the explicit rules that can robustly be represented and applied. In this way, this line of research offers an account of how the "language of thought", at least with regard to explicit rules, may be shaped by experience, rather than being innate.

Much work remains. In addition to addressing the mechanisms underlying the longer term retention of explicit rules, as well as their retrieval, efforts to scale up previous simulations are necessary to demonstrate the viability of the proposed approach to rule-guided behavior. Existing models are far from addressing the challenge of Hadley (1999), allowing complex acquired rule-based skills to be mixed and stacked, guided by direct verbal instruction. Still, much progress toward this level of systematic generalization has been made, offering insights into how neural circuits can give rise to the profoundly generative performance sometimes observed in humans.

Author Note

The author thanks Gary Cottrell, Randy O'Reilly, Jonathan Cohen, Todd Braver, Alex Petrov, Nicolas Rougier, Mike Frank, and Trent Kriete for fruitful discussions on the topics discussed in this article over many years of collaborations and interactions.

REFERENCES

Ahsby, F.G., Alfonso-Reese, L.A., Turken, A.U. & Waldron, E.W. (1998) A neuropsychological theory of multiple systems in category learning. *Psychol. Rev.*, **105**(3), 442−481.

Amit, D.J. (1989) *Modeling Brain Function: The World of Attractor Neural Networks.* Cambridge University Press.

Anderson, J.R. & Lebiere, C. (1998) *The atomic components of thought.* Mahway, NJ: Erlbaum.

Aur, D., Jog, M. & Poznanski, R.R. (2011) Computing by physical interaction in neurons. *J. Integr. Neurosci.*, **10**(4), 413−422.

Berg, E.A. (1948) A simple objective test for measuring flexibility in thinking. *J. Gen. Psychol.*, **39**, 15−22.

Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S. & Cohen, J.D. (2001) Conflict monitoring and cognitive control. *Psychol. Rev.*, **108**(3), 624−652.

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 and J. Driver, eds. *Control of Cognitive Processes: Attention and Performance XVIII.* Cambridge, Massachusetts: MIT Press, pp. 713−737.

Bunge, S.A. & Wallis, J.D. (2008) Introduction. In: S.A. Bunge and J.D. Wallis, eds. *Neuroscience of Rule-Guided Behavior.* Oxford: Oxford University Press, pp. xiii−xxiii.

Cacha, L.A. & Poznanski, R.R. (2011) Associable representations as field of influence for dynamic cognitive processes. *J. Integr. Neurosci.*, **10**(4), 423−437.

Camperi, M. & Wang, X.-J. (1998) A model of visuospatial working memory in prefrontal cortex: Recurrent network and cellular bistability. *J. Comput. Neurosci.*, **5**, 383−405.

Cohen, J.D., Dunbar, K. & McClelland, J.L. (1990) On the control of automatic processes: A parallel distributed processing model of the stroop effect. *Psychol. Rev.*, **97**(3), 332−361.

Cohen, J.D., Forman, S.D., Braver, T.S., Casey, B.J., Servan-Schreiber, D. & Noll, D.C. (1994) Activation of prefrontal cortex in a nonspatial working memory task with functional MRI. *Hum. Brain Mapp.*, **1**, 293−304.

Cohen, J.D. & Servan-Schreiber, D. (1992) Context, cortex, and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychol. Rev.*, **99**(1), 45−77.

Cottrell, G.W. (ed.) (1996) *Proceedings of the 18th Annual Conference of the Cognitive Science Society.* La Jolla: Lawrence Erlbaum.

Cowan, N. (2001) The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behav. Brain Sci.*, **24**(1), 87−185.

Desimone, R. & Duncan, J. (1995) Neural mechanisms of selective visual attention. *Ann. Rev. Neurosci.*, **18**, 192−222.

D'Esposito, M., Postle, B.R. & Rypma, B. (2000) Prefrontal cortical contributions to working memory: Evidence from event-related fMRI studies. *Exp. Brain Res.*, **133**, 3−11.

Dias, R., Robbins, T.W. & Roberts, A.C. (1997) Dissociable forms of inhibitory control within prefrontal cortex with an analog of the wisconsin card sort test: Restriction to novel situations and independence from "on-line" processing. *J. Neurosci.*, **17**, 9285−9297.

Elman, J.L. (1990) Finding structure in time. *Cogn. Sci.*, **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. *J. Cogn. Neurosci.*, **17**(1), 51−72.

Frank, M.J. (2006) Hold your horses: A dynamic computational role for the subthalamic nucleus in decision making. *Neural Netw.*, **19**(8), 1120−1136.

Funahashi, S., Bruce, C.J. & Golman-Rakic, P.S. (1989) Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J. Neurophysiol*, **61**, 331−349.

Goldman-Rakic, P.S. (1987) Circuitry of the prefrontal cortex and the regulation of behavior by representational knowledge. In: F. Plum and V. Mountcastle, eds. *Handbook of Physiology.* Bethesda, MD: American Physiological Society, pp. 373−417.

Hadley, R.F. (1999) Connectionism and novel combinations of skills: Implications for cognitive architecture. *Minds and Machines*, **9**(2), 197−221.

Hazy, T.E., Frank, M.J. & O'Reilly, R.C. (2006) Banishing the homunculus: Making working memory work. *Neurosci.*, **139**(1), 105−118.

Hochreiter, S. & Schmidhuber, J. (1997) Long short-term memory. *Neural Comput.*, **9**, 1735−1780.

Hopfield, J.J. (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci.*, **79**, 2554−2558.

Klahr, D., Langley, P.W. & Neches, R.T. (eds). (1987) *Production System Models of Learning and Development.* Cambridge, Massachusetts: MIT Press.

Kriete, T. & Noelle, D.C. (2005) Impaired cognitive flexibility and intact cognitive control in autism: A computational cognitive neuroscience approach. In: *Proceedings of the 27th Annual Conference of the Cognitive Science Society.* Stresa, Italy: Lawrence Erlbaum, pp. 1190−1195.

Kriete, T. & Noelle, D.C. (2011) Generalisation benefits of output gating in a model of prefrontal cortex. *Connection Science*, **23**(2), 119−129.

Kriete, T., Noelle, D.C., Cohen, J.D. & O'Reilly, R.C. (2012) *Indirection and Generativity in the Prefrontal Cortex/basal Ganglia System.* (in preparation).

Laird, J.E. (2012) *The Soar Cognitive Architecture.* Cambridge, MA: MIT Press.

Levitt, J.B., Lewis, D.A., Yoshioka, T. & Lund, J.S. (1993) Topography of pyramidal neuron intrinsic connections in macaque monkey prefrontal cortex (areas 9 46). *J. Comp. Neurol.*, **338**, 360−376.

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

Miller, E.K. & Cohen, J.D. (2001) An integrative theory of prefrontal cortex function. *Ann. Rev. Neurosci.*, **24**(1), 167−202.

Montague, P.R., Dayan, P. & Sejnowski, T.J. (1996) A framework for mesencephalic dopamine systems based on predictive hebbian learning. *J. Neurosci.*, **16**, 1936−1947.

Noelle, D.C. & Cottrell, G.W. (1995) A connectionist model of instruction following. In J.D. Moore and J.F. Lehman, eds. *Proceedings of the 17th Annual Conference of the Cognitive Science Society.* Pittsburgh: Lawrence Erlbaum, pp. 369−374.

Noelle, D.C. & Cottrell, G.W. (1996a) In search of articulated attractors. In: G.W. Cottrell, ed. *Proceedings of the 18th Annual Conference of the Cognitive Science Society.* La Jolla: Lawrence Erlbaum, pp. 329−334.

Noelle, D.C. & Cottrell, G.W. (1996b) Modeling interference effects in instructed category learning. In: G.W. Cottrell, ed. *Proceedings of the 18th Annual Conference of the Cognitive Science Society.* La Jolla: Lawrence Erlbaum, pp. 475−480.

Noelle, D.C. & Cottrell, G.W. (2000) Individual differences in exemplar-based interference during instructed category learning. In: L.R. Gleitman and A.K. Joshi, eds. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society.* Philadelphia: Lawrence Erlbaum, pp. 358−363,

Noelle, D.C. & Zimdars, A.L. (1999) Methods for learning articulated attractors over internal representations. In: M. Hahn and S.C. Stones, eds. *Proceedings of the 21st Annual Conference of the Cognitive Science Society.* Vancouver: Lawrence Erlbaum, pp. 480−485.

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

O'Reilly, R.C. (1996) *The LEABRA Model of Neural Interactions and Learning in the Neocortex.* Unpublished doctoral dissertation, Carnegie Mellon University, Department of Psychology.

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

O'Reilly, R.C. & Munakata, Y. (2000) *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain.* Cambridge, Massachusetts: MIT Press.

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

O'Reilly, R.C., Noelle, D.C., Braver, T.S. & Cohen, J.D. (2002) Prefrontal cortex and dynamic categorization tasks: Representational organization and neuromodulatory control. *Cereb. Cortex*, **12**, 246−257.

Parks, R.W., Levine, D.S. & Long, D.L. (eds.) (1998) *Fundamentals of Neural Network Modeling: Neuropsychology and Cognitive Neuroscience*. Cambridge, Massachusetts: MIT Press.

Pinker, S. (1999) *Words and Rules: The Ingredients of Language*. New York: Basic Books.

Plaut, D.C., McClelland, J.L., Seidenberg, M.S. & Patterson, K. (1996) Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychol. Rev.*, **103**(1), 56−115.

Port, R.F. & van Gelder, T. (eds.) (1995) *Mind as Motion: Explorations in the Dynamics of Cognition*. Cambridge, Massachusetts: MIT Press.

Poznanski, R.R. (2002a) Dendritic integration in a recurrent network. *J. Integr. Neurosci.*, **1**(1), 69−99.

Poznanski, R.R. (2002b) Towards an integrative theory of cognition. *J. Integr. Neurosci.*, **1**(2), 145−156.

Pucak, M.L., Levitt, J.B., Lund, J.S. & Lewis, D.A. (1996) Patterns of intrinsic and associational circuitry in monkey prefrontal cortex. *J. Comparative Neurol.*, **376**, 614−630.

Roberts, A.C., Robbins, T.W. & Weiskrantz, L. (eds.) (1998) *The Prefrontal Cortex: Executive and Cognitive Functions*. Oxford: Oxford University Press.

Rougier, N.P., Noelle, D.C., Braver, T.S., Cohen, J.D. & O'Reilly, R.C. (2005) Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proc. Natl. Acad. Sci.*, **102**(20), 7738−7343.

Schultz, W., Dayan, P. & Montague, P.R. (1997) A neural substrate of prediction and reward. *Science*, **275**, 1593−1599.

Smolensky, P. (1988) On the proper treatment of connectionism. *Behav. Brain Sci.*, **11**(1), 1−74.

Spivey, M. (2008) *The Continuity of Mind*. Oxford: Oxford University Press.

Stroop, J.R. (1935) Studies of interference in serial verbal reactions. *J. Exp. Psychol.*, **28**, 643−662.

Sutton, R.S. (1988) Learning to predict by the methods of temporal differences. *Machine Learning*, **3**, 9−44.

Szatmáy, B. & Izhikevich, E.M. (2010) Spike-timing theory of working memory. *PLoS Comput. Biol.*, **6**(8), e1000879. (doi: 10.1371/journal.pcbi.1000879).

Taatgen, N.A. & Lee, F.J. (2003) Production compilation: A simple mechanism to model complex skill acquisition. *Human Factors*, **45**(1), 61−76.

Wallis, J.D., Anderson, K.C. & Miller, E.K. (2001) Single neurons in prefrontal cortex encode abstract rules. *Nature*, **411**, 953−956.

Wang, M., Vijayraghavan, S. & Goldman-Rakic, P.S. (2004) Selective D2 receptor actions on the functional circuitry of working memory. *Science*, **303**, 853−856.