A Neurocomputational Theory of how Rule-Guided Behaviors Become Automatic

Paul Kovacs[a], Sébastien Hélie[b],
Andrew N. Tran[a], & F. Gregory Ashby[a]
[a]University of California, Santa Barbara
[b]Purdue University

Abstract

This article introduces a biologically-detailed computational model of how rule-guided behaviors become automatic. The model assumes that initially, rule-guided behaviors are controlled by a distributed neural network centered in the prefrontal cortex, and that in addition to initiating behavior, this network also trains a faster and more direct network that includes projections from sensory association cortex directly to rule-sensitive neurons in premotor cortex. After much practice, the direct network is sufficient to control the behavior, without prefrontal involvement. The model is implemented as a biologically-detailed neural network constructed from spiking neurons and displaying a biologically plausible form of Hebbian learning. The model successfully accounts for single-unit recordings and human behavioral data that are problematic for other models of automaticity.

Keywords: rule-guided behavior; automaticity; prefrontal cortex; computational neuroscience

Introduction

After long periods of practice, almost any task can be executed quickly, accurately, and with little or no conscious deliberation. At this point, we say that the behavior has become automatic. A strong case can be made that most behaviors performed by adults are automatic. When we sit in a chair, pick up a cup of coffee, or swerve to avoid a pothole, our actions are almost always automatic.

As motivation for his well-known cognitive theory of automaticity, Logan (1988) noted that children initially learn to add single-digit numbers by counting – that is, by applying a time-consuming and effortful rule – but after long periods of practice they can produce the correct sum seemingly by rote. How does the transition occur from systematically

This research was supported by grant MH2R01-063760 from the National Institute of Mental Health.

applying an effortful rule to responding automatically? Neurobiologically-detailed theories that account for the transition from initial learning to automaticity exist for motor skills (e.g., Ashby, Ennis, & Spiering, 2007), but no such theories exist for rule-guided behaviors. This article aims at filling this gap in the literature. Specifically, we propose a neurobiologically-detailed theory of how automaticity develops for rule-guided behaviors. The theory is formalized as a computational model constructed from spiking neurons, and we show that this model successfully accounts for a variety of single-unit recording and behavioral phenomena that characterize automatic rule-guided behavior.

By rule, we mean a set of explicit instructions that produces the correct behavior and can be applied to a variety of different stimuli or scenarios (e.g., counting to add two numbers). Note that not all behaviors are rule guided. Cigar rollers do not automatize their intricate finger movements by repeatedly recalling an elaborate set of instructions (Crossman, 1959). Instead, the acquisition of motor skills relies on extended practice with feedback and procedural learning and memory. Many previous studies of automaticity have focused on behaviors that depend heavily on procedural learning for initial acquisition. This includes skilled typing (e.g., Logan, 1982; Long, 1976; Rabbitt, 1978; Sternberg, Monsell, Knoll, & Wright, 1978) and the serial reaction time task (e.g., Cohen & Poldrack, 2008; Poldrack et al., 2005). In contrast, far fewer studies have examined the development of automaticity for rule-guided behaviors. This difference is important because rule-guided and procedural-learning mediated behaviors depend on different neural networks, require different criteria to assess automaticity (Ashby & Crossley, 2012), and as we will see shortly, express at least some qualitatively different properties after automaticity has developed (Roeder & Ashby, 2016). For these reasons, different neuroscience-based theories are required to account for how automaticity develops in rule-guided and procedural-learning mediated behaviors.

Because automatic behaviors that were acquired via rule learning versus procedural learning exhibit at least some qualitative differences (Roeder & Ashby, 2016), it is important to test a theory of rule-guided automaticity against data from tasks in which acquisition depends on rule learning. As a result, much of the empirical literature on automaticity is inappropriate for testing the model proposed here. Even so, all automatic behaviors share features in common (e.g., speed and effortlessness), so we believe that our new model could account for many of the automaticity-related phenomena documented via the study of behaviors that were acquired, for example, via procedural learning. However, little is known about exactly which phenomena are shared across automatic rule-guided and procedural behaviors, and which phenomena are unique. Therefore, an initial test of any new model of rule-guided automaticity should be restricted to tests against data from rule-guided tasks.

What is a good experimental paradigm for studying rule-guided behaviors? If a rule is a set of explicit instructions that can be applied to a variety of different stimuli or scenarios, then note that this set of stimuli or scenarios could be used to define a category. In other words, a rule is a set of instructions that can be applied to any member of some category. Therefore, although rule-guided behavior could be studied in many different domains, one particularly attractive choice is perceptual categorization. There is now abundant evidence that humans learn perceptual categories in qualitatively different ways, including via rule and procedural learning (e.g., Ashby & Maddox, 2005, 2010; Love, Medin, & Gureckis, 2004; Reber, Gitelman, Parrish, & Mesulam, 2003). Although this is also true in other paradigms, one advantage of perceptual categorization is that reliable methods exist to identify the type of strategy that individual participants are using (Ashby & Valentin, 2018). These methods contrast performance in rule-based (RB) and information-integration (II) categorization tasks. In RB tasks, the optimal strategy is some simple logical rule (Ashby, Alfonso-Reese, Turken, & Waldron, 1998). For example, in the most common applications, only one stimulus dimension is relevant, but tasks in which the optimal strategy is a conjunction rule are also RB. In II tasks, no explicit rule succeeds and accuracy is maximized only if information from two or more incommensurable stimulus components is integrated at some predecisional stage (Ashby & Gott, 1988; Ashby et al., 1998). Considerable evidence suggests that success in RB tasks depends on rule learning, whereas success in II tasks depends on procedural learning (for reviews, see, e.g., Ashby & Maddox, 2005; Ashby & Valentin, 2017).

The neural basis of learning and automaticity is better understood for II than for RB tasks – perhaps because the kind of stimulus-response association (i.e., procedural) learning thought to dominate in II tasks is more amenable to study in non-human animals than the rule learning that dominates in RB tasks. In particular, the evidence is good that early II learning depends critically on the basal ganglia, and especially on the striatum (e.g., Ashby & Ennis, 2006; Seger & Miller, 2010). The idea is that plasticity at cortical-striatal synapses follows reinforcement learning rules with dopamine serving as the reward signal (Doya, 2007). When positive feedback is received, dopamine rises above baseline and active synapses are strengthened, whereas negative feedback causes dopamine to fall below baseline levels, which causes active synapses to weaken.

Ashby et al. (2007) proposed that in contrast, automatic II categorization is mediated entirely within cortex and that the development of II automaticity is associated with a gradual transfer of control from the striatum to cortical-cortical projections from the relevant sensory areas directly to the premotor areas that initiate the behavior. According to this account, a critical function of the basal ganglia is to train purely cortical representations of automatic behaviors (Hélie, Ell, & Ashby, 2015). Specifically, the basal ganglia learn to activate the correct post-synaptic target in premotor cortex via dopamine-mediated reinforcement learning (Cantwell, Crossley, & Ashby, 2015), which allows the appropriate cortical-cortical synapses to be strengthened via Hebbian learning[1]. Once the cortical-cortical synapses have been built, the basal ganglia are no longer required to produce the automatic behavior.

This theory accounts for many results that are problematic for other theories of automaticity. For example, it correctly predicts that people with Parkinson's disease, who have dopamine reductions and striatal dysfunction, are impaired in initial procedural learning (Soliveri, Brown, Jahanshahi, Caraceni, & Marsden, 1997; Thomas-Ollivier et al., 1999), but relatively normal in producing automatic skills (Asmus, Huber, Gasser, & Schöls, 2008). Also, it correctly predicts that blocking all striatal output to cortical motor and premotor areas does not disrupt the ability of monkeys to fluidly produce an overlearned motor sequence (Desmurget & Turner, 2010). Similarly, a neuroimaging study reported that ac-

[1] According to this account, cortical-cortical synaptic plasticity follows Hebbian learning rules because low levels of dopamine active transporter (DAT) in cortex prevent the rapid fluctuations in cortical dopamine levels needed for DA to serve as a reward signal during reinforcement learning. In contrast, the basal ganglia are rich in DAT, so dopamine levels fluctuate rapidly. As a result, dopamine serves as a trial-by-trial reward signal and synaptic plasticity in the basal ganglia follows reinforcement-learning rules.

tivation in the putamen was correlated with II performance early in training but not after automaticity developed (Waldschmidt & Ashby, 2011). Instead, automatic performance was only correlated with activity in cortical areas (i.e., preSMA and SMA).

While evidence continues to accumulate in support of this theory of how procedurally acquired skills become automatized (Ashby, Turner, & Horvitz, 2010; Hélie et al., 2015), there is still no comparable neural account of how rule-guided behaviors become automatized. This article proposes such a theory. First, we describe the neural structures and mechanisms that mediate the transition from recently learned to fully automatized rule-guided behaviors. Next, to test this theory more rigorously, we formulate it as a biologically-detailed neurocomputational network of spiking neurons. Finally, we show that the resulting model successfully accounts for single-unit recording and behavioral data that are problematic for other accounts of automaticity.

The Neural Basis of Automatic Rule-Guided Behavior

Initial Learning

To begin, there is overwhelming evidence that initial rule learning depends on working memory, executive attention, and the prefrontal cortex (PFC). Much of this evidence comes from the Wisconsin Card Sorting Test (WCST; Heaton, 1981), which is a well-known neuropsychological assessment used to detect frontal dysfunction, and especially, damage to the PFC (e.g., Kimberg, D'Esposito, & Farah, 1997). Stimuli in this task are cards containing geometric patterns that vary in color, shape, and the number of symbols that are depicted. The patient's task is to use trial-by-trial feedback to learn to assign each card to its correct category. In all cases, the correct categorization strategy is a simple one-dimensional rule. Many studies have reported that PFC lesions impair animals on a simplified version of the WCST (e.g., Joel, Weiner, & Feldon, 1997). Similarly, a number of neuroimaging studies have used the WCST or an alternative RB task, and all of these have reported task-related activation in the PFC (e.g. Konishi et al., 1999; Monchi, Petrides, Petre, Worsley, & Dagher, 2001; Rogers, Andrews, Grasby, Brooks, & Robbins, 2000).

The most extensively tested neurobiologically-detailed model of category learning, called COVIS, assumes that humans have separate rule-learning and procedural-learning systems (Ashby et al., 1998; Ashby & Valentin, 2017; Ashby & Waldron, 1999). The neural architecture of the COVIS rule-learning system is shown in Figure 1. COVIS assumes that performance improvements in RB tasks are mediated by this rule-learning system, which uses working memory and executive attention to discover the optimal rule and is mediated primarily by the anterior cingulate, the PFC, the hippocampus, and the head of the caudate nucleus. There are two main subnetworks in this model: one that generates or selects new candidate rules, and one that maintains candidate rules in working memory during the testing process and mediates the switch from one rule to another. The COVIS rule-learning system is similar to the neural network models of the WCST that were proposed by Monchi et al. (2001) and Amos (2000).

One of the key assumptions of the COVIS rule-learning model is that rule-sensitive units in PFC remain activated throughout testing of candidate rules. In the Figure 1 model, this persistent activation is facilitated by reverberating loops through the medial dorsal nucleus of the thalamus (Ashby, Ell, Valentin, & Casale, 2005). A number of studies

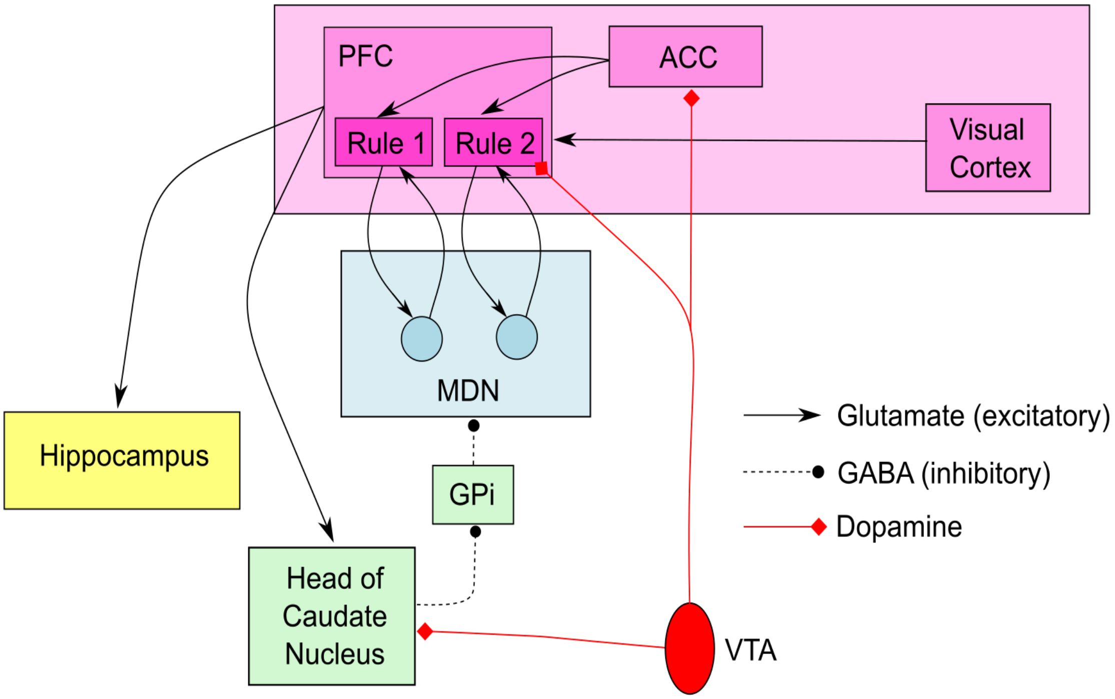

Figure 1. The COVIS rule-learning system. PFC = prefrontal cortex, ACC = anterior cingulate cortex, MDN = medial dorsal nucleus of the thalamus, GPi = internal segment of the globus pallidus, VTA = ventral tegmental area.

have reported evidence for such rule-sensitive neurons in PFC. In these studies, monkeys were trained to classify objects by applying either one rule (e.g., spatial) or another (e.g., associative) while single-unit recordings were collected from PFC neurons. Each trial began with a cue signaling the animal which rule to apply to the ensuing stimulus. Several studies using this paradigm reported many neurons in PFC that showed rule-specific activity – that is, they fired during application of one of the rules but not during the other, regardless of which stimulus was shown (Asaad, Rainer, & Miller, 2000; Hoshi, Shima, & Tanji, 2000; White & Wise, 1999).

Automaticity

Although there are many qualitative differences between initial RB and II learning (e.g., Ashby & Maddox, 2005; Ashby & Valentin, 2017), after automaticity develops, many of these differences disappear. For example, several studies have reported that switching the location of the response keys early in training interferes with II categorization performance but not with RB performance (Ashby, Ell, & Waldron, 2003; Maddox, Bohil, & Ing, 2004). However, Hélie, Waldschmidt, and Ashby (2010) reported that after more than 10,000 trials of practice, switching the location of the response keys produced interference in both tasks (on both accuracy and response time), and that there was almost no recovery from this interference over the course of 600 trials. Similarly, although a dual task that requires working memory interferes with initial RB learning much more than initial II learning (Waldron & Ashby, 2001; Zeithamova & Maddox, 2006), after extensive training this difference also disappears (Hélie, Waldschmidt, & Ashby, 2010). In particular, once RB and II categorization become automatic, there is no dual-task interference in either task.

Neuroimaging results also show convergence (Soto, Waldschmidt, Helie, & Ashby, 2013). During early learning, activation patterns for RB and II tasks are qualitatively different (Hélie, Roeder, & Ashby, 2010; Nomura et al., 2007; Waldschmidt & Ashby, 2011). For example, studies that scanned participants on four different days during 20 sessions of RB or II training reported that early RB performance was correlated with activation in PFC, the hippocampus, and the head of the caudate nucleus (Hélie, Roeder, & Ashby, 2010), whereas early II training depended heavily on the putamen (Waldschmidt & Ashby, 2011). By session 20 however, activation in all of these areas no longer correlated with performance. Instead, only cortical activation (e.g., in premotor cortex) was positively correlated with response accuracy in both tasks.

These results raise the question of whether the same model can account for RB and II automaticity. Despite their similarities, there is good evidence for at least some qualitative differences. For example, Roeder and Ashby (2016) reported evidence that stimulus-response (SR) mappings are automatized after extensive II training, whereas rules are automatized in RB tasks. Participants in this study completed more than 12,000 trials of RB or II categorization distributed across 21 different training sessions. Each participant practiced predominantly on a primary category structure, but every third session they switched to a secondary structure that used the same stimuli and responses. Importantly, half of the stimuli retained their same SR association when the secondary structures were practiced and half switched associations. Thus, if SR mappings are automatized, then the development of automaticity should be slowed on the stimuli that changed responses relative to stimuli that always maintained the same SR association. In contrast, if a rule is automatized there should be no difference between consistent and inconsistent stimuli since the same rule is applied an equal number of times to both types of stimuli. In fact, in the RB condition, there was no difference in accuracy or response time for consistent stimuli that maintained their category label in every session and inconsistent stimuli that switched labels in secondary category-structure sessions. In contrast, for the primary II categories, accuracy was higher and RT was lower for consistent than for inconsistent stimuli. These results strongly suggest that rules are automatized in RB tasks, whereas SR associations are automatized in II tasks.

Together, all these results suggest similar, but not identical, neural representations of automatic II and RB behaviors. As mentioned previously, Ashby et al. (2007) proposed that automatic II categorization is mediated entirely within cortex and that the development of II automaticity is associated with a gradual transfer of control from the striatum to cortical-cortical projections from the relevant sensory areas directly to units in areas of premotor cortex that initiate the behavior. According to this account, a critical function of the basal ganglia is to train purely cortical representations of automatic behaviors (Hélie et al., 2015). We propose a similar model for the development of automatic rule-guided behaviors. In particular, we propose that a key function of the rule-learning network illustrated in Figure 1 is to train automatic cortical-cortical projections from the relevant sensory areas to premotor areas of cortex. The primary difference from the automatization of procedural skills is that we propose that the premotor targets are rule-sensitive units, rather than units associated with a specific motor goal.

A variety of evidence supports this account of how rule-guided behaviors become automatic. Of course, a critical requirement of the theory is that rule-sensitive neurons exist in premotor cortex. Several studies have reported recording from such neurons (Muhammad, Wallis, & Miller, 2006; Wallis & Miller, 2003; Vallentin, Bongard, & Nieder, 2012). In addition, there is evidence that during extended rule-based training, behavioral control gradually passes from the PFC to premotor cortex. First, the neuroimaging data collected by Hélie, Roeder, and Ashby (2010) over the course of 20 sessions of RB categorization were consistent with this hypothesis. Second, Wallis and Miller (2003) recorded from single neurons in the PFC and premotor cortex while monkeys were making rule-based categorization responses (see also Muhammad et al., 2006). In agreement with the Figure 1 model, they found many neurons in the PFC that fired selectively to a particular rule. However, after training the animals for a year, they also found many premotor neurons that were rule selective, and even more importantly, these neurons responded on average about 100 ms before the PFC rule-selective cells. Thus, after categorization had become automatic, the PFC, although still active, was not mediating response selection. Instead, the single-unit data suggested that the automatic representation had moved to regions that included the premotor cortex. Third, within the PFC, several studies have reported that the more concrete the rule, the more caudal the representation (Badre, Kayser, & D'Esposito, 2010; Bunge & Zelazo, 2006; Christoff, Keramatian, Gordon, Smith, & Mädler, 2009). Based on evidence such as this, Hélie, Roeder, and Ashby (2010) proposed that as rules become more concrete with more extensive training, they are progressively re-coded more caudally in the PFC until eventually reaching the premotor cortex, at which time they become automatic.

Thus, according to this view, the primary goal of rule-learning circuits centered in PFC and procedural-learning circuits centered in the basal ganglia is to train automatic representations between sensory cortex and premotor cortex. If so, then the only difference between automaticity in RB and II tasks is that the terminal projection in RB tasks is onto premotor rule-sensitive neurons, whereas in II tasks the terminal projection is onto premotor response-sensitive neurons (Hélie et al., 2015). In other words, after extensive training, in RB tasks the sight of a familiar stimulus automatically triggers the appropriate rule, whereas in II tasks the sight of a familiar stimulus automatically triggers the appropriate motor response.

The neural architecture of the model, which we call the Cortex Automatizes Rules Model (CARM), is described in Figure 2. For clarity, this figure focuses exclusively on the neural structures that mediate the transition to automaticity, and it omits the structures that mediate initial learning. The complete model would combine Figures 1 and 2. For example, in the Figure 1 model, the ACC facilitates rule selection, the basal ganglia (head of the caudate and GPi) facilitate switching from one rule to another, and the hippocampus is critical for keeping track of which rules have already been tested and rejected. None of these processes are relevant for automaticity because the development of automaticity cannot begin until the correct rule has been discovered.

Figure 2 describes a hypothetical case where a selected rule – referred to as Rule 1 – is practiced enough so that its application eventually becomes automatic. In the Figure 2 scenario, each application of Rule 1 results in either an A or B response (e.g., a button

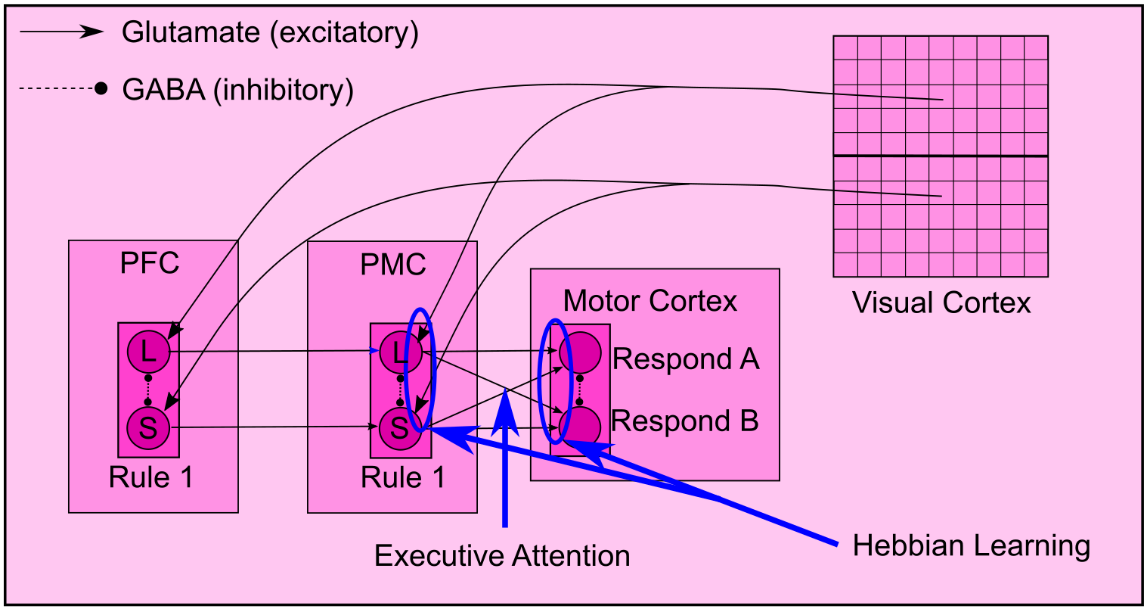

Figure 2. The neural architecture of CARM for an application to a one-dimensional categorization task in which the automatized rule is designated as Rule 1. According to this rule, response A is given if the presented stimulus has a large value on the single relevant dimension, and response B is given if the value is small. PFC = prefrontal cortex, PMC = premotor cortex.

press). Each rule unit includes two simulated neurons – one that signals that the stimulus has a large value on the selected dimension (the L unit), and one that signals a small value on this dimension (the S unit). For example, suppose Rule 1 is to decide if the orientation of an object (e.g., a line or grating) is steep or shallow. In this case orientation-sensitive units in visual cortex that respond to steep orientations would project to the PFC-L Rule 1 neuron, whereas visual cortical units that respond to shallow orientations would project to the PFC-S neuron. In this way, the L neuron responds to any steep orientation and the S neuron responds to any shallow orientation. We assume that rule units develop as a result of life-long practice with a rule. For example, before participating in a laboratory experiment, a person will have many years of practice deciding whether some orientation is steep or shallow.

The Rule 1 units in PFC and PMC are identical except for learning. Although the concepts of steep and shallow orientations are familiar to all adults, in any particular context, the criterion that separates steep from shallow is arbitrary. We assume that the PFC rule units can be quickly tuned to whatever criterion is currently relevant, whereas the PMC rule units adapt more slowly. Evidence supporting this assumption comes from the many studies showing that the PFC is critical for early rule learning. If the PMC motor units were also quickly adjustable, then the PFC would be unnecessary for rule learning.

We propose that the PMC rule units learn the relevant criterion via Hebbian learning at synapses between visual cortex and PMC, and that this learning is facilitated by input from PFC rule units. For example, consider an early-learning trial when the stimulus activates visual neurons that project to the PFC-L rule neuron. These same visual neurons will also project to both the PMC-L and PMC-S rule units because at this early stage of learning, the PMC will not yet have learned the criterion that separates large and small stimulus values. Initially, the PMC-L and S units will receive equally strong visual input. Even so, the PMC-L unit will receive much stronger PFC input than the PMC-S unit, and so there will be more overall activation in the PMC-L unit than in the PMC-S unit, allowing the correct motor response to be selected. Thus, initially, the PFC input is necessary for accurate responding. But note that the greater activation in the PMC-L unit will cause Hebbian learning to increase the strength of the synapses between visual cortex and PMC more in the L unit than in the S unit (i.e., because the post-synaptic activation is greater in the PMC-L unit). Eventually, the visual cortex to PMC rule unit projections will be strong enough that input from PFC is no longer needed for correct responding. At this point, rule application has become automatic.

In laboratory experiments, participants will be given explicit instructions to indicate their response in some way, for example by pressing the "A" or "B" keys. Of course, even though typical participants will have extensive prior experience with determining whether an orientation is steep or shallow, they will have no prior experience associating either steep or shallow orientations with any particular button presses. So whereas we assume that the projections from visual cortex to the PFC rule units are preset, and the projections from the PFC rule unit to the PMC rule unit are preset, we assume that there are no prior preferential connections between the PMC rule unit and units in motor cortex that initiate the selected motor response. Even so, note that participants instructed to press A and B keys do so without error from trial 1 (i.e., they typically do not press other keys incorrectly). Thus, we assume that the experimenter instructions to press key A or B are implemented via top-down executive attention directed at projections from PMC to primary motor cortex. We also assume that there is Hebbian learning at synapses between PMC and primary motor cortex. This Hebbian learning will strengthen the active connections – eventually allowing participants to execute the appropriate motor response without executive attention.

Computational Details

This section describes computational details of CARM.

Visual Cortex. We modeled visual cortex as either a 100 × 2 (Application 1) or 100 × 100 (Applications 2 and 3) grid of units. We assumed that each unit responds maximally when its preferred stimulus is presented and that its response decreases as a Gaussian function of the distance in stimulus-space between the stimulus preferred by that unit and the presented stimulus. In the present applications, we assumed an exceedingly simple model in which the activation of each visual cortical unit is either off (with activation 0) or equal to some positive constant value during the duration of stimulus presentation. Specifically, we assumed that when a stimulus is presented, the activation in sensory cortical unit K at time t equals

$$A_K(t) = 50 \ \exp\left[-\frac{d(K, stimulus)}{\omega}\right] \tag{1}$$

where ω is a constant that determines the width of the receptive field, and $d(K, stimulus)$

is the Euclidean distance (in stimulus space) between the stimulus preferred by unit K and the presented stimulus. Equation 1, which is an example of a radial basis function (Buhmann, 2003), is a popular method for modeling the receptive fields of sensory units in models of categorization (e.g., Ashby et al., 2007; Kruschke, 1992).

PFC, PMC, and Motor Cortex. We modeled all units in PFC, PMC, and primary motor cortex as Izhikevich (2003) regular-spiking neurons (based on results reported, e.g., by Connors, Gutnick, & Prince, 1982; Dégenètais, Thierry, Glowinski, & Gioanni, 2002). According to this model, the intracellular voltage in a unit at time t, denoted by $V(t)$, equals

$$\begin{aligned} 100\frac{dV(t)}{dt} &= I(t) + .07[V(t)+60][V(t)+40] - U(t) + \epsilon(t) \\ \frac{dU(t)}{dt} &= -.06[V(t)+60] - .03U(t), \end{aligned} \tag{2}$$

where $I(t)$ represents all inputs to the unit, $U(t)$ models slow changes in intracellular ion concentrations, and $\epsilon(t)$ is white noise (i.e., mean 0 and variance 1). Equation 2 models continuous changes in intracellular voltage. Therefore, to generate spikes, the voltage is reset to -50 mV (i.e., the resting potential) when $V(t) = 35$ mV. At the same time, $U(t)$ is reset to $U(t) + 100$.

There are two types of inputs – constants from visual cortex and spikes from units in PFC and PMC. We modeled the postsynaptic effects of each presynaptic spike using the alpha function (Ashby, 2018; Rall, 1967), which is a standard method for modeling the temporal smearing and delays that occur when the effects of a presynaptic spike cross a synapse. If a spike occurs at time $t = 0$ in the presynpatic neuron, then the input to the postsynaptic neuron is

$$\alpha(t) = \begin{cases} .05t \exp\left(\frac{20-t}{20},\right) & t > 0 \\ 0, & t < 0 \end{cases} \tag{3}$$

This function increases to a maximum value of 1.0 after 20 msec, and then decays back to 0. If the presynaptic neuron spikes at times $t_1, t_2, ..., t_N$, then the following input is delivered to the postsynaptic neuron:

$$F(t) = \sum_{i=1}^{N} \alpha(t - t_i). \tag{4}$$

Figure 2 shows that the only inputs to each PFC unit are from visual cortex and lateral inhibition from the other PFC unit. Each one-dimensional rule learned by CARM has the form “give one response if the stimulus has a large value on the selected dimension, and give the contrasting response if the stimulus has a small value on this dimension.” As mentioned earlier, we modeled each PFC rule unit with two neurons – one that receives input from visual units that respond to stimuli with large values on the selected dimension and one that receives input from visual units that respond to stimuli with small values on that dimension. Thus, the inputs to the PFC rule unit associated with large values on the selected dimension were

$$I_{PFC_L}(t) = \left[\sum_{K \in \mathrm{L}} A_K(t)\right] - F_{PFC_S}(t), \tag{5}$$

where L is the set of all visual cortical neurons that are maximally sensitive to stimuli with large values on the selected dimension, and $F_{PFC_S}(t)$ is as in Equation 4 where the spikes are from the PFC-S unit. The input to the other PFC unit is analogous (except with the set S replacing L).

Each PMC rule unit receives three types of input – excitatory input from visual cortex, excitatory input from the analogous rule unit in PFC, and lateral inhibition from the other PMC neuron (i.e., see Figure 2). Thus, for example, the input to the PMC-L rule unit was

$$I_{PMC_L}(t) = W_{VC \to PMC}\left[\sum_{\text{all } K} A_K(t)\right] + W_{PFC \to PMC} F_{PFC_L}(t) - F_{PMC_S}(t), \tag{6}$$

where $W_{VC \to PMC}$ represents the strength of the synapse between visual cortex and PMC and$W_{PFC \to PMC}$ represents the strength of the synapse between PFC and PMC.

Finally, the units in motor cortex receive excitatory input from both PMC neurons and lateral inhibition from the other motor unit. Thus, for example

$$\begin{aligned} I_{Motor_A}(t) = W_{PMC_L \to Motor_A}\Phi_{LA}F_{PMC_L}(t) \; &+ \; W_{PMC_S \to Motor_A}\Phi_{SA}F_{PMC_S}(t) \\ &- \; F_{Motor_B}(t), \end{aligned} \tag{7}$$

where Φ_{LA} and Φ_{SA} represent the attentional gains on the projections from the premotor L and S units to motor unit A, respectively. For example, suppose participants are instructed to press response button A when the stimulus is in category A and button B when the stimulus is in category B, and consider a task in which category A stimuli have large values on the relevant stimulus dimension and category B stimuli have small values. After initial category learning is complete, the premotor L unit will cross threshold before the premotor S unit on trials when the stimulus belongs to category A. To complete this response, the participant needs to execute a motor program that causes the finger to depress the A button. This association – between category A and the motor program that causes the A button to be depressed – is not the result of trial-by-trial learning, but rather is the immediate consequence of the experimenter's instructions. We model the effects of these instructions by setting $\Phi_{LA} = .9$ and $\Phi_{SA} = .1$. Furthermore, we assume that the gains on projections from the premotor L and S units to any motor units other than A and B are zero (e.g., the gain equals zero on the projection from the premotor L unit to the motor unit that causes the participant to press the Z button).

Hebbian Learning. As described earlier, Hebbian learning occurs at synapses between visual cortex and PMC and at synapses between PMC and motor cortex. Following standard Hebbian rules, we assumed that plasticity at these synapses depends only on the product of synapse-specific pre- and post-synaptic activation. Specifically, we assumed that strengthening of the synapse required post-synaptic NMDA receptor activation. Activation below this threshold weakened the synapse.

Let $W_{\mathrm{A,B}}(n)$ denote the strength of the synapse on trial n between presynaptic unit A and postsynaptic unit B, and let $V_{\mathrm{J}}(t)$ denote the intracellular activation in unit J (J = A or B) at time t. The key variables to compute are the integrated alpha functions of units A and B. Suppose the time between stimulus presentation and response is T. Then define

$$G_{\mathrm{J}}(T) = \int_0^T F_{\mathrm{J}}(t)\mathrm{dt}, \tag{8}$$

for J = A or B, and where $F_{\mathrm{J}}(t)$ is as in Equation 4 with the spikes generated in unit J. Note that $G_{\mathrm{J}}(T)$ describes the total postsynaptic effect of all spikes produced by unit J during the duration of the trial. Given these definitions, we used the following difference equation to adjust the strength of $W_{\mathrm{A,B}}(n)$.

$$\begin{aligned} W_{\mathrm{A,B}}(n+1) &= W_{\mathrm{A,B}}(n) \\ &+ \alpha_W \; G_{\mathrm{A}}(T)\,[G_{\mathrm{B}}(T) - \theta_{\mathrm{NMDA}}]^{+}\,[W_{\max} - W_{\mathrm{A,B}}(n)] \\ &- \alpha_W \; G_{\mathrm{A}}(T)\,[\theta_{\mathrm{NMDA}} - G_{\mathrm{B}}(T)]^{+}\,W_{\mathrm{A,B}}(n), \end{aligned} \tag{9}$$

where θ_{NMDA} denotes the threshold for activation of postsynaptic NMDA receptors. The terms α_W, θ_{NMDA}, and $W_{\max}$ are all constants. The function $[f(t)]^{+}$ equals $f(t)$ when $f(t) > 0$, and 0 when $f(t) \leq 0$. Thus, $[G_{\mathrm{B}}(T) - \theta_{\mathrm{NMDA}}]^{+}$ measures the total amount of post-synaptic activation above NMDA activation threshold. $[W_{\max} - W_{\mathrm{A,B}}(n)]$ is a rate-limiting term that prevents synaptic strength from exceeding $W_{\max}$. The constant α_W is the learning rate. In brain regions that are targets of dopamine but that lack fast dopamine reuptake, such as frontal cortex, α_W might be assumed to fluctuate with dopamine levels.

The second (positive) term describes the conditions under which LTP occurs – that is, when postsynaptic activation is great enough to activate NMDA receptors. Note that this term guarantees that the increase in synaptic strength is proportional to the product of the pre- and postsynaptic activations (and the final rate limiting term that prevents the strength of the synapse from exceeding $W_{\max}$). The third (negative) term describes conditions that produce LTD (postsynaptic activation below the threshold for NMDA activation). Most Hebbian learning rules do not include any mechanism to decrease synaptic strength, so this last term is unusual.[2] First, note that this term equals 0 except when total postsynaptic activation is below the NMDA-receptor threshold. The $W_{\mathrm{A,B}}(n)$ at the end prevents synaptic strength from dropping below 0.

Equation 9 required some slight modification for the synapses between PMC and motor cortex. We assumed that plasticity at these synapses follows the same Hebbian rules as synapses between visual cortex and PMC. However, note that Equation 9 is not synapse specific. For example, consider two different synapses on the same postsynaptic neuron – one that receives weak presynaptic input that by itself is not strong enough to drive the postsynaptic neuron above threshold for NMDA receptor activation, and one that receives input that is strong enough to activate postsynaptic NMDA receptors. Note that Equation 9 would strengthen both of these synapses because activation in the postsynaptic neuron is

[2] While including a negative term in Hebbian learning is rare in computational neuroscience applications, its has a long history in the traditional connectionist modeling literature (e.g., contrastive Hebbian learning, anti-Hebbian learning). A selected review of this history and its computational role can be found in Ross, Chartier, and Hélie (2017).

above NMDA threshold. However, in the mammalian brain, synaptic plasticity is synapse specific. Specifically, in a real brain, only the synapse receiving strong presynaptic input would be strengthened.

This is not a problem for synapses between visual cortex and PMC. Visual units that respond strongly to the presented stimulus initially project to both PMC rule units, but only the PMC rule unit that triggers the correct response will have strong postsynaptic activation (i.e., because it also receives strong PFC input). Therefore, by Equation 9, synaptic strengthening will primarily occur only at synapses between visual cortex and the correct PMC rule unit.

On the other hand, Equation 9 does not properly adjust the strength of synapses between PMC and motor cortex. As shown in Figure 2, there are four such synapses in the model. The PMC-L unit projects to both the motor-A and motor-B units, which for shorthand we call the LA and LB synapses, and there are similar SA and SB synapses. To illustrate the problem, consider an early training trial in which the stimulus has a large value on the selected dimension and the correct response is A. After the correct rule has been discovered, presynaptic activity on this trial will be high in PMC-L and low in PMC-S, whereas postsynaptic activity will be high in motor-A and low in motor-B (because of executive attentional biasing). Therefore, the only synapse where pre- and postsynaptic activation will both be high is LA. Thus, according to current models of long-term potentiation, this is the only synapse that should be strengthened. However, because postsynaptic activation is high in motor-A unit, Equation 9 will strengthen both LA and SA. For this reason, we need to replace Equation 9 with a Hebbian learning scheme that strengthens LA, but not SA, LB, or SB.

Our solution was to remove the postsynaptic term from Equation 9 and make plasticity at each synapse depend only the postsynaptic effect of the presynaptic activation. However, the effects of premotor activation on activity in the motor cortex units depends not only on activity within the premotor units, but also on the strength of the premotor-to-motor synapse and on the attentional gain. Therefore, at synapses between PMC unit J and motor cortex unit I, we modified synaptic strength as follows.

$$\begin{aligned} W_{PMC_J \rightarrow Motor_I}(n+1) = {} & W_{PMC_J \rightarrow Motor_I}(n) \\ & + \alpha_w \; [W_{PMC_J \rightarrow Motor_I}(n) \; \Phi_{JI} \; G_{\mathrm{A}}(T) - \theta_{\mathrm{NMDA}}]^{+} \, [W_{\max} - W_{PMC_J \rightarrow Motor_I}(n)] \\ & - \alpha_w \; [\theta_{\mathrm{NMDA}} - W_{PMC_J \rightarrow Motor_I}(n) \; \Phi_{JI} \; G_{\mathrm{A}}(T)]^{+} \, W_{\mathrm{A,B}}(n). \end{aligned} \tag{10}$$

The constant θ_{NMDA} still denotes the threshold for postsynaptic NMDA-receptor activation, but Equation 10 now strengthens the synapse only if input at the synapse between premotor unit J and motor unit I is strong enough to drive the postsynaptic activation above this threshold.

To see how this model works, consider the same trial as before in which the stimulus has a large value on the selected dimension and the correct response is A. On this trial, there will be strong presynaptic activation only at the LA synapse. Presynaptic activation will be weak at the other three synapses (e.g., it is weak at LB because the attentional gain Φ_{LB} is small). Thus, in agreement with current models of LTP and LTD, Equation 10 only strengthens the LA synapse.

Initial Category Learning. This article proposes a novel theory of how automaticity develops in rule-guided tasks. Of course, automaticity can only develop after the correct rule is discovered, so the theory proposed here focuses on neural changes that occur after rule discovery is complete.[3] Even so, to simulate the entire learning process – from initial rule discovery to automaticity – we augmented CARM with the COVIS model of rule learning (Ashby et al., 1998; Ashby, Paul, & Maddox, 2011) and the FROST model of working memory maintenance (Ashby et al., 2005). We call this augmented model CARM$^+$. A schematic of the neural structures of CARM$^+$, when applied to a dual-task experiment, is shown in Figure 5 below.

FROST assumes that representations of all items that are active in working memory – including the current categorization rule – are maintained via persistent activations in separate PFC working-memory units. Activation in these units is maintained during delay periods via reverberating activation between PFC and the medial dorsal nucleus (MDN) of the thalamus. An excitatory signal from the PFC to the head of the caudate nucleus during the time when working memory is needed causes the internal segment of the globus pallidus to disinhibit the MDN. FROST assumes no upper limit on the number of PFC working memory units that can be active simultaneously. Even so, as the working memory load increases, so does the number of active working memory units. FROST assumes lateral inhibition among these units, so the more units that are active, the more lateral inhibition there is on each unit. Ashby et al. (2005) showed that this model accurately accounts for limitations on working memory span (e.g., the magic number 7 ± 2).

The COVIS model of rule learning (Ashby et al., 2011) was used to model the initial rule-discovery process. This model assigns a weight to each alternative rule that depends on initial salience and the rule's past history of success. In addition, the active rule receives a bonus because of the natural human tendency to perseverate, and the model mimics exploration by increasing the weight of a randomly selected rule by a random amount. The probability that each rule is then used on the upcoming trial is proportional to its assigned weight. This algorithm was used to select a rule for application on each trial, and then the selected rule was implemented via the CARM$^+$ architecture. After the correct rule is discovered, which occurs within the first 100 trials or so of the first training session in the applications considered below, COVIS perseverates on this rule and no more rule switching occurs. Therefore, in the applications below, COVIS only affects performance of CARM$^+$ during the initial block or two of the first training session.

Empirical Tests of Model

This section describes empirical tests of the model. Before considering detailed applications, note that the model naturally predicts increases in accuracy and decreases in RT as training continues. Accuracy increases because of synaptic strengthening on the units that initiate the correct response, and RT decreases for two reasons. Responding gets faster because of synaptic strengthening, but more importantly, RT decreases because the PFC plays an ever diminishing role in response selection as training progresses. Eventually it plays no role, and instead, PMC activation in the correct rule unit is driven above response threshold via visual input alone. The model responds considerably faster without PFC in-

[3] We treat "rule discovery" and "rule learning" as synonyms in this article.

volvement because under these conditions, the pathway from visual cortex to motor cortex is more direct (with fewer synapses; see Figure 2).

Many current models predict increases in accuracy and decreases in RT as training progresses, so rather than documenting these well-studied effects, our focus will be on empirical phenomena that are problematic for standard cognitive models of automaticity, such as the instance model (Logan, 1988), the exemplar-based random walk model (Nosofsky & Palmeri, 1997), or the component power laws model (Rickard, 1997). This section considers three such phenomena – one electrophysiological and two behavioral. First, we show that the model correctly predicts that early in training PFC rule neurons fire before PMC rule neurons, but that this ordering reverses after automaticity has developed. Second, we show that the model correctly predicts that a dual task that requires executive attention and working memory interferes with early rule learning but not with automatic rule-guided behavior. Finally, we show that the model correctly predicts that during early rule learning, switching the location of the response keys has little or no effect on RB categorization, but the same switch causes significant interference after automaticity has developed. We know of no other models of automaticity that can account for these phenomena.

Application 1: Electrophysiology

Wallis and Miller (2003) reported the results of an experiment in which two rhesus monkeys practiced applying two rules every day for several months. On each trial, a visual image was displayed, and then the animals were given a cue that signaled whether they should apply a *same* rule or a *different* rule. Next, a second image was displayed. If the cue to apply the *same* rule was presented, then the task was to respond if the images were the same (by releasing a lever) and not to respond if the images were different. If the cue to apply the *different* rule was presented, then the task was to respond if the images were different and not to respond if they were the same. Each monkey completed approximately 700 correct trials per day for several months. Later, Muhammad et al. (2006) reported results from a third monkey who completed the same training.

After training was complete, Wallis and Miller (2003) collected single-unit recordings from neurons in PFC and PMC that were rule selective – that is, from neurons that fired during application of one of the rules but not during the other, regardless of what stimulus was shown and which cue was used to signal the rule. The right column of Figure 3 shows the estimated likelihood that a randomly sampled rule-selective neuron in PFC (panel e) or PMC (panel f) produce any given latency, where latency is defined as the time between cue onset and a significant increase in firing. Note that on average, the rule-selective PMC neurons fired *before* rule-selective neurons in PFC. Specifically, the median onset of rule-selective neurons was 270 ms in PMC and 330 ms in PFC.

This result is surprising since it implies that after automaticity has developed, rule selection in the PMC may not be driven by PFC input. PFC neurons cannot be causing PMC activation if they fire after the onset of PMC firing. Wallis and Miller (2003) did not collect similar recordings during early training, but as mentioned previously, a wealth of data suggests that initial rule learning depends heavily on PFC. So presumably, similar recordings from early training sessions would show PFC rule-selective neurons firing *before* PMC neurons. Thus, these data suggest that one property of automaticity is that during its development, control is gradually transferred from PFC to PMC.

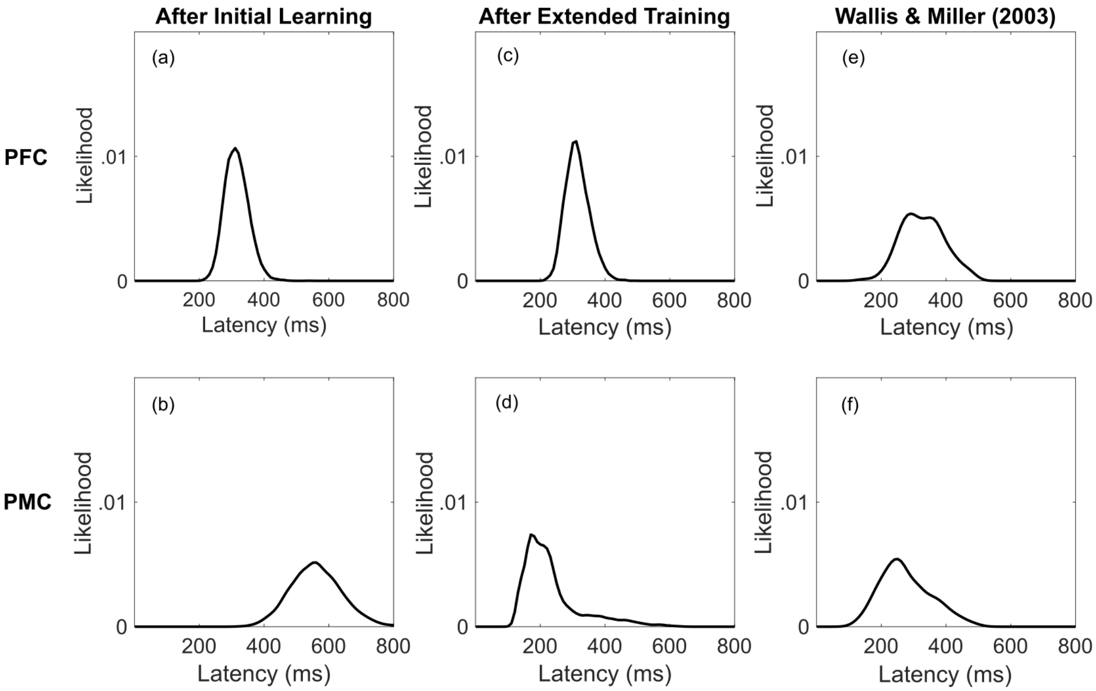

Figure 3. Probability density function estimates from rule-selective neurons in PFC (first row) and PMC (second row). Panels a – d show predictions of CARM, and panels e and f show estimates for single neurons that were reported by Wallis and Miller (2003). In the case of CARM, the estimates are the likelihood that the same neuron would produce any given latency during multiple independent simulations of the task. In the case of the Wallis and Miller (2003) data, the estimates are the likelihood that a randomly sampled rule-selective neuron in PFC (panel e) or PMC (panel f) would produce any given latency.

We modeled the Wallis and Miller (2003) task by training CARM to apply a *same* or *different* rule to pairs of visual images. The images were 12 grayscale photographs selected from the internet[4] and recorded with a resolution of 300 × 300 pixels. On half the trials, two copies of the same image were presented, and on the remaining trials two randomly selected different images were presented. Independent noise was added to each pixel value on every trial. Like the monkeys, CARM was trained to respond if the images were the same on *same-rule* trials and not to respond if the images were different. On *different-rule* trials, CARM was trained to respond if the images were different and not to respond if the images were the same (again, same as the monkeys).

The input to each *same-rule* unit was a perceived similarity value and the input to each *different-rule* unit was a perceived dissimilarity value.[5] We assumed that similarity

[4] Six of the photographs were of animals, 3 were outdoor scenes, 2 were abstract images, and 1 was a human face.

[5] Note that many psychological theories assume that a variety of different perceptual and cognitive decisions are based on such similarity values, and therefore, all of these theories assume that visual similarities are computed in some brain region. Within the categorization literature, a prominent example is exemplar theory (e.g., Nosofsky, 1986).

and dissimilarity were computed in some region of visual association cortex (or prefrontal cortex; see Davis, Goldwater, & Giron, 2017) that projects to the PFC rule units. Because the images were chosen to all be highly dissimilar from each other, the metric used to compute similarity and dissimilarity is relatively unimportant. Any metrics that produce higher similarity and lower dissimilarity values for same than for different pairs should produce similar results to those reported in this section. Therefore, for convenience, we chose the metrics used in the most popular versions of representational similarity analysis when applied to fMRI data (e.g., Ashby, 2019; Kriegeskorte, Mur, & Bandettini, 2008). Specifically, we defined the similarity between two images as the Pearson correlation between their 90,000 (i.e., 300^2) pixel values (each an integer between 0 and 256), and we defined their dissimilarity as one minus this value. The model included 200 units in the visual-cortical similarity/dissimilarity region. Half of these units responded to a specific preferred similarity value and half responded to a specific preferred dissimilarity value. In both cases, the preferred values ranged from .01 to 1 (in units of .01), and as described above, the tuning curve of each unit was modeled with a radial basis function. As in all other applications, the initial visual projections were selective to PFC units and nonselective to PMC units. For example, the visual units that responded to similarity projected selectively to the appropriate unit in the PFC *same-rule* complex and nonselectively to both units in the PMC *same-rule* complex.

To examine predictions of CARM in the Wallis and Miller (2003) experiment, we trained the model using the same experimental procedures as Wallis and Miller. We divided the data into three phases: 1) an initial baseline phase to estimate PFC and PMC activity before extended rule training, 2) a training phase of extended practice during which automaticity develops, and 3) a final post-training test phase. The baseline phase assumed that rule discovery was complete – that is, that the model had discovered the correct categorization rule, but that this correct rule had not yet received any extensive practice. To estimate pre-training activity, we set the Hebbian learning rates to 0. On each baseline trial, we recorded the time it took for rule units in the PFC and PMC to reach a threshold level of activation.[6]

The training phase models the development of automaticity. During these trials, the Hebbian learning rate was set to a positive value ($\alpha_W = 1 \times 10^{-8}$), and the model completed 10,000 trials of categorization. The animals in the Wallis and Miller (2003) experiment likely completed more than 10,000 trials of training, although the precise number was not reported.

The test phase was designed to estimate model performance after automaticity had developed. This phase included 300 trials with the Hebbian learning rate set to 0 to mimic standard categorization transfer conditions in which no feedback is provided to the participant. For more methodological details, see the Appendix.

Results are shown in Figure 3. The left column shows the predicted response latency probability density functions during the pre-learning baseline phase and the middle column shows these same estimates from the test phase after automaticity had developed. The first row shows predictions for PFC rule neurons and the second row shows predictions for PMC rule neurons. Note that the time taken for the presented stimulus to drive the relevant PFC

[6] For both regions, we computed the integral of Equation 4 and set the threshold on this integral to 700. Baseline predictions were generated by averaging across 300 such trials.

rule unit above threshold does not vary with training. However, the response latency of PMC rule units decreases dramatically as training progresses – from an average of around 550 msec during baseline to just over 200 msec after automaticity has developed. During the pre-learning phase, note that the PMC rule units fire well after the PFC units. This is because activation in the PMC units is largely driven by PFC input during this early stage of training. In contrast, after automaticity develops, note that the PMC units fire approximately 300 msec *before* the PFC units. After 10,000 trials of training, the PMC units are driven almost exclusively by input from neurons in visual cortex.

As mentioned earlier, Wallis and Miller (2003) did not collect any recordings before automaticity developed. But the substantial literature showing that initial rule learning is mediated largely within the PFC implies that PMC activation during early training is almost certainly driven by input from PFC (e.g., Durstewitz, Vittoz, Floresco, & Seamans, 2010; Strange, Henson, Friston, & Dolan, 2001). Therefore, Figure 3 shows that CARM accounts for a highly non-intuitive electrophysiological phenomenon – namely, that during early learning activation in PFC rule neurons precedes activation in PMC rule neurons, but after automaticity develops this ordering reverses. CARM is the first computational model that can account for this result.

Application 2: Dual Task Interference

During early learning, a simultaneous dual task that requires executive attention and working memory significantly interferes with RB learning and performance (Crossley, Paul, Roeder, & Ashby, 2016; Waldron & Ashby, 2001; Zeithamova & Maddox, 2006). However, after automaticity develops, the same dual task does not interfere with RB categorization (Hélie, Waldschmidt, & Ashby, 2010). In fact, this pattern of results – dual-task interference during early training but not after extended training – is a well-known diagnostic that is often used as a criterion that a behavior has become automatized (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977).

CARM naturally predicts this phenomenon because during early learning it assumes that PFC rule units are necessary for accurate responding, whereas after automaticity develops, PFC participation is no longer needed. More specifically, CARM assumes that activity in the PFC rule units is maintained via working memory. As a result, any allocation of working memory to a dual task will reduce working memory resources available for rule learning. In fact, Ashby et al. (2011) showed that the COVIS component of CARM$^+$ accurately accounts for the dual-task interference during early learning that was reported by Waldron and Ashby (2001). However, this was an abstract computational model that included no neuroscientific detail.

Unfortunately, we know of no studies that examined the effects of a dual task on categorization performance after both initial and extended training in the same group of participants. As a result, this section examines the ability of CARM$^+$ to account for dual task effects by comparing its performance to that of participants in the experiments reported by Zeithamova and Maddox (2006) and Hélie, Waldschmidt, and Ashby (2010). Zeithamova and Maddox (2006) examined the effects of a dual task on initial category learning, whereas Hélie, Waldschmidt, and Ashby (2010) examined dual-task effects on categorization performance after extended categorization training (i.e., 20 sessions). The two studies used the same categorization stimuli (i.e., Gabor disks) and the same dual task (a numerical Stroop task). Figure 4 shows the categories used in the two studies. Although these were somewhat different, note that the same rule maximizes accuracy in both cases.

In both studies, the categorization stimulus was centered between two single-digit numbers that varied across trials in numerical value and physical size. A Stroop-like interference occurs when the physically larger number is numerically smaller (e.g., as in Figure 5). The numbers disappeared and participants then made a categorization response. Next, a cue was presented that informed participants to report either the physically or numerically larger number. Therefore, during categorization, participants were required to maintain the numerical value and physical size of each digit in working memory.

The architecture of CARM$^+$ on a hypothetical dual-task trial of these experiments is shown in Figure 5. The model assumes that representations of the categorization rule and the two dual-task numbers are maintained via persistent activations in separate PFC working-memory units that is facilitated by reverberating activation between PFC and thalamus. As described above, the COVIS model of rule learning (Ashby et al., 2011) was used to model the initial rule-discovery process. On each trial, the rule selected by COVIS was implemented via the architecture shown in Figure 5.

We used this same model to simulate the effects of the dual task on category learning in the experiment described by Zeithamova and Maddox (2006), and in the experiment described by Hélie, Waldschmidt, and Ashby (2010). The only difference in the two simulations was in the stimuli that defined the two contrasting categories (and the amount of training the model received). The results for the Zeithamova and Maddox (2006) experiment are shown in Figure 6, whereas the results for the Hélie, Waldschmidt, and Ashby (2010) experiment are shown in Figure 7. For computational details, see the Appendix. Note that the model accurately accounts for the impaired learning that occurs when the dual task is added to the first session of categorization training, and it also correctly predicts the absence of a dual-task effect on performance after automaticity has developed. It is important to note that exactly the same model was used in both applications, and even the same parameter values. Thus, for example, the amount of lateral inhibition on PFC rule units caused by the dual task was identical during early and late learning.

Our simulations also showed that the model predicts that after automaticity develops, there is no effect of the dual task on response time. This is consistent with classic notions of automaticity (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977). On the other hand, Hélie, Waldschmidt, and Ashby (2010) reported that response times increased under dual-task conditions, despite the absence of any decrease in accuracy.[7] We believe there are two plausible accounts of this discrepancy. One possibility is that the response-time interference reported by Hélie, Waldschmidt, and Ashby (2010) might disappear with more training. We believe a more likely possibility, however, is that the response-time interference occurred because Hélie, Waldschmidt, and Ashby (2010) instructed their participants to maximize accuracy, but they provided no instructions about response time. As a result, the Hélie, Waldschmidt, and Ashby (2010) participants had no motivation to minimize their response times during the dual-task session. More research is needed to investigate these possibilities.

Figure 8 explains why the model predicts this dissociation. The top two panels show predicted categorization accuracy and mean RT for CARM$^+$ across 12,000 trials of the Hélie,

[7] Response times were not reported by Zeithamova and Maddox (2006) or in any of the other previous dual-task category-learning studies.

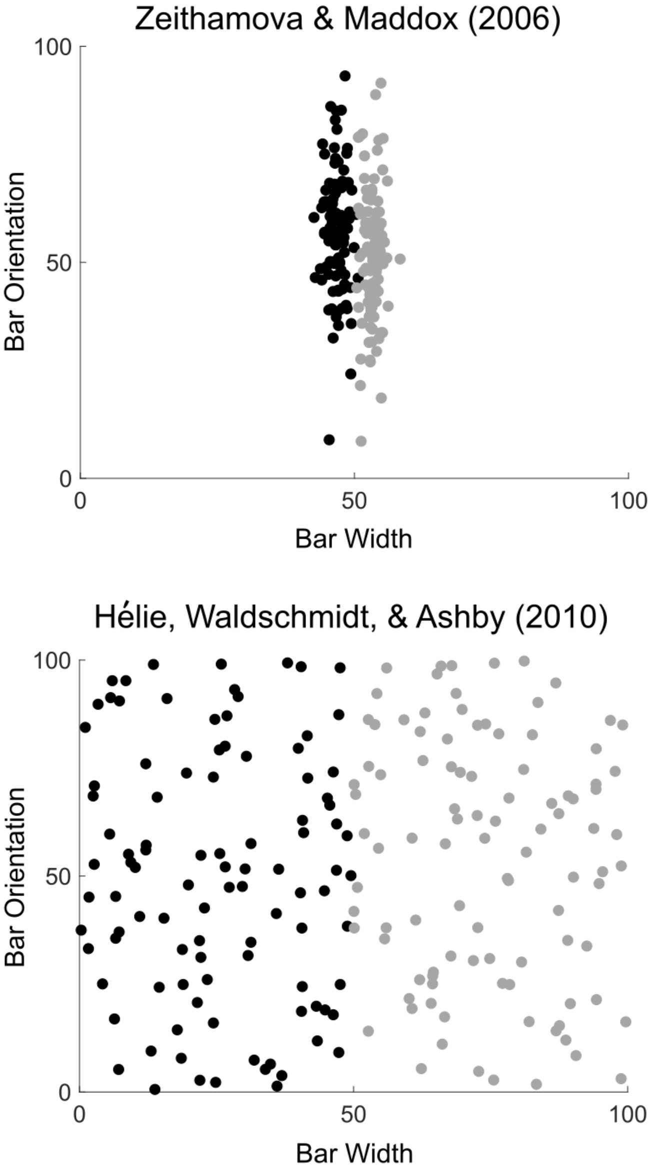

Figure 4. Categories used in the dual-task studies of Zeithamova and Maddox (2006) and Hélie, Waldschmidt, and Ashby (2010). Stimuli in both studies were circular sine-wave gratings that varied in bar width (i.e., spatial frequency) and bar orientation. Black dots denote bar width and orientation of category A exemplars, and gray dots identify exemplars of category B.

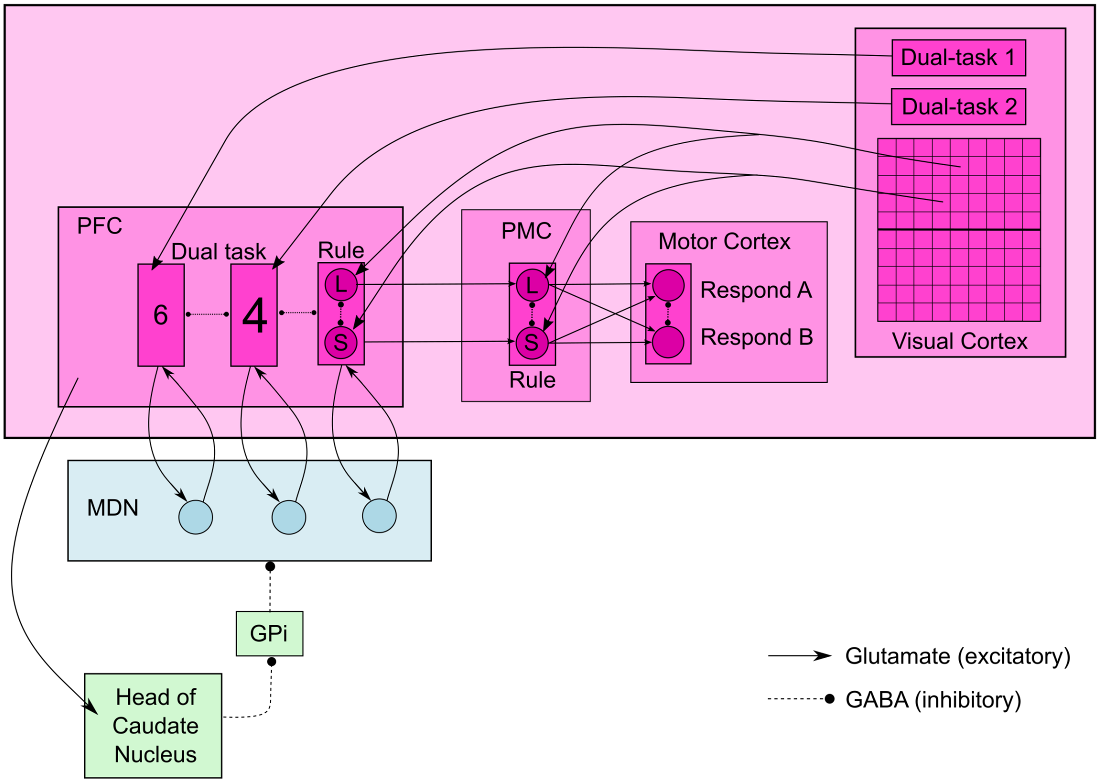

Figure 5. The architecture of CARM$^+$ on a dual-task trial when the numbers that flank the categorization stimulus are a physically small 6 and a physically large 4. PFC = prefrontal cortex, PMC = premotor cortex, MDN = medial dorsal nucleus of the thalamus, GPi = internal segment of the globus pallidus.

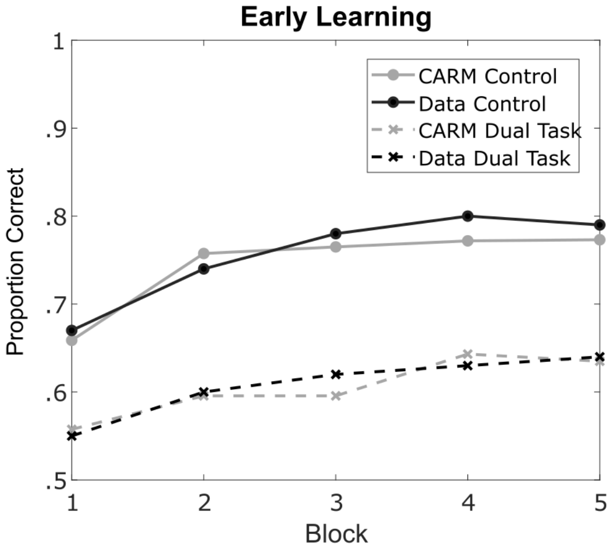

Figure 6. Learning curves reported by Zeithamova and Maddox (2006) for their single-task control and dual-task conditions. Also shown are results from CARM$^+$ in the same two conditions.

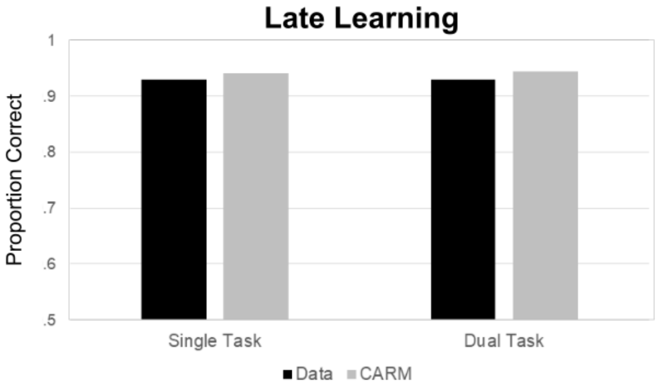

Figure 7. Proportion correct for the Hélie, Waldschmidt, and Ashby (2010) participants and for CARM$^+$ during the last session of training (Single Task) and under dual-task conditions.

Waldschmidt, and Ashby (2010) experiment. The bottom panel shows the proportion of the total activation in the PMC unit that controlled the categorization response that comes from the PFC. Note from Figure 5 that the PMC receives excitatory input from visual cortex and PFC. Initially, the synaptic strength of the visual cortex to PMC projection is weak, so activation in PMC units comes mostly from PFC. As training progresses however, Hebbian learning at the visual cortex/PMC synapses improves the ability of visual cortex to activate PMC. The bottom panel of Figure 8 quantifies this effect. A categorization response is generated by CARM when total activation in either PMC unit (i.e., integrated over the course of the trial) first crosses a response threshold. The bottom panel of Figure 8 shows the proportion of that total activation that came from PFC on each trial. Note that the proportion coming from visual cortex is just one minus the PFC value. As can be seeen, the model predicts a gradual transfer of control from PFC to visual cortex that takes approximately 8,000 trials to complete. Early in training, the categorization response is driven almost entirely by PFC and therefore a dual task that consumes PFC resources impairs categorization learning and performance. After automaticity develops however, the categorization response is driven almost entirely by input from visual cortex. As a result, a dual task that affects PFC has no effect on categorization performance.

The bottom panel of Figure 8 also reinforces the widely held view that the development of automaticity is a gradual process that takes thousands of trials to complete (e.g., Logan, 1985). Note that CARM predicts that a signature of this process should be a long-lasting, but ever diminishing contribution of the rule that mediated initial learning. Some data support this prediction. For example, consider simple addition. In support of the counting rule, the response times of young children increase linearly with the magnitude of the smaller of the two addends and the slope of this linear regression is about 400 ms per unit. As predicted by CARM, typical adults show a similar pattern, except with a much smaller slope (of around 20 ms; e.g., Groen & Parkman, 1972; LeFevre, Sadesky, & Bisanz, 1996). Note that CARM also predicts that this effect should continue to decrease with additional training.

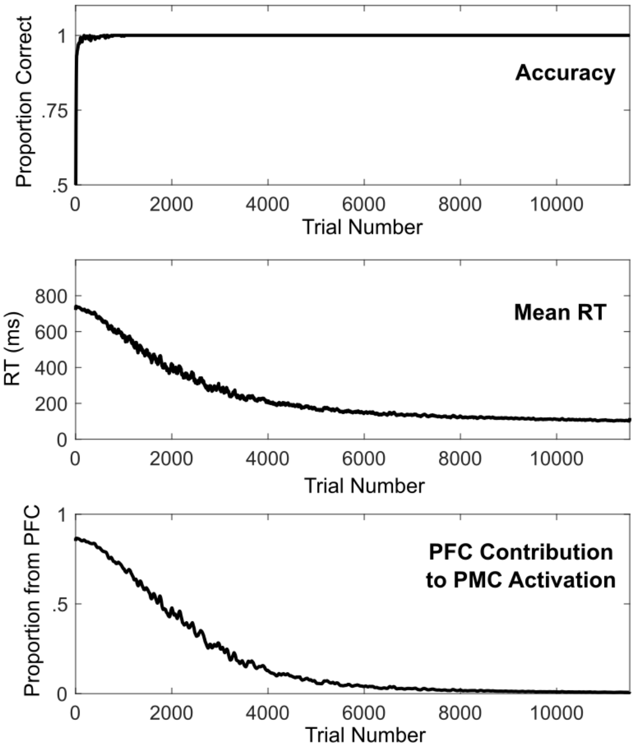

Figure 8. Various CARM predictions in the Hélie, Waldschmidt, and Ashby (2010) experiment. The top panel shows the mean proportion of correct categorization responses across 12,000 trials, the middle panel shows mean categorization RT for these same trials, and the bottom panel shows, for the premotor cortex unit that controlled the categorization response, the proportion of the total activation that came from PFC.

Application 3: Button-Switch Interference

Another popular diagnostic criterion that is often used to determine whether a behavior has become automatized is behavioral inflexibility – that is, automatic behaviors are often disrupted when the behavioral requirements are changed in any way (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977). For example, a number of studies have trained participants on RB or II categories, and then switched the location of the response keys. Participants are instructed in these experiments that the stimuli and categories are identical, but that the location of the two response keys was reversed. Switching the location of the keys after one session of training interferes with II categorization but not with RB categorization (Ashby et al., 2003; Maddox et al., 2004; Spiering & Ashby, 2008). In contrast, this same button switch causes a significant decrease in accuracy and increase in RT in both RB and II tasks if it is first implemented after automaticity has developed (Hélie, Waldschmidt, & Ashby, 2010).

At first glance, this result seems incompatible with CARM. In II categorization tasks, stimulus-response mappings are automatized (Roeder & Ashby, 2016), so switching the response keys interferes with what was learned. But in RB tasks, the rule is automatized

(Roeder & Ashby, 2016), and the rule is independent of the response keys. Put another way, after one session of training, RB categorization is rule guided and switching the response keys causes no interference. After 12,000 trials of training, RB categorization is still rule guided. So why should there now be a button-switch interference?

According to CARM, the development of a button-switch interference after extended training is due to the Hebbian learning that occurs at synapses between PMC and motor cortex. Although the model assumes that rules are automatized in RB tasks (mediated by PMC rule units), it also assumes that pressing the "A" button for every category L stimulus and the "B" button for every S stimulus strengthens associations between the PMC-L unit and the Motor-A unit and between the PMC-S and Motor-B units. The idea is that these associations become strong enough after thousands of trials of practice that top-down executive attention is unable to reverse them completely.

As described earlier, prior to the experiment, participants have no association between stimuli that have a large value on the critical stimulus dimension and any button presses, or between stimuli with small values on this dimension and any button presses. Even so, after given explicit instructions that they should respond by pressing the "A" or "B" buttons, most participants reliably press only these two buttons beginning from the first trial of the experiment. We assume that these response instructions to participants are mediated by top-down executive attention, which we implemented as a gain on projections between PMC and motor cortex (i.e., see Equation 7). In this implementation, instructions to press the "A" or "B" button on each trial causes the gain on projections from PMC to all other possible motor responses to be set to 0 (including all other possible button presses). Furthermore, we assume that "press the A button on L trials and the B button on S trials" is a different rule from "press the A button on S trials and the B button on L trials." And since they are different rules, the development of automaticity cannot begin until the participant has discovered which one is correct. As mentioned earlier, we modeled this rule discovery process using COVIS.

The rule units in PMC are not naturally associated with any button press, so to model the rule "press the A button on L trials and the B button on S trials" we assume that executive attention sets a large gain on the projection from the PMC-L unit to the Motor-A unit (i.e., $\Phi_{LA} = 0.9$) and a small gain on the projection from PMC-L to Motor-B (i.e., $\Phi_{LB} = 0.1$). When instructed that the location of the response keys has switched – that is, that participants should now press the opposite button to indicate their response – we assume that these attentional gains also switch (i.e., from $\Phi_{LA} = 0.9$ and $\Phi_{LB} = 0.1$ to $\Phi_{LA} = 0.1$ and $\Phi_{LB} = 0.9$).

The CARM$^+$ predictions for the effects of a button switch at the end of the first session of training are shown in Figure 9. The "before" block includes the last 100 trials before the button switch and the "after" block includes the 100 trials immediately after the switch. Note that the model correctly predicts no interference if the response buttons are switched at the end of one training session. Initially, the strengths of the four synapses between PMC and motor cortex that are shown in Figure 2 are all equal. The few training trials that occur during initial learning are not enough to cause slow Hebbian learning to change these strengths in any substantial way. Thus, when the button-switch instructions are given, the switch of the attentional gains allows accurate responding to continue with no drop in accuracy.

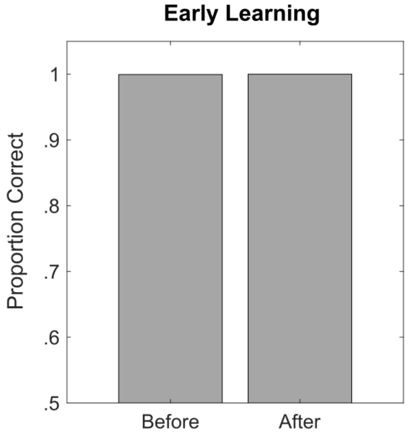

Figure 9. Predicted accuracy of CARM$^+$ during the block of trials before and after a button switch, when the switch occurs at the end of the first session of training.

However, the model does predict that the same button switch after extended training causes a significant drop in accuracy. Figure 10 shows the accuracy of participants across thirteen 50-trial blocks of the experiment reported by Hélie, Waldschmidt, and Ashby (2010). Also shown are predictions from CARM. The accuracies shown at block 0 are the terminal accuracies after approximately 11,000 trials of training. Participants were instructed to switch response buttons between blocks 0 and 1, and these same switched response mappings remained in place for the entire 600-trial experimental session. The participants' drop in accuracy after the button switch was statistically significant, and note that recovery was not complete even after 600 trials of practice.

Figure 11 shows the response times reported by Hélie, Waldschmidt, and Ashby (2010) and predicted by CARM. These should be interpreted with caution because Hélie, Waldschmidt, and Ashby (2010) provided no response-time instructions to their participants. Even so, note that, in agreement with classical notions of automaticity (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977), response times increased by around 66 ms after the button switch, and that CARM predicts a similar increase.

The model predicts the button-switch interference shown in Figures 10 and 11 because after 11,000 trials of pushing the A button on L trials and the B button of S trials, Hebbian learning – even very slow Hebbian learning – significantly increases the strengths of the $PMC_L \rightarrow Motor_A$ and $PMC_S \rightarrow Motor_B$ synapses, relative to the opposite synaptic strengths. Thus, when the attentional gains reverse after the button switch instructions are given, the imbalance in synaptic strengths is great enough to cause a drop in accuracy. Simulation details can be found in the Appendix.

General Discussion

This article proposed a biologically-detailed account of how rule-guided behaviors become automatic. The model successfully predicts many well-known, general automaticity-related phenomena. Included in this list are that 1) accuracy increases and response time decreases with extended practice; 2) initial rule learning is impaired by a simultaneous dual

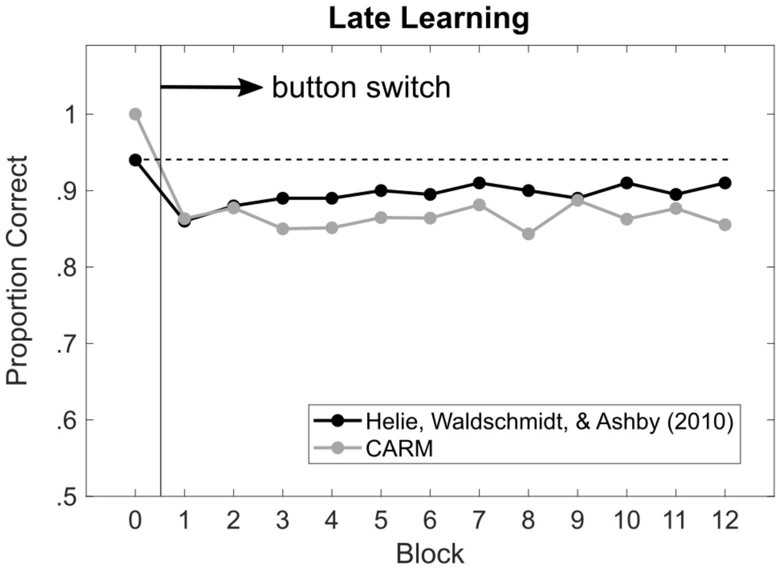

Figure 10. Accuracy of participants in each 50-trial block of the experiment reported by Hélie, Waldschmidt, and Ashby (2010). Block 0 shows terminal accuracy following approximately 11,000 trials of training. Participants were instructed to switch response buttons between blocks 0 and 1. The dotted line indicates expected accuracy in the absence of a button switch. Also shown are predictions of CARM under these same experimental conditions.

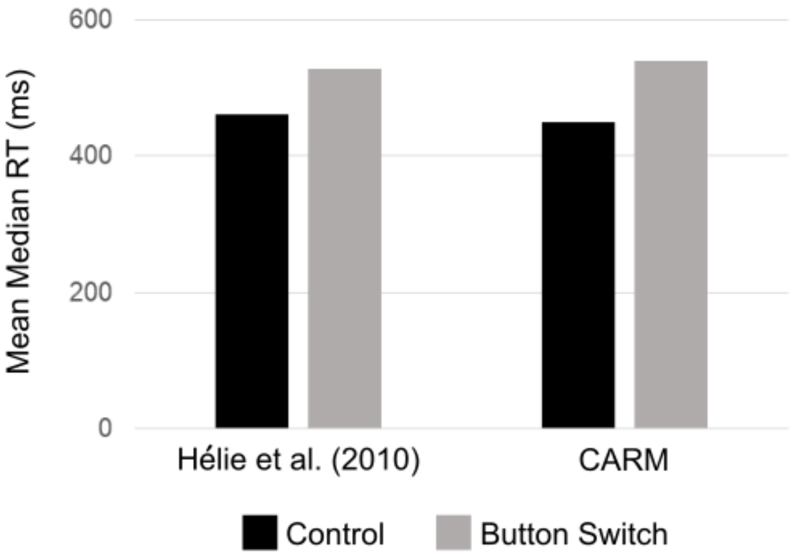

Figure 11. Across-participant means of median response times (RTs) reported by Hélie, Waldschmidt, and Ashby (2010) and predicted by CARM. Control RTs are averaged across the last block before the button switch. A motor time of 364 ms was added to the RTs predicted by CARM.

task, but automatic rule application is immune to dual-task interference; and 3) switching the locations of the response keys has little or no effect on initial rule application, but significantly interferes with automatic performance. Although all of these phenomena have been well-known signatures of automaticity for more than 40 years (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977), to our knowledge, CARM is the first theory to account for all these results simultaneously. In addition, we also showed that CARM successfully accounts for a seemingly counterintuitive neuroscience finding – namely, that rule-sensitive neurons in premotor cortex fire *before* PFC rule neurons when the behavior is automatic, even though these same premotor neurons fire after the PFC neurons during early learning.

In addition to these formal tests of the model, it is important to note that CARM is consistent with many other empirical automaticity phenomena. First, it accounts for the functional neuroimaging data reported by Hélie, Roeder, and Ashby (2010). In this experiment, participants practiced either a simple one-dimensional RB task or an RB task in which the optimal rule was a logical disjunction. Each participant completed more than 11,000 trials of practice, distributed across 20 different experimental sessions. Four of these sessions occurred inside an MRI scanner (sessions 1, 4, 10, and 20). As predicted by CARM, the correlation (across participants) between categorization accuracy and activation decreased with training in both the hippocampus and basal ganglia for both types of RB tasks. In contrast, these correlations increased with training for both tasks in (ventral) premotor cortex.

Second, CARM also accounts for the results of Roeder and Ashby (2016). Recall that in this study, participants practiced on a primary category structure long enough for the behavior to become automatic (i.e., 8,400 trials distributed across 14 sessions). Interspersed with this practice were occasional sessions in which participants practiced on a secondary category structure in which half of the stimuli retained their same stimulus-response (SR) associations (consistent stimuli) as in the primary categories and half switched associations (inconsistent stimuli). When II categories were used for both structures, accuracy was higher and RT was lower for consistent stimuli than for inconsistent stimuli, which suggests that SR associations are automatized in II tasks. However, when RB categories were used for both structures, accuracy and RT did not differ between the two types of stimuli. As noted earlier, this result strongly suggests that rules are automatized in RB tasks.

CARM accounts for the Roeder and Ashby (2016) results because one set of PFC and PMC rule units would be active on days when the primary category is practiced and a different set of rule units would be active on secondary category-structure days (i.e., because the rules were different on these days). Thus, practice on the secondary days would have no effect on the neural representation of the correct rule on primary days, and so the model predicts the same performance on consistent and inconsistent stimuli.

Relation to Earlier Theoretical Work on Automaticity

Neuroscience Accounts. CARM is most similar to the SPEED model of procedural automaticity (Ashby et al., 2007). Both models assume that the development of automaticity is a gradual transfer of control from neural networks that mediate initial learning to direct projections between sensory association areas of cortex and premotor cortex. There are three primary differences between the models. First, and most importantly, they are models of different behaviors. CARM is a model of how automaticity develops for rule-guided behaviors, whereas SPEED models the development of automatic behaviors that were acquired via procedural learning. Second, whereas SPEED assumes the training of these cortical-cortical projections is facilitated by a basal ganglia-mediated procedural-learning system, CARM assumes the facilitation is by a PFC-mediated rule-learning system. Third, SPEED assumes the terminal projections in premotor cortex are onto neurons that instantiate abstract motor goals, whereas CARM assumes the critical premotor targets are rule-sensitive neurons. This latter difference allows SPEED to correctly account for the Roeder and Ashby (2016) II results. The inconsistent stimuli in that study strengthen SR associations in SPEED that are incorrect for the primary category structures and as a result, SR associations are weaker for inconsistent than for consistent stimuli.

Note that both models assume that a primary function of PFC-mediated declarative learning and memory systems and basal ganglia-mediated procedural systems is to train automatic cortical-cortical projections (Hélie et al., 2015). The idea behind both models is that these cortical-cortical networks are incapable, by themselves, of using trial-by-trial feedback to guide learning. This is because there are negligible concentrations of dopamine active transporter (DAT) in cortex (e.g., Varrone & Halldin, 2014), and so dopamine is slow to clear cortical synapses. For example, the delivery of a single food pellet to a hungry rat increases PFC dopamine levels for approximately 30 minutes (Feenstra & Botterblom, 1996). Therefore, cortical dopamine levels are likely to remain above baseline during an entire training session, which means that all active synapses in cortex will get strengthened, even those leading to incorrect responses and negative feedback. For this reason, synaptic plasticity in cortex follows Hebbian, rather than reinforcement learning rules (Feldman, 2009). As a result, sensory cortical-to-premotor networks can only acquire behaviors for which errors are common during initial learning if they are supervised, at least up until errors become sufficiently rare. CARM assumes that for rule-guided behaviors this supervision is provided by a PFC network, whereas SPEED assumes that for procedural-learning mediated behaviors, the supervision is provided by the basal ganglia.

The transfer from the initial learning systems to the automatic sensory-premotor cortical systems is computationally efficient because response time is reduced after the transfer is complete, and because it frees the learning systems for new tasks. Learning requires a high degree of flexibility and plasticity, whereas responding automatically does not. For these reasons, it is inefficient to use the slower learning systems to execute automatic responses.

Despite their similarities, SPEED and CARM have many differences. In the current applications, we augmented CARM with the rule-learning module of COVIS and the FROST model of working memory maintenance to develop a complete model that can account for initial learning and automatic rule-guided behavior. We called this model CARM$^+$. The analogue for SPEED would be to augment it with the procedural-learning module of COVIS, and we can refer to this model as SPEED$^+$.

CARM$^+$ and SPEED$^+$ make many qualitatively different predictions about learning and performance in RB and II tasks. Currently, more than 30 such qualitative differences have been identified and confirmed empirically, and many of these dissociations were replicated in independent labs (for a review, see Ashby & Valentin, 2017). Importantly, virtually all of these are predicted a priori by CARM$^+$ and SPEED$^+$. As just one example, SPEED$^+$ predicts that procedural learning is mediated by dopamine-dependent synaptic plasticity at cortical-striatal synapses. Because the striatum has high concentrations of DAT, striatal dopamine levels that rise after positive feedback return to baseline after just a few seconds. Therefore, SPEED$^+$ predicts that delaying feedback by just a few seconds will impair II learning, whereas CARM$^+$ predicts that such delays will not affect RB learning because of its access to working memory. A variety of independent studies have confirmed these predictions (Crossley & Ashby, 2015; Dunn, Newell, & Kalish, 2012; Maddox, Ashby, & Bohil, 2003; Maddox & Ing, 2005). As another example, we have already seen that CARM$^+$ and SPEED$^+$ correctly predict the RB versus II dissociation in automatic performance reported by Roeder and Ashby (2016).

The models are also anatomically different. They share initial visual areas and motor cortex because they rely on the same eyes for sensory input and effectors for motor output. But otherwise, they mostly rely on distinct neural networks. For example, CARM$^+$ and SPEED$^+$ both assign roles to the basal ganglia, but CARM$^+$ depends on the head of the caudate nucleus, whereas SPEED$^+$ depends on the body and tail of the caudate and on the posterior putamen (Cantwell et al., 2015). The most uncertainty about the models is in the precise location of their premotor targets. The models predict that the premotor units in the two models are different, since CARM$^+$ assumes these are rule-sensitive units, whereas SPEED$^+$ assumes they are units that respond to motor goals. However, more neuroscience research is needed to clarify their exact locations within premotor cortex. Even so, note that the premotor rule units in CARM$^+$ must receive prominent input from PFC, whereas the premotor response units in SPEED$^+$ must receive prominent input from the ventral-lateral nucleus of the thalamus, which is the target of posterior putamen.

Cognitive Accounts. The most widely known cognitive models of automaticity assign prominent roles to memory representations associated with single trials or instances. Included in this list are the instance theory of Logan (1988) and the EBRW model of Nosofsky and Palmeri (1997). CARM is fundamentally different from such models in that it assumes that no instances are ever recalled or activated. Instead, CARM only applies to rule-guided behaviors, and it assumes that for such behaviors, learning is a process of discovering the explicit rule that is optimal for the task. Once this rule is discovered, CARM assumes it is applied on every trial without any reference to specific previous instances.

On the other hand, it is important to acknowledge that there is good evidence that memory representations of specific instances sometimes play a key role in category learning – especially during the initial phases of learning (Smith & Minda, 1998) or if the to-be-learned categories include distinct exceptions (Davis, Love, & Preston, 2011). Even so, these studies did not use RB categorization tasks, and the role that the memory of specific instances play in the learning of rule-guided behaviors is unclear. CARM actually predicts faster responding to previously seen stimuli – because of Hebbian learning between visual cortex and PMC – even though the model does not store or activate any instance-based memories. Clearly though, the role that the memory of previous instances plays in rule-guided behaviors is an important topic for future research.

CARM assumes that the development of automaticity is a gradual transfer of control from rule application to behavior that is elicited simply by visual access to the stimulus. The EBRW assumes that the same process is used to respond on every trial. Responding is faster after extensive training only because there are more stored instances available to guide responding. So CARM and the EBRW are fundamentally different. In contrast, the instance model also assumes that the development of automaticity is a gradual transfer of control from one process to another. In this sense then, CARM could be viewed as a sort of neural interpretation of the instance model.

New Predictions and Future Work

Unlike previous cognitive models of automaticity (e.g., Logan, 1988; Nosofsky & Palmeri, 1997), CARM makes strong predictions about the neural networks and processes that mediate the transfer to automaticity. Therefore, compared to cognitive models, CARM has the potential to account for a much greater variety of data (Ashby & Helie, 2011). Whereas the cognitive models are limited to making predictions about response accuracy and response time, CARM makes predictions about these same behavioral data, but in addition, it also can be tested against a wide variety of neuroscience data. This includes single-unit recordings, but it could also be rigorously tested against fMRI BOLD data (via model-based fMRI methods) and EEG recordings. In addition, unlike cognitive models, CARM could be used to make predictions about how transcranial magnetic stimulation, neuropsychological disease, or pharmacological intervention might affect the development of automaticity in rule-guided tasks (for an example application with sequence production, see, e.g., Hélie, Roeder, Vucovich, Rünger, & Ashby, 2015). Future work should be devoted to such tests.

Another interesting prediction of CARM is that both the PFC and PMC contain rule-sensitive neurons. In each brain area, rules were represented using multiple simulated neurons, each corresponding to discrete (qualitative) values on the rule dimension. For example, if a rule specifies that long lines belong to category A while short lines belong to category B, CARM would include two units representing that rule in both the PFC and PMC (one for long lines and another for short lines). While simple rules of this form were useful for the initial tests of the model described in this article, rules can be arbitrarily complex and so future work should focus on establishing how rule complexity affects their representation.

One intriguing hypothesis is that rule-sensitive neurons in the PFC implement the rule and respond to perceptually similar stimuli (Freedman, Riesenhuber, Poggio, & Miller, 2003), whereas rule-sensitive neurons in the PMC represent the categories and respond to consequential regions (Tenenbaum & Griffiths, 2001). For example, Hélie, Waldschmidt, and Ashby (2010) had participants learn two categories of sine-wave gratings defined by a disjunctive rule that included three perceptually distinct regions: gratings with wide or narrow bars were in category A, whereas gratings with bars of medium width were in category B. In this case, CARM would include three rule-sensitive units in the PFC – one for wide bars, one for medium bars, and one for narrow bars. However, because there are only two categories and therefore only two consequential regions, only two rule-sensitive units would be included in the PMC, one for category A and one for category B. Likewise, consider a conjunction rule of the type "respond A if the stimulus has a large value on dimensions 1 and 2; otherwise respond B" (e.g., Hélie & Cousineau, 2015). In this case, CARM would include four rule-sensitive units in the PFC – one for small values on dimension 1, one for large values, one for small values on dimension 2, and one for large values. In contrast, PMC would include only two rule-sensitive neurons – one for category A and one for category B. In other words, CARM assumes PFC representations are truly rule-based, whereas the PMC representations are category-based.

Although this hypothesis about differences between rule-sensitive neurons in PFC and PMC is speculative, it is consistent with current data and theory. First, Hélie, Waldschmidt, and Ashby (2010) showed that with a common set of stimuli, disjunctive categorization rules take longer to learn than one-dimensional rules. Second, Hélie, Roeder, and Ashby (2010) showed important differences in PMC BOLD signals after 20 sessions of training for disjunctive and one-dimensional categorization rules. Third, Hélie, Shamloo, and Ell (2020) tested the ability of participants to compositionally join categories that have already been learned. The results showed that joining categories that are perceptually similar is easier, which CARM predicts is because perceptually similar categories require fewer rule-sensitive neurons in the PFC. Finally, the proposed framework suggests that rules are initially more sensitive to perceptual similarity and gradually become more sensitive to consequential similarity. This is consistent with the proposal of Tenenbaum and Griffiths (2001) relating Bayesian inference and generalization. Future work should be devoted to designing experiments that directly test these predictions and fit the model to the resulting data.

Finally, we should return to the first example considered in this article, which described how children initially learn to add single-digit numbers by applying a counting rule, whereas adults produce the correct sum automatically (or nearly automatically). How would CARM account for the automatization of more complex rules such as this? One complication is that with mental arithmetic, there is no automatized behavior because the same sum could be expressed orally, in writing, via typing, or only in thought. CARM is a theory of how rule-guided *behaviors* become automatized, so some revisions would be needed to account for the automaticity of mental arithmetic.[8] Even so, we hypothesize that similar processes would be in play, with the primary exception that the analogue of the CARM PMC rule units would likely not be in PMC. One candidate is the intraparietal sulcus (e.g., Dehaene, Dehaene-Lambertz, & Cohen, 1998; Venkatraman, Ansari, & Chee, 2005). Similarly, rather than relying on visual input, the representation of the summands in a problem such as "3 + 2 =" might also be in the intraparietal sulcus. Wherever these input and output units are, however, CARM predicts that activation of the "3" and "2" input units in a problem such as "3 + 2 =" would automatically activate the output unit representing "5" after sufficient training. The critical prediction of CARM is that during initial learning, a counting rule mediated in PFC would activate the "5" unit on "3 + 2" trials, causing more activation in the "5" output unit than, for example, in the "3" or "6" units, which would cause Hebbian learning to strengthen the synapse between the "3 + 2" input units and the "5" output unit enough so that eventually the PFC is no longer needed to produce the correct sum. Computationally, the model would be almost identical to the version of CARM proposed here. The PFC rule units would operate in a similar (but more complex) way, but the neuroanatomical location of the PMC rule units and of the visual input would likely differ. Generalizing CARM to these more complex rules should be a goal of future research.

[8] This is largely because most neuroscience studies that generate data about neural changes that occur as automaticity develops use non-human animals (e.g., as in Wallis & Miller, 2003).

Conclusions

This article proposed a new theory of the neural changes that occur as rule-guided behaviors become automatized. The theory was instantiated as a biologically-detailed computational model that makes predictions about behavior at the highest level, and single-neuron firing data at the lowest level. The theory proposes that initially, rule-guided behaviors are controlled by a distributed neural network centered in the prefrontal cortex, and that in addition to initiating behavior, this network also trains a faster and more direct network that includes projections from sensory association cortex directly to rule-sensitive neurons in premotor cortex. After much practice, the direct network is sufficient to control the behavior, without prefrontal involvement. The model successfully accounts for a variety of empirical phenomena that are problematic for other models of automaticity.

References

Amos, A. (2000). A computational model of information processing in the frontal cortex and basal ganglia. *Journal of Cognitive Neuroscience*, *12*(3), 505–519.

Asaad, W. F., Rainer, G., & Miller, E. K. (2000). Task-specific neural activity in the primate prefrontal cortex. *Journal of Neurophysiology*, *84*(1), 451–459.

Ashby, F. G. (2018). Computational cognitive neuroscience. In W. Batchelder, H. Colonius, E. Dzhafarov, & J. Myung (Eds.), *New handbook of mathematical psychology, volume 2* (pp. 223–270). New York: New York: Cambridge University Press.

Ashby, F. G. (2019). *Statistical analysis of fMRI data, Second edition.* Cambridge, MA: MIT Press.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442-481.

Ashby, F. G., & Crossley, M. J. (2012). Automaticity and multiple memory systems. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(3), 363–376.

Ashby, F. G., Ell, S. W., Valentin, V. V., & Casale, M. B. (2005). FROST: A distributed neurocomputational model of working memory maintenance. *Journal of Cognitive Neuroscience*, *17*(11), 1728–1743.

Ashby, F. G., Ell, S. W., & Waldron, E. M. (2003). Procedural learning in perceptual categorization. *Memory & Cognition*, *31*(7), 1114-1125.

Ashby, F. G., & Ennis, J. M. (2006). The role of the basal ganglia in category learning. *Psychology of Learning and Motivation*, *46*, 1-36.

Ashby, F. G., Ennis, J. M., & Spiering, B. J. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review*, *114*(3), 632-656.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 33-53.

Ashby, F. G., & Helie, S. (2011). A tutorial on computational cognitive neuroscience: Modeling the neurodynamics of cognition. *Journal of Mathematical Psychology*, *55*(4), 273-289.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149-178.

Ashby, F. G., & Maddox, W. T. (2010). Human category learning 2.0. *Annals of the New York Academy of Sciences*, *1224*, 147-161.

Ashby, F. G., Paul, E. J., & Maddox, W. T. (2011). COVIS. In E. M. Pothos & A. Wills (Eds.), *Formal approaches in categorization* (pp. 65–87). New York: Cambridge University Press.

Ashby, F. G., Turner, B. O., & Horvitz, J. C. (2010). Cortical and basal ganglia contributions to habit learning and automaticity. *Trends in Cognitive Sciences*, *14*, 208–215.

Ashby, F. G., & Valentin, V. V. (2017). Multiple systems of perceptual category learning: Theory and cognitive tests. In *Handbook of categorization in cognitive science* (pp. 157–188). Elsevier.

Ashby, F. G., & Valentin, V. V. (2018). The categorization experiment: Experimental design and data analysis. In E. J. Wagenmakers & J. T. Wixted (Eds.), *Stevens' handbook of experimental psychology and cognitive neuroscience, Fourth edition, Volume five: Methodology* (Vol. 5, pp. 1–41). New York: Wiley.

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review*, *6*(3), 363-378.

Asmus, F., Huber, H., Gasser, T., & Schöls, L. (2008). Kick and rush: Paradoxical kinesia in parkinson disease. *Neurology*, *71*(9), 695.

Badre, D., Kayser, A. S., & D'Esposito, M. (2010). Frontal cortex and the discovery of abstract action rules. *Neuron*, *66*(2), 315-326.

Buhmann, M. D. (2003). *Radial basis functions: Theory and implementations* (Vol. 12). Cambridge, MA: Cambridge University Press.

Bunge, S. A., & Zelazo, P. D. (2006). A brain-based account of the development of rule use in childhood. *Current Directions in Psychological Science*, *15*(3), 118-121.

Cantwell, G., Crossley, M. J., & Ashby, F. G. (2015). Multiple stages of learning in perceptual categorization: Evidence and neurocomputational theory. *Psychonomic Bulletin & Review*, *22*, 1598–1613.

Christoff, K., Keramatian, K., Gordon, A. M., Smith, R., & Mädler, B. (2009). Prefrontal organization of cognitive control according to levels of abstraction. *Brain Research*, *1286*, 94-105.

Cohen, J. R., & Poldrack, R. A. (2008). Automaticity in motor sequence learning does not impair response inhibition. *Psychonomic Bulletin & Review*, *15*(1), 108–115.

Connors, B. W., Gutnick, M. J., & Prince, D. A. (1982). Electrophysiological properties of neocortical neurons in vitro. *Journal of Neurophysiology*, *48*(6), 1302–1320.

Crossley, M. J., & Ashby, F. G. (2015). Procedural learning during declarative control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(5), 1388–1403.

Crossley, M. J., Paul, E. J., Roeder, J. L., & Ashby, F. G. (2016). Declarative strategies persist under increased cognitive load. *Psychonomic Bulletin & Review*, *23*(1), 213–222.

Crossman, E. R. F. W. (1959). A theory of the acquisition of speed-skill. *Ergonomics*, *2*(2), 153-166.

Davis, T., Goldwater, M., & Giron, J. (2017). From concrete examples to abstract relations: The rostrolateral prefrontal cortex integrates novel examples into relational categories. *Cerebral Cortex*, *27*(4), 2652–2670.

Davis, T., Love, B. C., & Preston, A. R. (2011). Learning the exception to the rule: Model-based fMRI reveals specialized representations for surprising category members. *Cerebral Cortex*, *22*(2), 260–273.

Dégenètais, E., Thierry, A.-M., Glowinski, J., & Gioanni, Y. (2002). Electrophysiological properties of pyramidal neurons in the rat prefrontal cortex: An in vivo intracellular recording study. *Cerebral Cortex*, *12*(1), 1–16.

Dehaene, S., Dehaene-Lambertz, G., & Cohen, L. (1998). Abstract representations of numbers in the animal and human brain. *Trends in Neurosciences*, *21*(8), 355–361.

Desmurget, M., & Turner, R. S. (2010). Motor sequences and the basal ganglia: Kinematics, not habits. *Journal of Neuroscience*, *30*(22), 7685–7690.

Doya, K. (2007). Reinforcement learning: Computational theory and biological mechanisms. *HFSP Journal*, *1*, 30–40.

Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 840–859.

Durstewitz, D., Vittoz, N. M., Floresco, S. B., & Seamans, J. K. (2010). Abrupt transitions between prefrontal neural ensemble states accompany behavioral transitions during rule learning. *Neuron*, *66*(3), 438–448.

Feenstra, M. G., & Botterblom, M. H. (1996). Rapid sampling of extracellular dopamine in the rat prefrontal cortex during food consumption, handling and exposure to novelty. *Brain Research*, *742*(1-2), 17–24.

Feldman, D. E. (2009). Synaptic mechanisms for plasticity in neocortex. *Annual Review of Neuroscience*, *32*, 33–55.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *Journal of Neuroscience*, *23*(12), 5235–5246.

Groen, G. J., & Parkman, J. M. (1972). A chronometric analysis of simple addition. *Psychological Review*, *79*(4), 329–343.

Heaton, R. K. (1981). *A manual for the Wisconsin Card Sorting Test*. Odessa, FL: Psychological Assessment Resources.

Hélie, S., & Cousineau, D. (2015). Differential effect of visual masking in perceptual categorization. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(3), 816–825.

Hélie, S., Ell, S. W., & Ashby, F. G. (2015). Learning robust cortico-cortical associations with the basal ganglia: An integrative review. *Cortex*, *64*, 123-135.

Hélie, S., Roeder, J. L., & Ashby, F. G. (2010). Evidence for cortical automaticity in rule-based categorization. *Journal of Neuroscience*, *30*(42), 14225-14234.

Hélie, S., Roeder, J. L., Vucovich, L., Rünger, D., & Ashby, F. G. (2015). A neurocomputational model of automatic sequence production. *Journal of Cognitive Neuroscience*, *27*, 1412–1426.

Hélie, S., Shamloo, F., & Ell, S. W. (2020). The impact of training methodology and category structure on the formation of new categories from existing knowledge. *Psychological Research*, *84*(4), 990–1005.

Hélie, S., Waldschmidt, J. G., & Ashby, F. G. (2010). Automaticity in rule-based and information-integration categorization. *Attention, Perception, & Psychophysics*, *72*(4), 1013-1031.

Hoshi, E., Shima, K., & Tanji, J. (2000). Neuronal activity in the primate prefrontal cortex in the process of motor selection based on two behavioral rules. *Journal of Neurophysiology*, *83*(4), 2355–2373.

Izhikevich, E. M. (2003). Simple model of spiking neurons. *IEEE Transactions on Neural Networks*, *14*(6), 1569–1572.

Joel, D., Weiner, I., & Feldon, J. (1997). Electrolytic lesions of the medial prefrontal cortex in rats disrupt performance on an analog of the Wisconsin Card Sorting Test, but do not disrupt latent inhibition: Implications for animal models of schizophrenia. *Behavioural Brain Research*, *85*(2), 187–201.

Kimberg, D. Y., D'Esposito, M., & Farah, M. J. (1997). Effects of bromocriptine on human subjects depend on working memory capacity. *Neuroreport*, *8*(16), 3581–3585.

Konishi, S., Kawazu, M., Uchida, I., Kikyo, H., Asakura, I., & Miyashita, Y. (1999). Contribution of working memory to transient activation in human inferior prefrontal cortex during performance of the Wisconsin Card Sorting Test. *Cerebral Cortex*, *9*(7), 745–753.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 1–28.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22-44.

LeFevre, J.-A., Sadesky, G. S., & Bisanz, J. (1996). Selection of procedures in mental addition: Reassessing the problem size effect in adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 216–230.

Logan, G. D. (1982). On the ability to inhibit complex movements: A stop-signal study of typewriting. *Journal of Experimental Psychology: Human Perception and Performance*, *8*(6), 778–792.

Logan, G. D. (1985). Skill and automaticity: Relations, implications, and future directions. *Canadian Journal of Psychology*, *39*(2), 367–386.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*(4), 492–527.

Long, J. (1976). Visual feedback and skilled keying: Differential effects of masking the printed copy and the keyboard. *Ergonomics*, *19*(1), 93–110.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309-332.

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 650-662.

Maddox, W. T., Bohil, C. J., & Ing, A. D. (2004). Evidence for a procedural-learning-based system in perceptual category learning. *Psychonomic Bulletin & Review*, *11*(5), 945-952.

Maddox, W. T., & Ing, A. D. (2005). Delayed feedback disrupts the procedural-learning system but not the hypothesis testing system in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 100-107.

Monchi, O., Petrides, M., Petre, V., Worsley, K., & Dagher, A. (2001). Wisconsin Card Sorting revisited: Distinct neural circuits participating in different stages of the task identified by event-related functional magnetic resonance imaging. *Journal of Neuroscience*, *21*(19), 7733–7741.

Muhammad, R., Wallis, J. D., & Miller, E. K. (2006). A comparison of abstract rules in the prefrontal cortex, premotor cortex, inferior temporal cortex, and striatum. *Journal of Cognitive Neuroscience*, *18*(6), 974-989.

Nomura, E., Maddox, W., Filoteo, J., Ing, A., Gitelman, D., Parrish, T., ... Reber, P. (2007). Neural correlates of rule-based and information-integration visual category learning. *Cerebral Cortex*, *17*(1), 37-43.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39-57.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*(2), 266-300.

Poldrack, R. A., Sabb, F. W., Foerde, K., Tom, S. M., Asarnow, R. F., Bookheimer, S. Y., & Knowlton, B. J. (2005). The neural correlates of motor skill automaticity. *Journal of Neuroscience*, *25*(22), 5356–5364.

Rabbitt, P. (1978). Detection of errors by skilled typists. *Ergonomics*, *21*(11), 945–958.

Rall, W. (1967). Distinguishing theoretical synaptic potentials computed for different soma-dendritic distributions of synaptic input. *Journal of Neurophysiology*, *30*(5), 1138-1168.

Reber, P. J., Gitelman, D. R., Parrish, T. B., & Mesulam, M. (2003). Dissociating explicit and implicit category knowledge with fMRI. *Journal of Cognitive Neuroscience*, *15*(4), 574–583.

Rickard, T. C. (1997). Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General*, *126*(3), 288–311.

Roeder, J. L., & Ashby, F. G. (2016). What is automatized during perceptual categorization? *Cognition*, *154*, 22–33.

Rogers, R. L., Andrews, T. K., Grasby, P., Brooks, D., & Robbins, T. (2000). Contrasting cortical and subcortical activations produced by attentional-set shifting and reversal learning in humans. *Journal of Cognitive Neuroscience*, *12*(1), 142–162.

Ross, M., Chartier, S., & Hélie, S. (2017). The neurodynamics of categorization: Critical challenges and proposed solutions. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science. 2nd edition* (pp. 1053–1076). Oxford: Elsevier.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(1), 1-66.

Seger, C. A., & Miller, E. K. (2010). Category learning in the brain. *Annual Review of Neuroscience*, *33*, 203-219.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127–190.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411–1436.

Soliveri, P., Brown, R. G., Jahanshahi, M., Caraceni, T., & Marsden, C. D. (1997). Learning manual pursuit tracking skills in patients with Parkinson's disease. *Brain: A Journal of Neurology*, *120*(8), 1325–1337.

Soto, F. A., Waldschmidt, J. G., Helie, S., & Ashby, F. G. (2013). Brain activity across the development of automatic categorization: A comparison of categorization tasks using multi-voxel pattern analysis. *Neuroimage*, *71*, 284–897.

Spiering, B. J., & Ashby, F. G. (2008). Response processes in information–integration category learning. *Neurobiology of Learning and Memory*, *90*(2), 330-338.

Sternberg, S., Monsell, S., Knoll, R. L., & Wright, C. E. (1978). The latency and duration of rapid movement sequences: Comparisons of speech and typewriting. In G. E. Stelmach (Ed.), *Information processing in motor control and learning* (pp. 117–152). New York: Academic Press.

Strange, B., Henson, R., Friston, K., & Dolan, R. J. (2001). Anterior prefrontal cortex mediates rule learning in humans. *Cerebral Cortex*, *11*(11), 1040–1046.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–640.

Thomas-Ollivier, V., Reymann, J., Le Moal, S., Schück, S., Lieury, A., & Allain, H. (1999). Procedural memory in recent-onset Parkinson's disease. *Dementia and Geriatric Dognitive Disorders*, *10*(2), 172–180.

Vallentin, D., Bongard, S., & Nieder, A. (2012). Numerical rule coding in the prefrontal, premotor, and posterior parietal cortices of macaques. *Journal of Neuroscience*, *32*(19), 6621–6630.

Varrone, A., & Halldin, C. (2014). Human brain imaging of dopamine transporters. In P. Seeman & B. Madras (Eds.), *Imaging of the human brain in health and disease* (pp. 203–240). Amsterdam: Elsevier.

Venkatraman, V., Ansari, D., & Chee, M. W. (2005). Neural correlates of symbolic and non-symbolic arithmetic. *Neuropsychologia*, *43*(5), 744–753.

Waldron, E. M., & Ashby, F. G. (2001). The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. *Psychonomic Bulletin & Review*, *8*(1), 168-176.

Waldschmidt, J. G., & Ashby, F. G. (2011). Cortical and striatal contributions to automaticity in information-integration categorization. *Neuroimage*, *56*(3), 1791-1802.

Wallis, J. D., & Miller, E. K. (2003). From rule to response: Neuronal processes in the premotor and prefrontal cortex. *Journal of Neurophysiology*, *90*(3), 1790-1806.

White, I. M., & Wise, S. P. (1999). Rule-dependent neuronal activity in the prefrontal cortex. *Experimental Brain Research*, *126*(3), 315–335.

Zeithamova, D., & Maddox, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition*, *34*(2), 387-398.

Appendix: Modeling Details

The numerical values of the parameters used in all simulations are listed in Tables 1 and 2.

Table 1
General Model Parameters

Parameter	Value
Equation 1	
ω	0.8
Equation 6	
$W_{PFC \rightarrow PMC}$	9
$W_{VC \rightarrow PMC}$	.08
Equation 7	
$W_{PMC \rightarrow Motor}$	100
Equation 9	
α_w	1×10^{-8}
θ_{NMDA}	400
W_{max}	5
COVIS	
Δ_C	1
Δ_E	1
γ	1
λ	1
Decision Points	
Motor cortex response threshold	700

Note. Weights on all modifiable synapses are initial weights.

Electrophysiological Simulations

In simulations of the Wallis and Miller (2003) experiment, CARM included separate *same* and *different* rule units in PFC and PMC. Each PFC rule unit included two neurons – one that responded to low values of similarity and one that responded to high values. The resulting model was trained on 300 initial baseline trials, 11,520 training trials, and 300 post-training test trials. During each phase, half the trials were *same-rule* trials and half were *different-rule* trials.

Because of noise, successive runs of the model under the same conditions yield variable results. For this reason, the results in Figure 3 are based on 20 independent model simulations. For each of these runs, we recorded the time until the first spike on all baseline (i.e., pre-automaticity) and test (i.e., post-automaticity) trials. The probability density functions shown in Figure 3 were then estimated using the MATLAB kernel density estimator algorithm (ksdensity).

As described in the text, the stimuli were 12 grayscale photographs recorded at a resolution of 300 × 300 pixels. On each *same* trial, two copies of the same randomly selected photograph were presented to CARM (with independent noise added to each) and on *different* trials, two randomly selected different photographs were presented (again with independent noise added to each). On each trial, the similarity and dissimilarity of the two photographs were computed, where similarity was defined as the pixel-by-pixel correlation in grayscale values, and dissimilarity was computed as one minus this value. The model included 100 units (assumed to reside in some area of visual association cortex) that were

Table 2

Application Specific Parameters

Parameter	Value
Electrophysiological	
α_w	1×10^{-10}
Threshold for PFC activation	400
Threshold for PMC activation	400
noise standard deviation	3
θ_{NMDA}	300
Dual Task	
α_w	1×10^{-9}
Amplitude of visual input from dual-task units	250
$W_{PFC_{STROOP} \to PFC_{CAT}}$	50
All other FROST weights	1
Button Switch	
α_w in PMC	1×10^{-9}
α_w in motor cortex	2.45×10^{-8}
$W_{PMC \to MC}$	1
W_{max} (PMC-MC Synapses)	10
θ_{NMDA}	450
Figure 8	
ω	0.5
α_w	1×10^{-10}
W_{max}	100
$W_{PFC \to PMC}$	50
$W_{PMC \to Motor}$	.01

Note. Weights on all modifiable synapses are initial weights.

sensitive to perceptual similarity, and 100 units that were sensitive to perceptual dissimilarity. Each unit was tuned to a preferred value that ranged from .01 to 1 in units of .01 (so .01, .02, .03, etc.), and we modeled the tuning curve of each unit using radial basis functions. The high-similarity neuron in the PFC *same-rule* unit received input from similarity-sensitive visual units that were tuned to high similarities and the low-similarity neuron received input from similarity-sensitive visual units that were tuned to low similarities. Similarly, the high-dissimilarity neuron in the PFC *different-rule* unit received input from dissimilarity-sensitive visual units that were tuned to high dissimilarities and the low-dissimilarity neuron received input from dissimilarity-sensitive visual units that were tuned to low dissimilarities.

Dual-Task Simulations

As described in the text, we simulated dual-task experiments by augmenting CARM with the COVIS model of initial rule learning and the FROST model of working memory maintenance (see Figure 5). Specifically, the model included three active PFC working-memory units on each trial – one that maintained the current categorization rule (containing two neurons), and two units that maintained representations of the two dual-task numbers. Each dual-task unit received its own (square-wave) visual input that was on as long as the corresponding number was visually present. All PFC working memory units laterally inhibited each other. Because the delay between presentation of the dual-task stimuli and the categorization response was typically less than a second, the reverberating loops in FROST that maintain PFC activation during delay periods (e.g., of 30 sec or longer) played little or no functional role in our simulations. Instead, the model predicts that the primary effect of the dual task is to increase lateral inhibition on the PFC categorization rule unit. We modeled all FROST units using the same (Izhikevich) regular-spiking neuron model as in CARM. The synaptic weights are listed in Table 2. For more details see Ashby et al. (2005). The rule-learning model of COVIS was only used to select which categorization rule was active on each trial.

We modeled the Zeithamova and Maddox (2006) and Hélie, Waldschmidt, and Ashby (2010) experiments using exactly the same version of CARM, since the two studies used the same stimuli and the same dual task. The only difference was in the different category structures used in the two experiments (shown in Figure 4), and in the different amounts of training each participant received. On each simulated trial, one randomly selected stimulus from a randomly selected category was presented to CARM. To model perceptual noise, the coordinates of the stimulus (in the 100 × 100 space shown in Figure 4) were randomly perturbed by adding noise sampled from a normal distribution (with mean 0 and standard deviation 6) to the position of the visual stimulus along each dimension.

To simulate the Zeithamova and Maddox (2006) experiment, we sampled stimuli from the same multivariate normal distributions used by Zeithamova and Maddox (2006). Since theirs was a one-session experiment, CARM was run through 400 trials a total of 20 separate times and the results were averaged across the 20 simulations. The control condition (no dual task) was simulated in exactly the same way, except no dual-task working memory units were active. The results are shown in Figure 6.

The Hélie, Waldschmidt, and Ashby (2010) experiment was simulated with exactly the same model, except using the same uniformly distributed categories as Hélie, Waldschmidt, and Ashby (2010) and for the same extended training used in that study. Specifically, we ran 20 identical simulations of 12,120 trials each. The dual task was in effect only on trials 11,521 – 12,120. On each run, we calculated the mean accuracy on the trials that corresponded to the last training session in the Hélie, Waldschmidt, and Ashby (2010) experiment (trials 10,921 – 11,520) and the mean accuracy of the dual-task session (trials 11,521 – 12,120). The results are shown in Figure 7.

To calculate the PFC contribution to PMC activation that is shown in Figure 8, on each trial we separately computed the total PMC activation and the total PMC activation coming from PFC by numerically integrating over the duration of the trial (i.e., beginning with stimulus onset and ending when activation in the controlling PMC unit crossed the response threshold). Each "proportion from PFC" value plotted in Figure 8 is the ratio of these two values on the indicated trial.

Button-Switch Simulations

The exact same version of CARM was used to model the early-learning and late-learning effects of a button switch. The only difference was the extra Hebbian learning that occurred as a result of the extended training in the late-learning simulations. To estimate the effects on accuracy of a button switch after initial learning, we simulated 20 independent runs of 600 trials each followed by 100 button-switch trials. We computed mean accuracy during trials 501-600 for each simulation, and averaged all 20 means to estimate pre-button switch accuracy. We also computed the mean accuracy over trials 601-700 for each simulation, and then averaged all 20 of these means to estimate post-button switch accuracy. The results are shown in Figure 9.

To simulate the effects of a button switch after extended training, we simulated 20 independent runs of 11,520 trials each, followed by 600 post-button-switch trials (as in the Hélie, Waldschmidt, & Ashby, 2010 experiment). For each run, we split the 600 post-button-switch trials into twelve 50-trial blocks and calculated an average accuracy for each block. We then calculated an average block accuracy across all 20 runs of the model. The results are shown in Figure 10.