University of California Santa Barbara

Cortex Automatizes Rules Model: A Novel Neurocomputational Model of Rule Based Automaticity

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor in Philosophy

in

Psychological and Brain Sciences

by

Paul H. Kovacs

Committee in Charge:

Professor Gregory Ashby, Committee Chair

Professor Miguel Eckstein

Professor Scott Grafton

Professor Michael Miller

June 2022

The Dissertation of Paul H. Kovacs is approved.

Professor Miguel Eckstein

Professor Michael Miller

Professor Scott Grafton

Professor Greg Ashby, Committee Chair

June 2022

Abstract

Cortex Automatizes Rules Model: A Novel Neurocomputational Model of Rule Based Automaticity

by

Paul H. Kovacs

This dissertation 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.

The dissertation also presents the results from two experiments investigating the nature of what is automatized after lengthy practice with a rule-guided behavior. The results of both experiments suggest that an abstract rule, if interpreted as a verbal-based strategy, was not automatized during training, but rather the automatization linked a set of stimuli with similar values on one visual dimension to a common motor response. The experiments were designed to test the Cortex Automatizes Rules Model. The present results support this model and suggest that the projections from visual cortex to prefrontal and premotor cortex are restricted to visual representations of the relevant stimulus dimension only.

Curriculum Vitae

Paul H. Kovacs
1709 San Pascual St.
Santa Barbara
CA, 93101
805-570-2026
p88kovacs@gmail.com

Summary Statement

I am a graduate student in the UCSB Psychological Brain Sciences department working in Greg Ashby's "Computational Cognitive Neuroscience Lab". My main research interest is in the neuroscientific basis of automatic behavior.

Education

2003 to 2007 B.A. In Philosophy University of California, Santa Cruz

2014 to 2016 B.A. in Biology with a focus on Neurobiology College of Creative Studies, UCSB

2016 to 2022 PhD student in the Psychological and Brain Sciences department at UCSB Computational Cognitive Neuroscience

Dissertation Title: "Cortex Automatizes Rules Model: A Novel Neurocomputational Model of Rule Based Automaticity"

Dissertation Committee
Dr. Greg Ashby (chair)
Dr. Miguel Eckstein
Dr. Stan Grafton
Dr. Michael Miller

Teaching Experience (as TA)

Spring 2022 - Laboratory in Biopsychology (with Dr. Karen Szumlinski)

Winter 2022 - Retinal Development (with Dr. Ben Reese)

Fall 2021 - Introduction to Biopsychology (with Dr. Samantha Scudder)

Summer 2021 - The Psychology of Self (with Dr. Stan Klein)

Spring 2021 - FMRI Analysis (with Dr. Greg Ashby)

Winter 2021 - Introduction to Biopsychology (with Dr. Samantha Scudder)

Fall 2020 - Introduction to Psychology (with Dr. Tamsin German)

Spring 2020 - Learning and Motivation (with Dr. Ron Keiflin)

Winter 2020 - Complex Systems (with Dr. Skirmantas Janusonis)

Winter 2019 - Brain Cell Analysis (with Dr. Skirmantas Janusonis)

Summer 2019 - Statistical Methods (with Anudhi Munasinghe)

Fall 2018 - Introductory Statistics (with Dr. Erin Horowitz)

Summer 2018 - Statistical Methods (with Dr. Jason Anderson)

Winter 2018 - Introductory Statistics (with Dr. Jonna LaJoy)

Fall 2017 - Neuroantomy Lab (with Dr. Ben Reese)

Summer 2017 - Cognitive Neuroscience (with Dr. Allison Shapiro)

Summer 2017 - Neuropharmacology (with Dr. Adam Klein)

Summer 2017 - Psychopharmacology of Therapeutic Drugs (Dr. Vanessa Woods)

Spring 2017 - Introductory Statistics (with Dr. Jeff Bowen)

Fall 2016 - Introduction to Psychology (with Dr. Alan Friedlund)

Conferences/Talks

Spring 2022 - Presented the CARM model and the results from two experiments at UCSB Psychological and Brain Sciences CPCN Seminar

Summer 2018 - Presented my CARM (Cortex Automatizes Rules) Model of Rule-Based Automaticity at Math Psych. The talk included a presentation of the structure of the model as well as data modeling work with CARM.

Poster Presentations

Spring 2015 - URCA Poster Presentation at the UCSB Undergraduate Research Colloquium on a project "using a two-way choice behavioral screen to identify ligands of the Drosophila gustatory receptor GR89a"

Fall 2015 - SURF Poster Presentation at the College of Creative Studies Science Week on a project "using a two-way choice behavioral screen to identify ligands of the Drosophila gustatory receptor GR89a"

Research Experience

June 2016 - June 2022 Graduate Researcher in Dr. Gregory Ashby's Computational Cognitive Neuroscience Lab. Looking at automatic behavior in the context of category learning

September 2015 - June 2016. Undergraduate Research Assistant in Dr. Tod Kippin's Lab Investigating the neurobiological basis of Cocaine and Methamphetamine addiction in rats.

Spring 2014 - September 2015 Undergraduate Researcher in Dr. Craig Montell's Lab. Investigating the function of Gustatory Receptors in the fruit fly.

Summer 2013. INSET Summer Internship through the Center for Nanotechnology in Society looking at the societal implications of nanotechnology including regulatory challenges and safety concerns.

Grants/Fellowships

Awarded Fall of 2014: URCA Grant for a project using a two-way choice behavioral screen to find the ligands of the Drosophila gustatory receptor GR89a.

Awarded Summer SURF Fellowship for a project investigating the role of the gustatory receptor GR77a in male Drosophila reproduction.

Publications

Kovacs, P., Hélie, S., Tran, A. N., Ashby, F. G. (2021). A neurocomputational theory of how rule-guided behaviors become automatic. *Psychological review.*

Kovacs, P., Ashby, F. G. (2022). On what it means to automatize a rule. *Cognition*, 226, 105168.

Contents

Chapter 1

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 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 dissertation aims at filling this gap in the literature. Specifically, I 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 I show that this model successfully accounts for a variety of single-unit recording and behavioral phenomena that characterize automatic rule-guided behavior.

By rule, I 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, 1988; 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 I believe that this 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 activation 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

[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.

with activity in cortical areas (i.e., preSMA and SMA).

1.1 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 1.1. 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

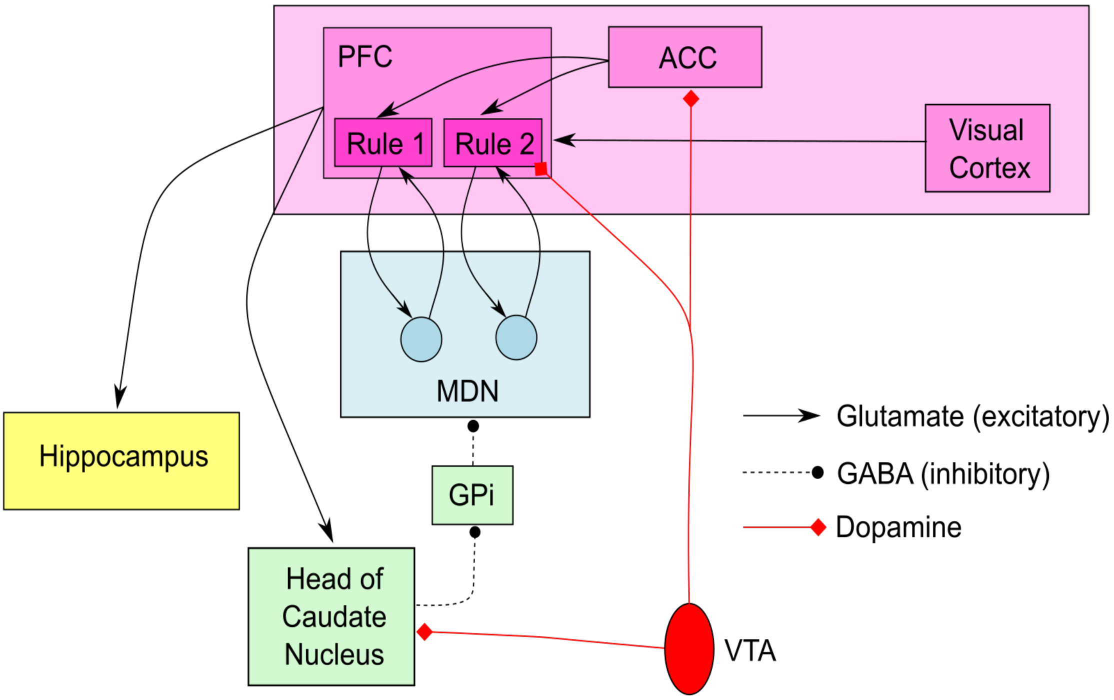

Figure 1.1: The COVIS rule-learning system. PFC = prefrontal cortex, ACC = anterior cingulate cortex, MDN = medial dorsal nucleus of the thalamus, GPi = internal segment of the globus pallidus, VTA = ventral tegmental area. Figure reprinted from "A neurocomputational theory of how rule-guided behaviors become automatic" by Kovacs, P., Hélie, S., Tran, A. N., & Ashby, F. G., *Psychological Review*. Copyright 2021 with permisssion from the American Psychological Association.

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 1.1 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 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).

1.2 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. Roeder and Ashby interpret these results to suggest that rules are automatized in RB tasks, whereas SR associations are automatized in II tasks (2016). However, this dissertation presents results from a new experiment that challenges this interpretation.

1.3 Overview of Dissertation

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 dissertation proposes such a theory and presents modeling and experimental results intended to test the theory.

In Chapter 2 I propose a new theory that describes 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, I formulate it as a biologically-detailed neurocomputational network of spiking neurons. Finally, I show that the resulting model successfully accounts for single-unit recording and behavioral data that are problematic for other accounts of automaticity.

In Chapter 3 I present an experiment that falls naturally out of the struture of CARM. In this experiment subjects are automatized on a set training stimuli in a rule based task, and then tested on transfer stimuli that differ on either relevant dimension stimulus values or irrelevant dimension values. If stimuli are represented as two dimensional gestalts then the model predicts that automaticity will be lost in both conditions, however if stimuli are represented based only on relevant dimension values, then automaticity will be lost only in the relevant dimension transfer condition. I observed the later which suggests that the perceptual stimuli being automatized are relevant dimension representations and not multi-dimensional gestalts.

In Chapter 4 I present an experiment that tests a novel prediction of the new theory. The experiment in chapter 4 is a modification of an experiment by Roeder and Ashby described previously (2016). In this experiment Roeder and Ashby observed no interference on incongruent stimuli and on this basis concluded that abstract rules are automatized in rule based tasks, not stimulus response associations. The experiment presented in Chapter 4 has the same structure except the primary disjunctive rule task on primary days and simple 1D rule task on secondary days both had the same relevant dimension on bar width. In this experiment it was observed that when primary and secondary sessions used rules with the same relevant dimension, there was a significant interference on incongruent stimuli. I interpret this result as supporting evidence that rule based automaticity forms stimulus response associations between one dimensional perceptual representations and behavioral responses.

In Chapter 5 I discuss the implications of the results from the experiments presented in chapters 3 and 4.

1.4 Permissions and Atributions

Chapters 1-5 draw heavily from the following publications:

Kovacs, P., Hélie, S., Tran, A. N., Ashby, F. G. (2021). A neurocomputational theory of how rule-guided behaviors become automatic. *Psychological review.*

Kovacs, P., Ashby F. G. (In Press). On what it means to automatize a rule. *Cognition.*

Chapter 2

Cortex Automatizes Rules Model (CARM)

2.1 Introduction

The literature 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). I propose a similar model for the development of automatic rule-guided behaviors. In particular, I propose that a key function of the rule-learning network illustrated in Figure 1.1 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 I 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 1.1 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.

2.2 Neural Architecture of CARM

The neural architecture of the model, which I call the Cortex Automatizes Rules Model (CARM), is described in Figure 2.1. 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 1.1 and 2.1. For example, in the Figure 1.1 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 2.1 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 2.1 scenario, each application of Rule 1 results in either an A or B response (e.g., a button 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. I assume that rule units develop as a result of life-long practice with a rule. For example, before participating in a laboratory experiment, a

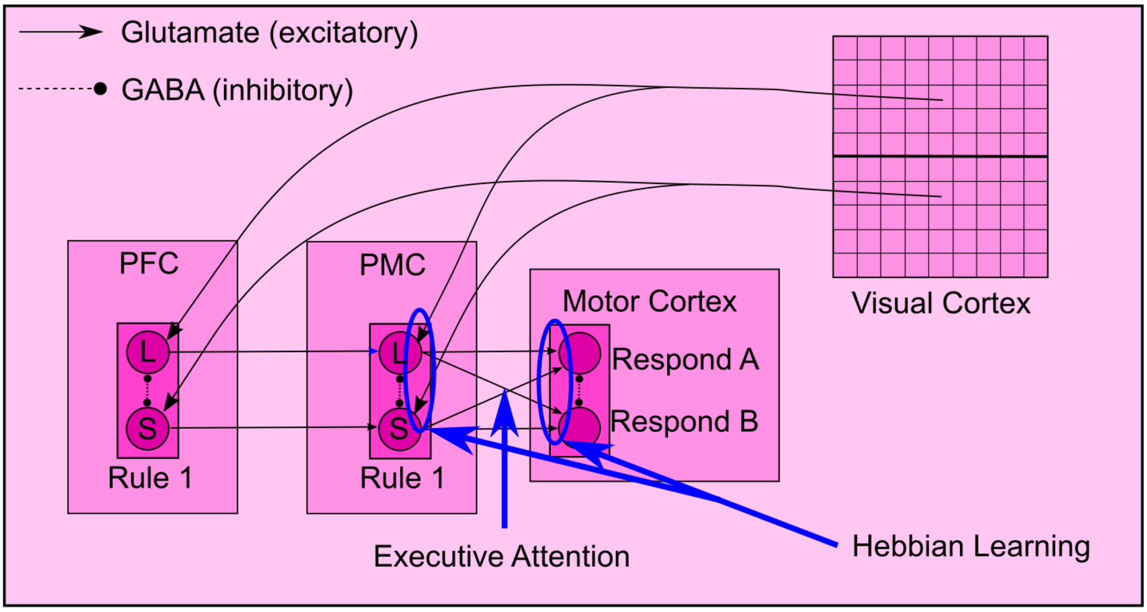

Figure 2.1: The neural architecture of CARM for an application to a one-dimensional categorization task in which the automatized rule is designated as Rule 1. According to this rule, response A is given if the presented stimulus has a large value on the single relevant dimension, and response B is given if the value is small. PFC = prefrontal cortex, PMC = premotor cortex. Figure reprinted from “A neurocomputational theory of how rule-guided behaviors become automatic” by Kovacs, P., Hélie, S., Tran, A. N., & Ashby, F. G., *Psychological Review.* Copyright 2021 with permisssion from the American Psychological Association.

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. I 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.

I 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 I 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, I 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, I 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. I 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.

2.2.1 Visual Cortex

I modeled visual cortex as either a 100×2 (Application 1) or 100×100 (Applications 2 and 3) grid of units. I 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, I 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, I assumed that when a stimulus is presented, the activation in sensory cortical unit K at time t equals

$$A_K(t) = 50 \ \exp\left[-\frac{d(K, stimulus)}{\omega}\right] \tag{2.1}$$

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).

2.2.2 PFC, PMC, and Motor Cortex

I 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

$$\begin{aligned} 100\frac{dV(t)}{dt} &= I(t) + .07[V(t)+60][V(t)+40] - U(t) + \epsilon(t) \\ \frac{dU(t)}{dt} &= -.06[V(t)+60] - .03U(t), \end{aligned} \tag{2.2}$$

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 2.2 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. I 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

$$\alpha(t) = \begin{cases} .05t\exp\left(\frac{20-t}{20},\right) & t > 0 \\ 0, & t < 0 \end{cases} \tag{2.3}$$

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:

$$F(t) = \sum_{i=1}^{N} \alpha(t - t_i). \tag{2.4}$$

Figure 2.1 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, I 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

$$I_{PFC_L}(t) = \left[\sum_{K \in \mathrm{L}} A_K(t)\right] - F_{PFC_S}(t), \tag{2.5}$$

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 2.4 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 2.1). Thus, for example, the input to the PMC-L rule unit was

$$I_{PMC_L}(t) = W_{VC \to PMC}\left[\sum_{\text{all } K} A_K(t)\right] + W_{PFC \to PMC} F_{PFC_L}(t) - F_{PMC_S}(t), \tag{2.6}$$

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

PMC and$W_{PFC \rightarrow PMC}$ 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

$$\begin{aligned} I_{Motor_A}(t) = W_{PMC_L \rightarrow Motor_A} \Phi_{LA} F_{PMC_L}(t) \; &+ \; W_{PMC_S \rightarrow Motor_A} \Phi_{SA} F_{PMC_S}(t) \\ &- \; F_{Motor_B}(t), \end{aligned} \tag{2.7}$$

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. I model the effects of these instructions by setting $\Phi_{LA} = .9$ and $\Phi_{SA} = .1$. Furthermore, I 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).

2.2.3 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, I assumed that plasticity at these synapses depends only on the product of synapse-specific pre- and post-synaptic activation. Specifically, I 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

$$G_{\mathrm{J}}(T) = \int_0^T F_{\mathrm{J}}(t)\mathrm{dt}, \tag{2.8}$$

for J = A or B, and where $F_{\mathrm{J}}(t)$ is as in Equation 2.4 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, I used the following difference equation to adjust the strength of $W_{\mathrm{A,B}}(n)$.

$$\begin{aligned} W_{\mathrm{A,B}}(n+1) &= W_{\mathrm{A,B}}(n) \\ &+ \alpha_W\ G_{\mathrm{A}}(T)\,[G_{\mathrm{B}}(T) - \theta_{\mathrm{NMDA}}]^+\,[W_{\max} - W_{\mathrm{A,B}}(n)] \\ &- \alpha_W\ G_{\mathrm{A}}(T)\,[\theta_{\mathrm{NMDA}} - G_{\mathrm{B}}(T)]^+\,W_{\mathrm{A,B}}(n), \end{aligned} \tag{2.9}$$

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.[1] First, note that this term equals 0 except when total postsynaptic activation is below the NMDA-receptor threshold. The $W_{\text{A,B}}(n)$ at the end prevents synaptic strength from dropping below 0.

Equation 2.9 required some slight modification for the synapses between PMC and motor cortex. I assumed that plasticity at these synapses follows the same Hebbian rules as synapses between visual cortex and PMC. However, note that Equation 2.9 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 2.9 would strengthen both of these synapses because activation in the postsynaptic neuron is 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 2.9, synaptic strengthening will primarily occur only at synapses between visual cortex and the correct PMC rule unit.

On the other hand, Equation 2.9 does not properly adjust the strength of synapses

[1] 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).

between PMC and motor cortex. As shown in Figure 2.1, 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 I 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 2.9 will strengthen both LA and SA. For this reason, I need to replace Equation 2.9 with a Hebbian learning scheme that strengthens LA, but not SA, LB, or SB.

My solution was to remove the postsynaptic term from Equation 2.9 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, I modified synaptic strength as follows.

$$\begin{aligned} W_{PMC_J \to Motor_I}(n+1) = \; & W_{PMC_J \to Motor_I}(n) \\ & + \alpha_w \; [W_{PMC_J \to Motor_I}(n) \; \Phi_{JI} \; G_{\mathrm{A}}(T) - \theta_{\mathrm{NMDA}}]^+ \, [W_{\max} - W_{PMC_J \to Motor_I}(n)] \\ & - \alpha_w \; [\theta_{\mathrm{NMDA}} - W_{PMC_J \to Motor_I}(n) \; \Phi_{JI} \; G_{\mathrm{A}}(T)]^+ \, W_{\mathrm{A,B}}(n). \end{aligned} \tag{2.10}$$

The constant θ_{NMDA} still denotes the threshold for postsynaptic NMDA-receptor activation, but Equation 2.10 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 2.10 only strengthens the LA synapse.

2.2.4 Initial Category Learning

This dissertation 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.[2] Even so, to simulate the entire learning process – from initial rule discovery to automaticity – I 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). I call this augmented model CARM$^+$. A schematic of the neural structures of CARM$^+$, when applied to a dual-task experiment, is shown in Figure 2.4 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

[2] I treat "rule discovery" and "rule learning" as synonyms in this dissertation.

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.

2.3 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 involvement because under these conditions, the pathway from visual cortex to motor cortex is more direct (with fewer synapses; see Figure 2.1).

Many current models predict increases in accuracy and decreases in RT as training progresses, so rather than documenting these well-studied effects, my 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, I 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, I 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, I 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. I know of no other models of automaticity that can account for these phenomena.

2.3.1 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 record-

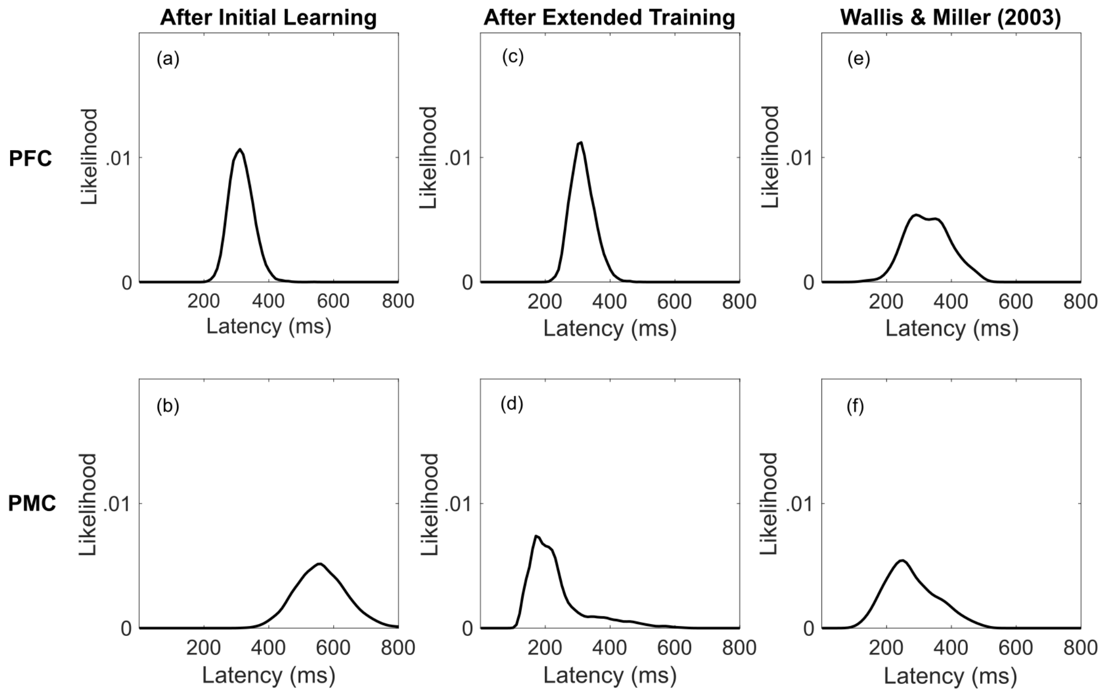

Figure 2.2: Probability density function estimates from rule-selective neurons in PFC (first row) and PMC (second row). Panels a – d show predictions of CARM, and panels e and f show estimates for single neurons that were reported by Wallis and Miller (2003). In the case of CARM, the estimates are the 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. Figure reprinted from "A neurocomputational theory of how rule-guided behaviors become automatic" by Kovacs, P., Hélie, S., Tran, A. N., & Ashby, F. G., *Psychological Review*. Copyright 2021 with permisssion from the American Psychological Association.

ings 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 2.2 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.

I 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[3] 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.[4] I assumed that similarity 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.

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

[4] 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).

Therefore, for convenience, I 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, I 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 I 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, I trained the model using the same experimental procedures as Wallis and Miller. I 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, I set the Hebbian learning rates to 0. On each baseline trial, I recorded the time it took for rule units in the PFC and PMC to reach a threshold level of activation.[5]

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

[5] For both regions, I computed the integral of Equation 2.4 and set the threshold on this integral to 700. Baseline predictions were generated by averaging across 300 such trials.

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 2.2. 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 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 2.2 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.

2.3.2 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, I 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 2.3 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 2.4). 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 2.4. 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 2.4.

I 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 2.5, whereas the results for the Hélie, Waldschmidt, and Ashby (2010) experiment are shown in Figure 2.6. 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.

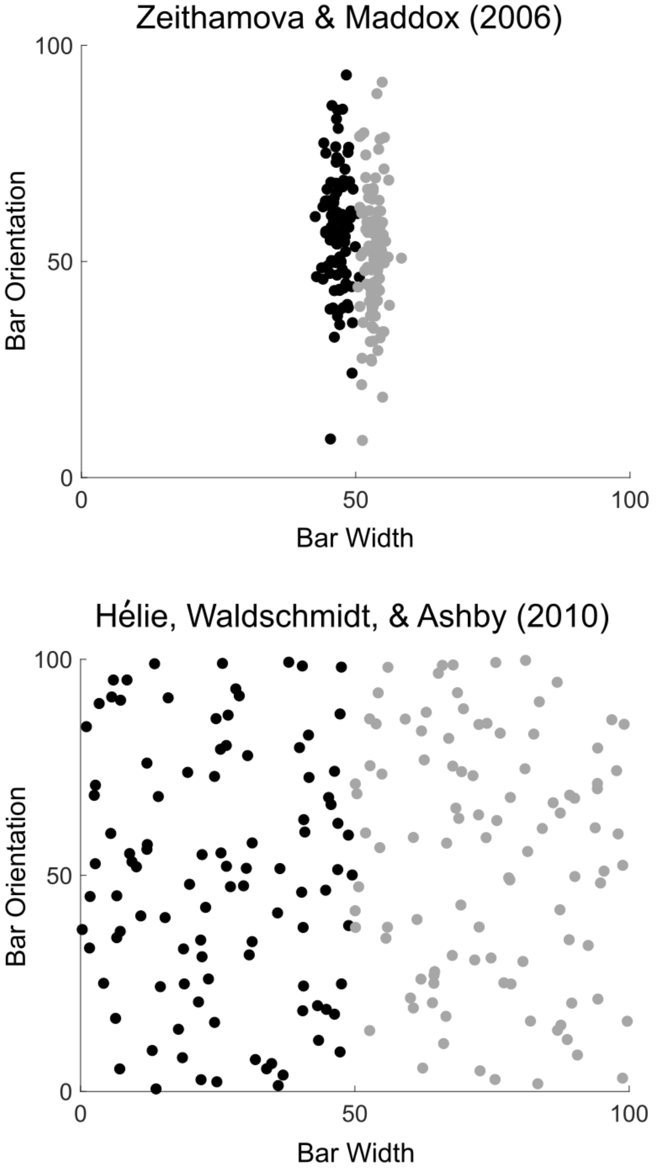

Figure 2.3: Categories used in the dual-task studies of Zeithamova and Maddox (2006) and Hélie, Waldschmidt, and Ashby (2010). Stimuli in both studies were circular sine-wave gratings that varied in bar width (i.e., spatial frequency) and bar orientation. Black dots denote bar width and orientation of category A exemplars, and gray dots identify exemplars of category B. Figure reprinted from "A neurocomputational theory of how rule-guided behaviors become automatic" by Kovacs, P., Hélie, S., Tran, A. N., & Ashby, F. G., *Psychological Review.* Copyright 2021 with permisssion from the American Psychological Association.

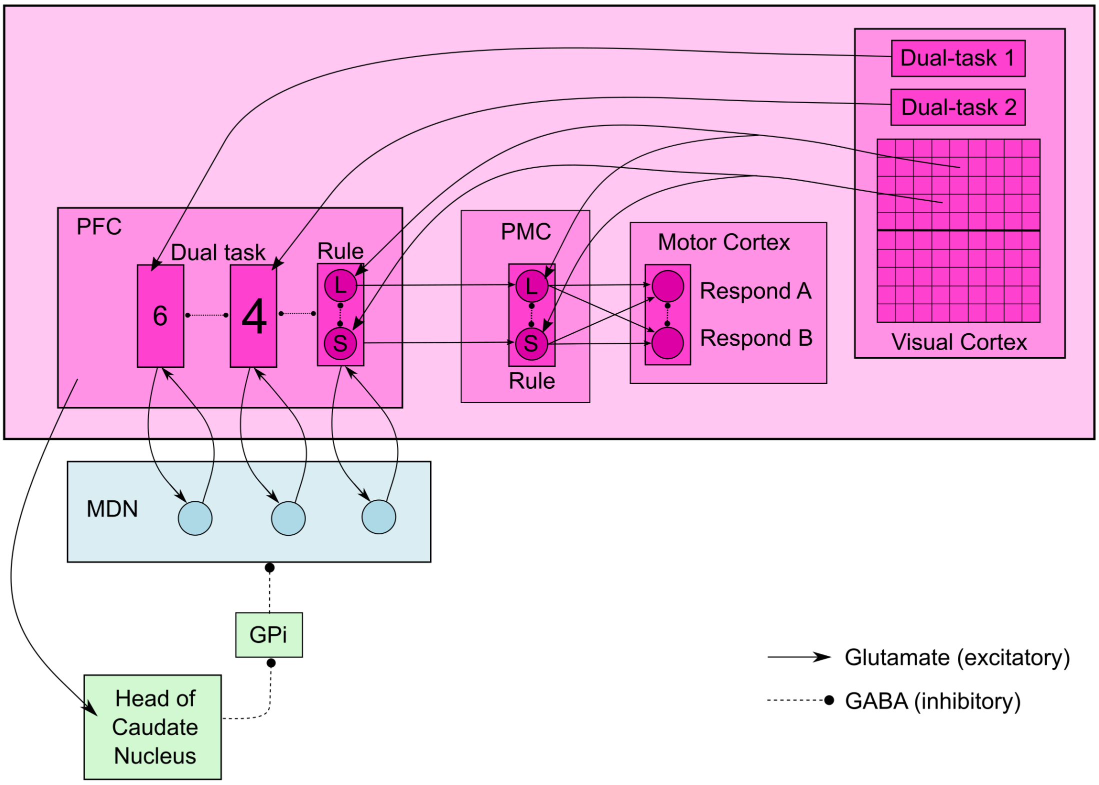

Figure 2.4: The architecture of CARM$^+$ on a dual-task trial when the numbers that flank the categorization stimulus are a physically small 6 and a physically large 4. PFC = prefrontal cortex, PMC = premotor cortex, MDN = medial dorsal nucleus of the thalamus, GPi = internal segment of the globus pallidus. Figure reprinted from "A neurocomputational theory of how rule-guided behaviors become automatic" by Kovacs, P., Hélie, S., Tran, A. N., & Ashby, F. G., *Psychological Review*. Copyright 2021 with permisssion from the American Psychological Association.

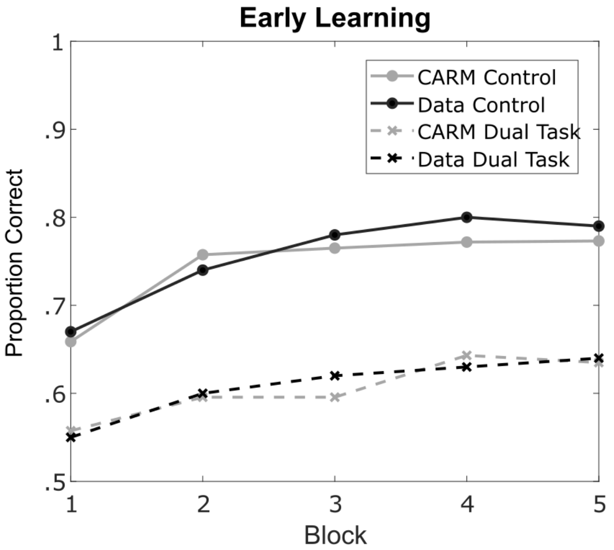

Figure 2.5: Learning curves reported by Zeithamova and Maddox (2006) for their single-task control and dual-task conditions. Also shown are results from CARM$^+$ in the same two conditions. Figure reprinted from “A neurocomputational theory of how rule-guided behaviors become automatic” by Kovacs, P., Hélie, S., Tran, A. N., & Ashby, F. G., *Psychological Review.* Copyright 2021 with permisssion from the American Psychological Association.

Thus, for example, the amount of lateral inhibition on PFC rule units caused by the dual task was identical during early and late learning.

My 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.[6] I 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. I 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

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

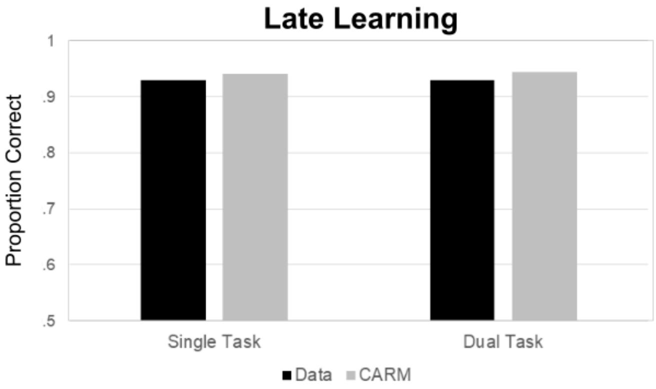

Figure 2.6: Proportion correct for the Hélie, Waldschmidt, and Ashby (2010) participants and for CARM$^+$ during the last session of training (Single Task) and under dual-task conditions.

dual-task session. More research is needed to investigate these possibilities.

Figure 2.7 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, 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 2.4 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 2.7 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 2.7 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

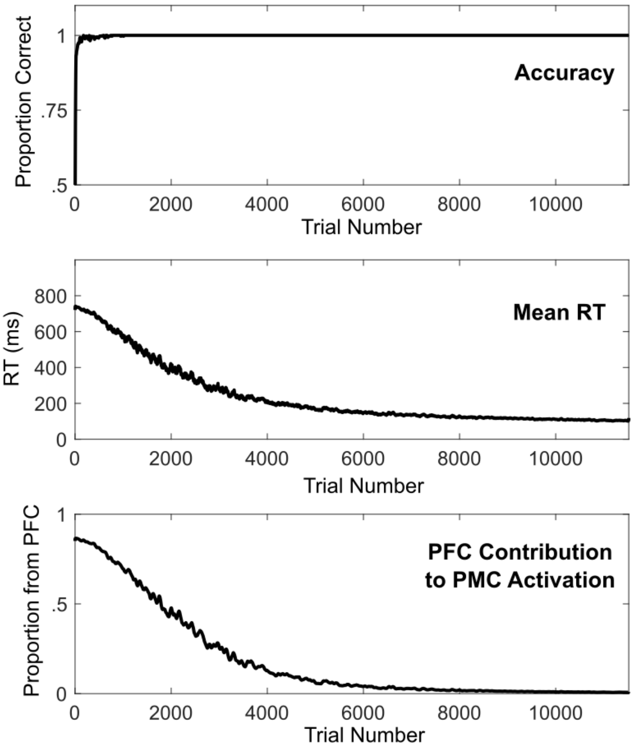

Figure 2.7: Various CARM predictions in the Hélie, Waldschmidt, and Ashby (2010) experiment. The top panel shows the mean proportion of correct categorization responses across 12,000 trials, the middle panel shows mean categorization RT for these same trials, and the bottom panel shows, for the premotor cortex unit that controlled the categorization response, the proportion of the total activation that came from PFC. Figure reprinted from "A neurocomputational theory of how rule-guided behaviors become automatic" by Kovacs, P., Hélie, S., Tran, A. N., & Ashby, F. G., *Psychological Review*. Copyright 2021 with permisssion from the American Psychological Association.

visual cortex. As a result, a dual task that affects PFC has no effect on categorization performance.

The bottom panel of Figure 2.7 also reinforces the widely held view that the development of automaticity is a gradual process that takes thousands of trials to complete (e.g., ?, ?). 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., ?, ?, ?). Note that CARM also predicts that this effect should continue to decrease with additional training.

2.3.3 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. I assume that these response instructions to participants are mediated by top-down executive attention, which I implemented as a gain on projections between PMC and motor cortex (i.e., see Equation 2.7). 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, I 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, I 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" I 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 – I assume that these attentional gains also switch (i.e., from

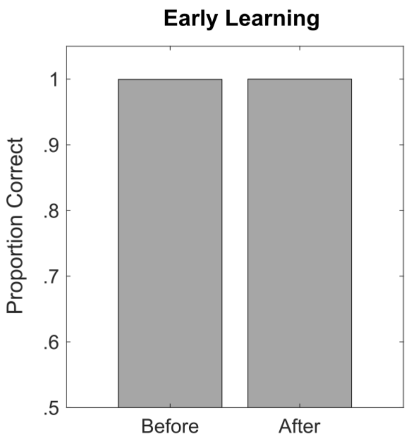

Figure 2.8: Predicted accuracy of CARM$^+$ during the block of trials before and after a button switch, when the switch occurs at the end of the first session of training. Figure reprinted from “A neurocomputational theory of how rule-guided behaviors become automatic” by Kovacs, P., Hélie, S., Tran, A. N., & Ashby, F. G., *Psychological Review.* Copyright 2021 with permisssion from the American Psychological Association.

$\Phi_{LA} = 0.9$ 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 2.8. 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 2.1 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.

However, the model does predict that the same button switch after extended training causes a significant drop in accuracy. Figure 2.9 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

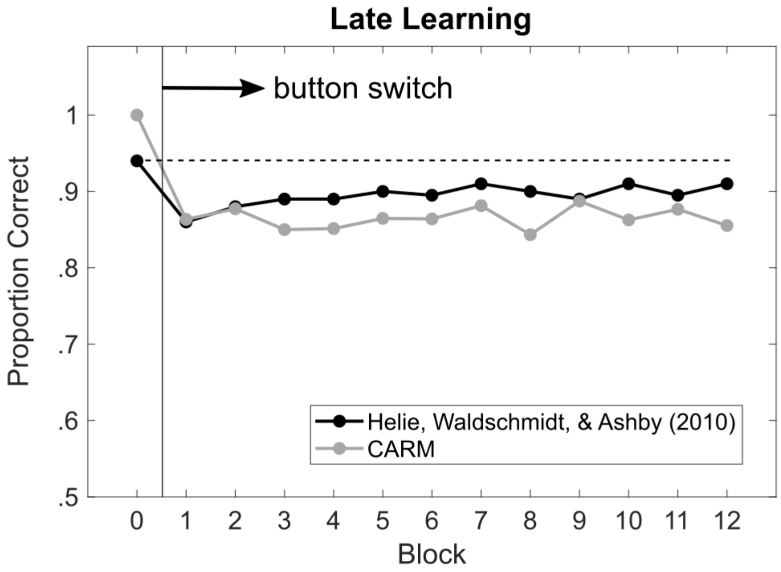

Figure 2.9: Accuracy of participants in each 50-trial block of the experiment reported by Hélie, Waldschmidt, and Ashby (2010). Block 0 shows terminal accuracy following approximately 11,000 trials of training. Participants were instructed to switch response buttons between blocks 0 and 1. The dotted line indicates expected accuracy in the absence of a button switch. Also shown are predictions of CARM under these same experimental conditions. Figure reprinted from "A neurocomputational theory of how rule-guided behaviors become automatic" by Kovacs, P., Hélie, S., Tran, A. N., & Ashby, F. G., *Psychological Review.* Copyright 2021 with permisssion from the American Psychological Association.

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 2.10 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 2.9 and 2.10 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

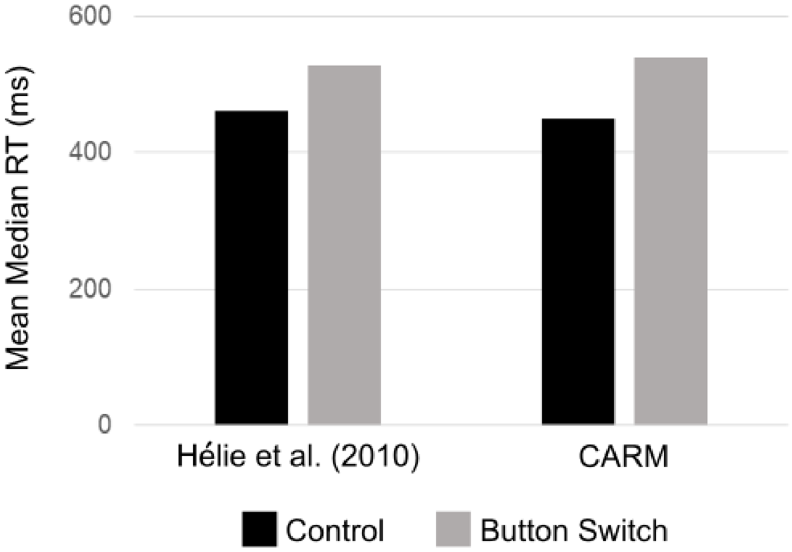

Figure 2.10: Across-participant means of median response times (RTs) reported by Hélie, Waldschmidt, and Ashby (2010) and predicted by CARM. Control RTs are averaged across the last block before the button switch. A motor time of 364 ms was added to the RTs predicted by CARM. Figure reprinted from "A neurocomputational theory of how rule-guided behaviors become automatic" by Kovacs, P., Hélie, S., Tran, A. N., & Ashby, F. G., *Psychological Review.* Copyright 2021 with permisssion from the American Psychological Association.

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.

2.4 General Discussion

This dissertation proposes 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 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 my knowledge, CARM is the first theory to account for all these results simultaneously. In addition, I also showed that CARM successfully accounts for a seemingly counter-intuitive 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.

2.4.1 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 (D. E. 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, I 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. I called this model CARM$^+$. The analogue for SPEED would be to augment it with the procedural-learning module of COVIS, and I 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 (?, ?, ?; Maddox, Ashby, & Bohil, 2003; Maddox & Ing, 2005). As another example, I 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 (?, ?) or if the to-be-learned categories include distinct exceptions (?, ?). 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.

2.4.2 Future Applications of CARM

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 dissertation, 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, ? (?) 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, I should return to the first example considered in this dissertation, 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.[7] Even so, I 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., ?, ?, ?). 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.

2.4.3 Conclusions

This dissertation 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

[7] 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).

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.

Chapter 3

Experiment 1: Automaticity Transfer in a Rule Based Perceptual Categorization Task

3.1 Introduction

Relatively little work has studied exactly what is automatized during the long period of practice that is required for automaticity. Among the first studies to examine this issue reported evidence that the nature of the knowledge that is automatized depends on the learning system used to acquire the behavior. In particular, Roeder and Ashby (2016) reported evidence that rules are automatized with rule-guided behaviors, whereas stimulus-response associations are automatized with skills that are acquired via procedural learning. Stimulus-response associations seem unambiguous, but a rule could be instantiated in many different ways. For example, is the automatized rule an abstract set of instructions that can be applied with equal facility to any relevant stimulus, or is it highly stimulus specific? And does it require selective attention to individual stimulus features or components, or can it operate on the stimulus gestalt?

This chapter describes the results of an experiment that investigated the nature of what is automatized after lengthy practice with a rule-guided behavior. The experiment was designed to test novel predictions of CARM (Kovacs, Hélie, Tran, & Ashby, 2021). The results of this experiment supports the predictions of the model and suggest that an abstract rule, if interpreted as a verbal-based strategy, was not automatized during training, but rather the automatization linked a set of stimuli with similar values on one visual dimension to a common motor response.

Experiment 1 trained 29 naive participants on novel categories of unfamiliar visual stimuli long enough so that their responses became automatic (i.e., 8,400 trials each). Next, each participant completed a final transfer session in which they categorized novel stimuli that they had never seen before. My analyses focused on how well their categorization training prepared them to categorize these novel stimuli. All of the novel stimuli presented during this transfer session could be categorized perfectly using the same strategy that was automatized during training. As a result, I expected transfer accuracy to be high. My main goal therefore, was to assess whether automaticity transferred to the novel stimuli. Specifically, the aim of the experiment was to determine whether participants categorized the transfer stimuli automatically or whether they appealed back to the more effortful categorization strategy they used during the early training sessions.

The stimuli were circular sine-wave gratings that varied across trials in bar width (spatial frequency) and bar orientation. Figure 3.1 illustrates the stimuli and categories used during training and transfer in both of my experimental conditions. There were two training categories and perfect performance could be achieved via the simple one-dimensional rule: "respond A if the orientation of the bars is shallow; otherwise respond B". Participants were given no instructions about the optimal strategy. They were simply told that there were two categories of disks, A and B, and their job was to use the trial-by-trial feedback to learn to assign each presented disk to its correct category.

The experiment included 15 sessions of 600 categorization trials each. Therefore, each participant completed a total of 9,000 categorization trials. The first 14 sessions

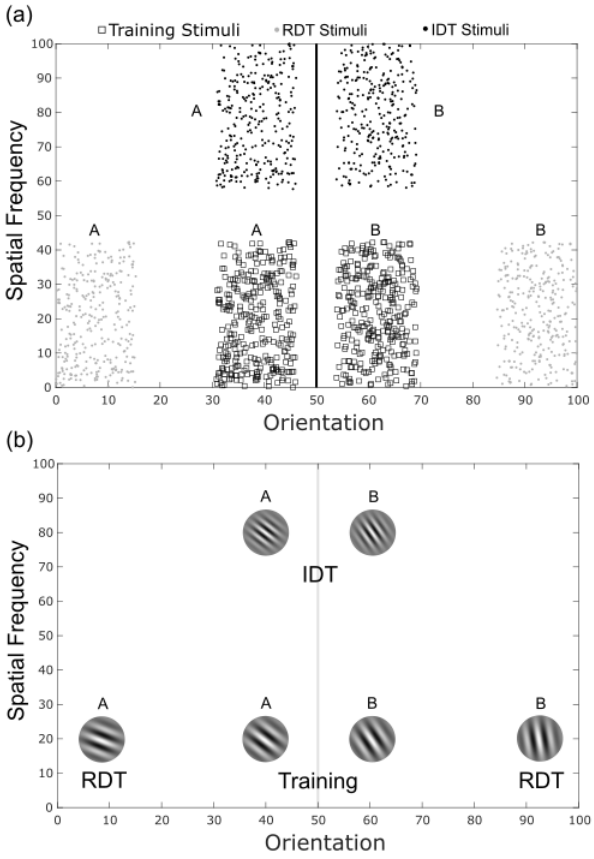

Figure 3.1: Stimuli and category structures used in Experiment 1. The optimal bound for all category structures is $x_1 = 50$. Panel (a) shows coordinate values of all stimuli used and panel (b) shows some example stimuli. Figure reprinted from “On what it means to automatize a rule” by Kovacs, P., & Ashby, F. G., *Cognition*. Copyright 2022 with permisssion from Elsevier.

were identical for all participants. Each of these 8,400 trials (i.e., 14 × 600) were standard categorization trials. The stimuli and categories used during training are denoted in Figure 3.1 by the open squares. The goal of the training sessions was to train participants on the categorization task long enough that their responses became automatic. Previous research with the same stimuli indicated that 8,400 trials of training was sufficient for automaticity to develop (Hélie, Waldschmidt, & Ashby, 2010).

The nature of the knowledge that participants acquired during training was assessed during the final transfer session (i.e., session 15). There were two conditions, with separate participants in each condition. In the Relevant-Dimension Transfer (RDT) condition, the stimuli presented to participants changed values on the relevant dimension (i.e., orientation of the bars), but not on the irrelevant dimension (bar width). The transfer stimuli in the RDT condition are denoted in Figure 3.1 by the light gray dots. Note that the separation between the category A and B exemplars in the RDT condition is greater during transfer than during training, and as a result, the transfer task is objectively easier than the training task. In the Irrelevant-Dimension Transfer (IDT) condition, the stimuli changed values on the irrelevant dimension (i.e., bar width), but not on the relevant dimension. The transfer stimuli in the IDT condition are denoted in Figure 3.1 by the black dots. Note that the separation between the category A and B exemplars in the IDT condition is the same as during training, so the IDT transfer task is objectively equal in difficulty to the training task.

Note that the simple one-dimensional rule that perfectly categorizes the training stimuli also works perfectly in both transfer conditions. As a result, based on previous research, I expected transfer accuracy to be high in both conditions (Casale, Roeder, & Ashby, 2012). For this reason, my primary goal was to determine whether automaticity transferred to the novel stimuli that participants categorized during the final session. To answer this question, I used two classic tests for assessing automaticity – the performance of automatic behaviors should be: 1) unaffected by having to perform a simultaneous dual task, and 2) impaired if the location of the response

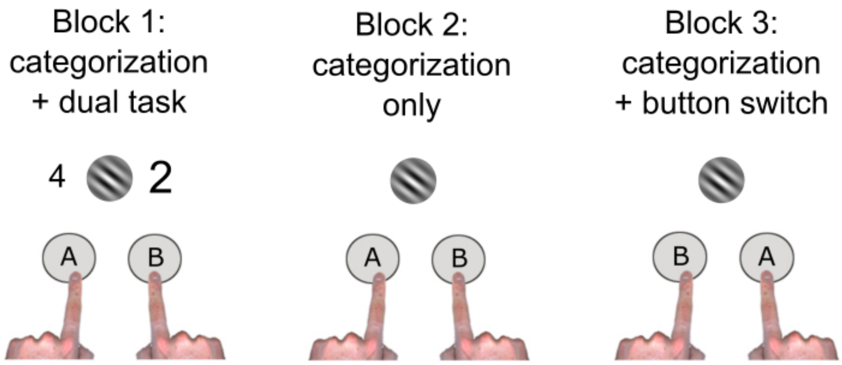
Block 1: categorization + dual task
Block 2: categorization only
Block 3: categorization + button switch
4
2
A
B
A
B
B
A

Figure 3.2: Description of the three 200-trials blocks of the 15th and final (transfer) session of Experiment 1. All stimuli during this session were either from the IDT or RDT categories shown in Figure 1. During the first 200 trials, participants categorized the novel stimuli while completing a simultaneous numerical Stroop dual task. During the second block of 200 trials, participants categorized the stimuli under the same procedures as during the first 14 training sessions. Finally, during the last block of 200 trials, participants categorized the stimuli in the usual manner, except the locations of the response buttons were reversed, and participants were explicitly instructed of this change. Figure reprinted from "On what it means to automatize a rule" by Kovacs, P., & Ashby, F. G., *Cognition*. Copyright 2022 with permisssion from Elsevier.

buttons is reversed (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977).

To implement these tests, the final session was divided into three separate blocks of 200 trials each. These are described in Figure 3.2. During the first 200 trials (block 1), participants categorized the novel transfer stimuli while simultaneously performing a dual task that required working memory and executive attention (i.e., a numerical Stroop task). During the third block of 200 trials, participants categorized the transfer stimuli using the same procedures as during training, except that the locations of the response buttons were switched. Participants were informed of this switch before the block began and cues were presented on the screen on every trial that signaled the new button locations. Therefore, no new learning was required. Finally, during the second block of 200 trials, participants categorized the transfer stimuli using the same procedures as during training. The data from these trials served as a baseline or control that was used to assess the effects of the dual task and button switch on performance. Therefore, in summary, the final session followed a 2 × 3 factorial design, in which 2 conditions (RDT, IDT) were crossed with 3 block types (categorization only, dual task, button switch).

3.2 Predictions of CARM

This dissertation proposes a neurocomputational model of rule-based automaticity entitled CARM. Figure 3.3 shows the model as it would look at the end of the 14 training sessions of Experiment 1. The model assumes that Hebbian learning will strengthen all active synapses in the Figure 3.3 network. The most critical of these for behavioral predictions are highlighted in the figure by the thicker projections. First consider the synapses between visual cortex and PMC. In Hebbian learning, synaptic strengthening is proportional to the product of the pre- and post-synaptic activations. During early training, much of the post-synaptic activation (i.e., the activation within the PMC units) is driven by input from PFC. As the visual cortex-to-PMC synaptic strength increases, it eventually becomes strong enough so that visual input alone is enough to cause the PMC unit to activate the appropriate target in motor cortex. The pathway through PFC is still active, but because it is longer, it no longer controls behavior. At this point, the behavior has become automatic.

Second, consider the synapses between PMC and primary motor cortex. Initially these are weak because participants have no prior association between shallow or steep orientations and A or B button presses. But after thousands of practice trials, Hebbian learning will strengthen these associations. The model therefore predicts that both transfer conditions will be susceptible to a button-switch interference. This is because the transfer conditions introduce novel stimuli, but the categorization rule and motor responses remain the same as during training.

The model successfully accounts for single-unit recordings and human behavioral data that are problematic for other models of automaticity. For example, it accounts for resistance to dual-task interference because the working memory circuits centered in PFC are not needed to initiate automatic behaviors, and it accounts for an interference when the response button locations are switched because Hebbian learning between PMC and primary motor cortex strengthens the motor associations during training so much that top-down executive attention is unable to reverse them completely after the switch occurs.

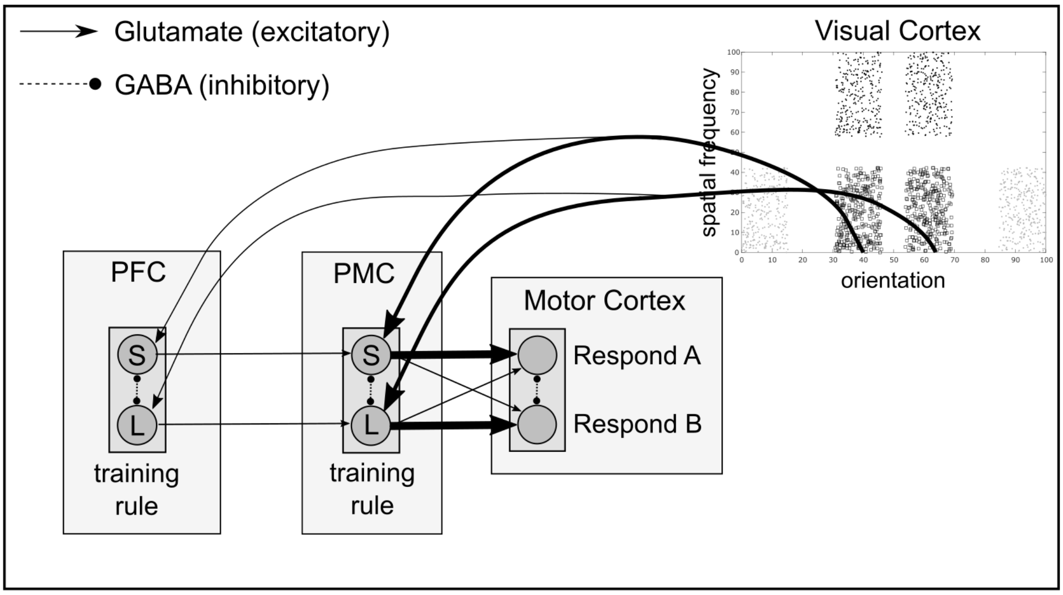

Figure 3.3: A schematic of the Kovacs et al. (2021) model as it would look at the end of the training sessions of Experiment 1. The thicker projections represent increases in synaptic strength that result from Hebbian learning. PFC = prefrontal cortex, PMC = premotor cortex, S and L refer to units that respond to stimuli with small and large orientations, respectively. Figure reprinted from "On what it means to automatize a rule" by Kovacs, P., & Ashby, F. G., *Cognition*. Copyright 2022 with permisssion from Elsevier.

This model predicts that automatic rule-guided behaviors are stimulus specific, but initial rule-guided behaviors are not. In particular, the model predicts that early rule-guided behaviors are mediated by abstract rules that are represented in PFC, whereas automatic rule-guided behaviors are mediated by direct projections from visual cortex to PMC units that control the behavior. Because of Hebbian learning, the associations between the stimulus representations in visual cortex and the motor associations in PMC eventually become strong enough to trigger the behavior without assistance from the abstract rule representations in PFC.

Now consider the predictions of the model for the RDT and IDT conditions of Experiment 1. In the RDT condition, the model predicts that the novel orientations of the transfer stimuli will activate visual cortical neurons that were never activated during training. As a result, their synapses into PMC will be weak (i.e., untrained), dropping the PMC response to visual input below the threshold needed to activate motor cortex. In this case, PFC input is needed to cause enough PMC activation to trigger a motor response. Accuracy should remain high, however, because the PFC retains the representation of the correct rule. Even so, because application of that rule now depends on working memory and executive attention (unlike automatic behaviors), transfer performance should be susceptible to dual-task interference. This is a strong prediction because the transfer categories are more widely separated in the RDT condition than the training categories (see Figure 3.1), and therefore the transfer task is objectively easier than the training task. Thus, the model predicts that even though the transfer categories are easier, participants should lose the ability to respond automatically to the RDT transfer stimuli.

Somewhat counterintuitively, however, the model also predicts that transfer performance during the button-switch block of the RDT condition should appear automatic, in the sense that it should be susceptible to button-switch interference. This is because the model predicts that no matter how the response is selected, response execution is mediated by the same PMC-to-primary motor cortex projections during both training and transfer. Therefore, even if control is passed back to PFC during the RDT blocks, the same PMC-to-primary motor projections must be used to initiate the motor response as during training, and therefore a button-switch interference should still occur. In summary then, the model makes a set of strong and novel predictions about transfer performance in the RDT condition: 1) accuracy should remain high, 2) performance during the simultaneous dual-task should appear non-automatic (i.e., susceptible to interference), and 3) performance after the button switch should appear automatic (also susceptible to interference).

Next, consider the IDT condition. The only difference between the IDT and RDT conditions is in the transfer stimuli. In both conditions, the categorization rule remains the same during training and transfer, and so do the response buttons. As a result, the model predicts that transfer accuracy should be high in both conditions and both conditions should be susceptible to a button-switch interference. But what about a dual-task interference? The IDT transfer stimuli differ from the training stimuli, but only on the irrelevant dimension. So the model predictions depend on what type of visual representation projects to PFC and PMC. If the projections from visual cortex to PMC are of the stimulus gestalt, then the visual inputs to PFC and PMC change in both conditions, so the model makes identical predictions in the RDT and IDT conditions. Abstract rule representations in PFC would be needed to initiate motor behaviors in both conditions, so IDT transfer responding should be susceptible to a dual-task interference. In contrast, if the projections from visual cortex to PMC are only of values on the relevant dimension, then the model predicts that dual-task and button-switch performance should both remain automatic because from the perspective of PMC, the visual representations received during transfer would be identical to the visual representations received during training (since the stimuli do not change on the relevant dimension).

Kovacs et al. (2021) made no assumptions about whether the visual representations used by the model were of stimulus gestalts or were restricted to the relevant stimulus dimensions only. Even so, there is reason to favor the hypothesis that the representations are of single dimensions. For example, humans learn categories like the ones used during the Experiment 1 training – in which the optimal strategy is a simple one-dimensional rule – much more quickly than categories that are identical except the stimulus space is rotated 45°, so that the optimal decision boundary is diagonal (e.g., Ashby, Smith, & Rosedahl, 2020). In contrast, pigeons and rats learn both types of categories at exactly the same rate (Broschard, Kim, Love, Wasserman, & Freeman, 2019; Qadri, Ashby, Smith, & Cook, 2019; Smith et al., 2011). This across-species difference supports the hypothesis that the human one-dimensional advantage is due to their ability to apply explicit rules with one-dimensional categories, and that pigeons and rats lack this ability. Critically though, both macaque and capuchin monkeys show a similar advantage to humans in the one-dimensional task, relative to the rotated diagonal-bound task (Smith, Beran, Crossley, Boomer, & Ashby, 2010; Smith, Crossley, et al., 2012; Smith et al., 2015). This result suggests that the human one-dimensional learning advantage is not necessarily language based, and instead may be due to an ability to selectively attend to the single relevant dimension – a skill that is closely tied to PFC (e.g., Miller & Cohen, 2001). If so, then it seems natural that the visual representations used by the PFC rule units would exploit this selective attention ability.

Experiment 1 tests some highly non-intuitive predictions of the Kovacs et al. (2021) theory – for example, that transfer performance in the RDT condition should appear automatic during the button-switch trials but non-automatic during the dual-task trials, and that this loss of automaticity in the presence of a dual task should occur even though the RDT transfer stimuli are objectively easier to categorize than the training stimuli (i.e., the RDT transfer categories are more widely separated than the training categories). In addition, it also tests whether the visual representations supporting explicit rule use are of gestalts or limited only to the relevant stimulus dimension.

3.3 Methods

Participants

Twenty-nine healthy undergraduate students at the University of California, Santa Barbara, participated in this experiment in exchange for class credit. Fourteen participants were randomly assigned to the RDT condition, and the remaining 15 participants were assigned to the IDT condition.

Stimuli and Apparatus

All stimuli were circular sine-wave gratings of constant contrast and size presented on a 21-in. monitor (1,280 × 1,024 resolution). Each stimulus was defined by a set of points (x_1, x_2) sampled from a 100 × 100 stimulus space and converted to a disk using the following equations: spatial frequency = $2^{(x_1/28)}$ cycles per disk and orientation = $9x_2/10 + 15$ degrees counterclockwise rotation from horizontal.

During training, stimuli in category A were uniformly distributed (in the 100×100 space) in the interval [30.77, 46.15] on the orientation dimension and [0, 42.31] on the spatial frequency dimension. Stimuli in category B were also uniformly distributed, over the intervals [53.85, 69.23] and [0, 42.31] for orientation and spatial frequency, respectively. The stimuli were generated with PsychoPy (2009), and subtended an approximate visual angle of 13°. Note that perfect accuracy is possible if participants use the simple one-dimensional decision rule: Respond A if the orientation is less than 50°; otherwise respond B.

During the transfer session, the stimulus values were the same as during training, except in the RDT condition, the stimulus values were shifted on the relevant dimension – that is, orientation – whereas in the IDT condition they were shifted on the irrelevant dimension (i.e., spatial frequency). In the RDT condition, the category A stimuli were uniformly distributed over the intervals [0, 15.35] and [0, 42.31] for orientation and spatial frequency, respectively, and the category B stimuli were uniformly distributed over the intervals [84.62, 100] and [0, 42.31] for orientation and spatial frequency, respectively. In the IDT condition, the category A stimuli were uniformly distributed over the intervals [30.77, 46.15] and [57.7, 100] for orientation and spatial frequency, respectively, and the category B stimuli were uniformly distributed over the intervals [53.85, 69.23] and [57.7, 100] for orientation and spatial frequency, respectively.

Stimulus presentation, feedback, response recording, and response time (RT) measurement were acquired and controlled using PsychoPy on a Macintosh computer. Responses were given on a standard Macintosh keyboard: the “D” key for an A categorization and the “K” key for a B categorization (sticker-labeled as either A or B). A participants who hit the right key saw the word “Correct” on the screen in green letters, and a participant who hit the wrong key saw the word “Incorrect” in red letters.

Procedure

The experiment lasted for 15 sessions over 15 consecutive workdays. The first 14 sessions were training, and the last session was transfer. Each session included 600 categorization trials. All together, each participant completed 8,400 trials of training and 600 trials of transfer.

On training days, participants were informed that they were taking part in a categorization experiment and were instructed to assign each stimulus to one of two categories, either A or B. A single trial proceeded as follows: The stimulus appeared in the center of the screen and remained on the screen until the participant made a response, correct or incorrect visual feedback appeared immediately on the screen and remained on the screen for 2 seconds.

On the transfer session day participants performed a total of 600 trials split into three blocks: 1) 200 trials of categorization with a concurrent numerical Stroop task, 2) 200 trials of categorization only, and 3) 200 trials of categorization with the locations or the response buttons switched.

During the 200 dual-task trials of the transfer session (block 1), two different digits were randomly chosen on every trial (ranging from 2 to 8), and displayed for 1 sec on the left and right of the center of the screen, with each offset by approximately 2° of visual angle. One of the digits was displayed in a larger font at 6 cm in height. The other digit was 3 cm in height. A "congruent" trial in the numerical Stroop task was defined as a trial in which the digit with the larger value was displayed in a larger font, whereas an "incongruent" trial was defined as a trial where the digit with the smaller value was displayed in the larger font. Incongruent trials produce a Stroop-like interference (Waldron & Ashby, 2001). The response keys and feedback for the numerical Stroop task were the same as for the categorization task. The D key (labeled A) was used to indicate left, and the K key (labeled B) was used to indicate right (matching their locations on a regular keyboard).

Participants were instructed to memorize the numerical value and physical size of the two digits. The digits disappeared followed by a blank screen for 300 msec, followed by the categorization stimulus. The categorization stimulus stayed on the screen until a categorization response was made. Feedback was given after 300 msec and stayed on the screen for 700 msec. After the feedback, the screen went blank for

300 msec followed by a cue, either the word “Size” or the word “Value” If the cue was “Size,” the participant needed to indicate whether the number presented in the larger font was on the right or the left of the screen. If the cue was “Value,” the participant needed to indicate whether the number of larger value was on the right or the left of the screen. The cue remained on the screen until the participant responded. Feedback was given in the same way as in the categorization task. As in the training sessions, half the categorization stimuli were from category A and half were from category B. In the numerical Stroop task, 170 trials were incongruent (85%), and the remaining 30 trials were congruent (15%). This manipulation aimed at drawing the analogy with the original Stroop task – that is, by opposing the natural bias of associating digit size with digit value. Half the correct responses were located on the left, and half on the right. Also, the digit with the larger value was located on the left for half the trials, and half the digits with the larger size were located on the left. Participants were instructed to focus on the numerical Stroop task and to perform the categorization task with the attentional resources they had left. Additionally, participants were instructed to respond as quickly as they could without sacrificing accuracy.

The trial-by-trial procedures for the 200 categorization-only trials of the transfer session (block 2) were identical to the training sessions. During the break between blocks 1 and 2, participants were again instructed to respond as quickly as they could without sacrificing accuracy.

During the 200 button-switch trials of the transfer session (block 3), categorization trials were identical to training trials except the categorization response key locations were switched. The letters “A” and “B” were displayed on the left and right side of the bottom of the screen in positions corresponding to the new locations of the response keys. Participants were instructed at the end of block two that everything in the next 200 trials would be the same except that the response keys would switch positions. They were also instructed to refer to the letters “A” and “B” displayed at the bottom of the screen to remind them of the new button locations. Additionally, participants were again instructed to respond as quickly as they could without sacrificing accuracy.

3.4 Results

Figure 3.4 shows the mean proportion correct averaged over participants during each session of training. As expected, accuracy increased quickly and plateaued at a high level of performance (above 90% correct). The means of each participant's median RTs are shown in Figure 3.5. Also as expected, note that RT gradually decreased over sessions, beginning at about 700 ms on session 1 and ending at 580 ms during the last training session (i.e., Session 14).

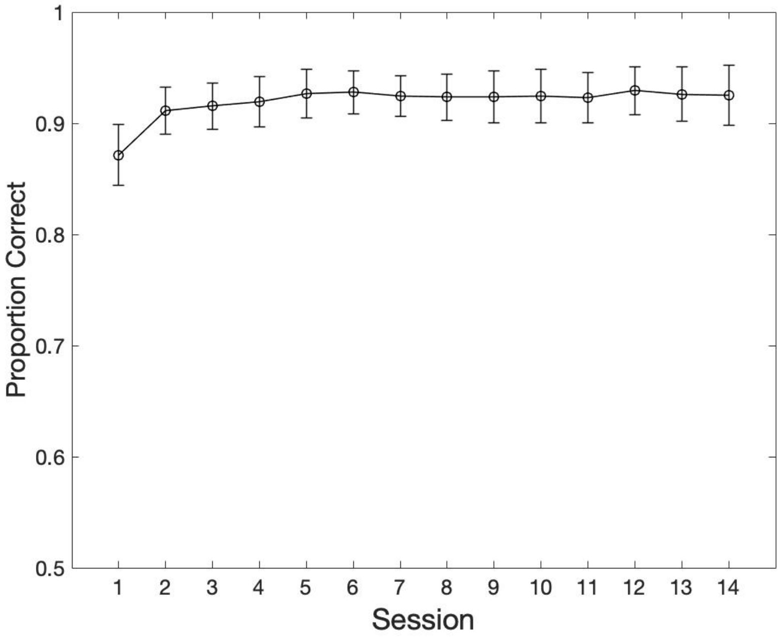

Figure 3.4: Mean proportion correct for all training sessions of Experiment 1 averaged across participants. The error bars are 95% confidence intervals. Figure reprinted from "On what it means to automatize a rule" by Kovacs, P., & Ashby, F. G., *Cognition.* Copyright 2022 with permisssion from Elsevier.

Standard Statistical Analysis

Results from the final transfer session are summarized in Figure 3.6. The data from the categorization-only trials (i.e., block 2) were used as controls.

As a first analysis, I analyzed the transfer session data using a series of generalized linear mixed models (GLMM). The accuracy analysis assumed a logistic link function, whereas the link function for the RT analysis was the identity. The main advantage

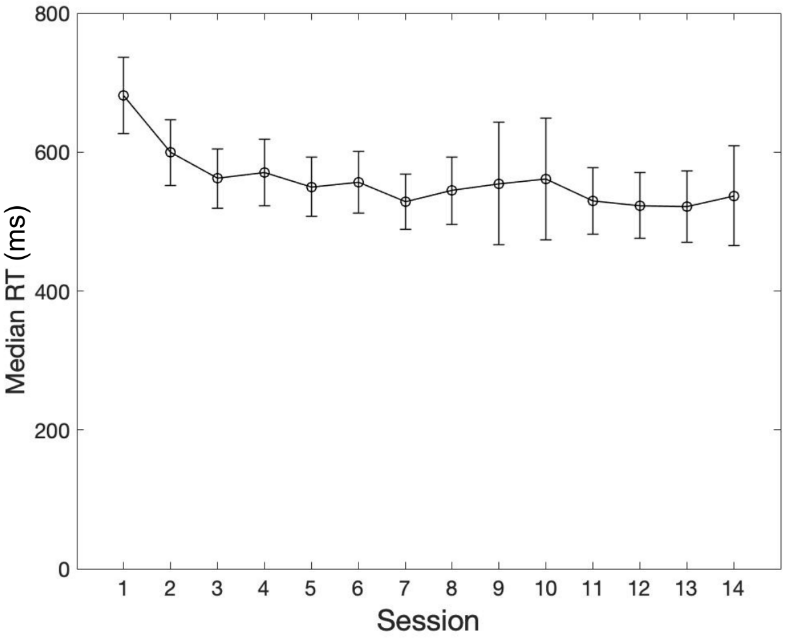

Figure 3.5: Median RTs for all training sessions of Experiment 1 averaged across participants. The error bars are 95% confidence intervals. Figure reprinted from “On what it means to automatize a rule” by Kovacs, P., & Ashby, F. G., *Cognition*. Copyright 2022 with permisssion from Elsevier.

of using a GLMM analysis instead of ANOVA is that, in the case of the trial-by-trial Bernoulli distributed accuracy data, the ANOVA assumption of normality is violated. However, I also analyzed the RTs using a standard ANOVA and the results were qualitatively identical.

Recall that the final transfer session followed a 2 × 3 factorial design, in which 2 conditions (RDT, IDT) were crossed with 3 block types (categorization only, dual task, button switch). Therefore, the GLMM analysis included all of the models that would be tested in a standard ANOVA. This includes a null model in which there are no main effects or interaction, a model that only includes a main effect of condition (model Cond), a model that only includes a main effect of block (model Block), a model that includes main effects of condition and block (model CondBlock), and a full model that includes both main effects and an interaction. Separate GLMM analyses were performed for accuracy and RT. The accuracy results are described in Table 3.4 and the RT results are shown in Table 3.2.

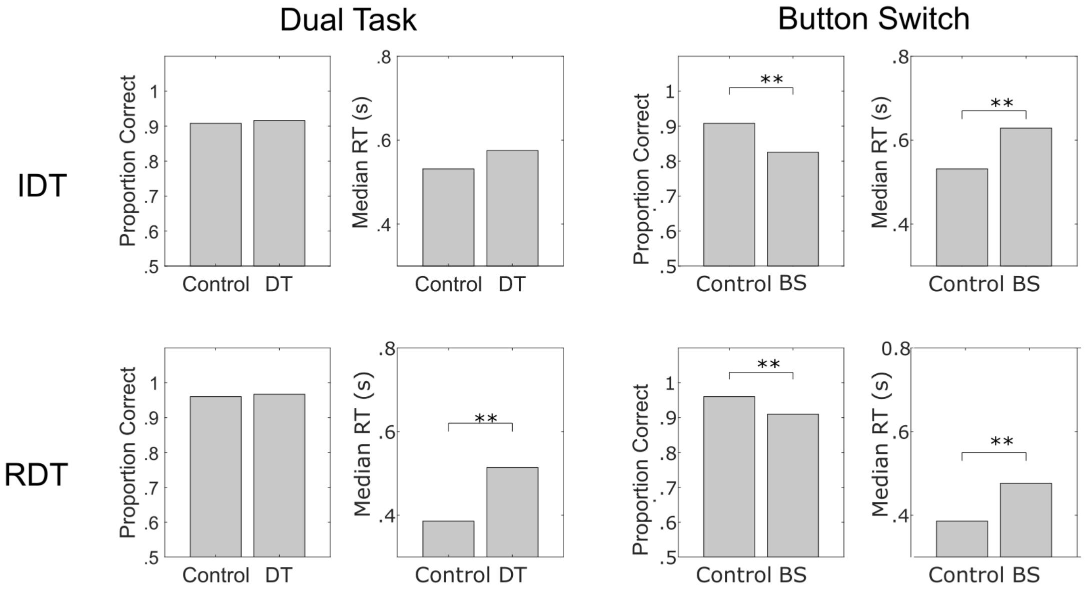

Figure 3.6: Results from the final transfer session of Experiment 1. Control results are from the categorization-only block (i.e., block 2). DT = data during a simultaneous dual task; BS = data while the response buttons have switched locations. Accuracy values are computed as a mean of each participant's proportion correct. RTs are the mean of each participant's median RT. Comparisons were performed with t-test (** indicates p ¡ 0.005). Figure reprinted from "On what it means to automatize a rule" by Kovacs, P., & Ashby, F. G., *Cognition.* Copyright 2022 with permisssion from Elsevier.

For the accuracy analysis, the best-fitting model was CondBlock, suggesting both main effects were significant, but not the interaction. The Bayes factors (BF) suggest that the evidence for both main effects is extreme, and the evidence that there is no interaction is also extreme (Lee & Wagenmakers, 2014). An examination of Figure 3.6 suggests that the main effect of condition is driven by the higher accuracy in the RDT condition than in the IDT condition. This is not surprising since the RDT transfer stimuli were objectively easier to categorize than the IDT transfer stimuli (i.e., see Figure 3.1). The main effect of block is driven by the lower accuracy during the button-switch block compared to the control or dual-task blocks, and the lack of an interaction suggests that the lower button-switch accuracy was similar in both conditions.

The RT analysis led to different conclusions. The evidence for both main effects was again extreme, but now the evidence for an interaction was also extreme. In

Model	Terms	Log L	BIC	BF
Null	β_0	5126	10263	1
Cond	$\beta_0 + C$	5015	10050	1.5e46
Block	$\beta_0 + B$	5009	10048	4.0e46
CondBlock	$\beta_0 + C + B$	4897	9832	2.9e93
Full	$\beta_0 + C + B + (C \times B)$	4895	9849	6.2e89

Table 3.1: GLMM results for the accuracy data from the Experiment 1 transfer session. This table reprinted from "On what it means to automatize a rule" by Kovacs, P., & Ashby, F. G., *Cognition.* Copyright 2022 with permisssion from Elsevier.

Model	Terms	Log L	BIC	BF
Null	β_0	13634	27288	1
Cond	$\beta_0 + C$	13388	26806	5.6e104
Block	$\beta_0 + B$	13441	26921	6.2e79
CondBlock	$\beta_0 + C + B$	13189	26427	9.5e186
Full	$\beta_0 + C + B + (C \times B)$	13149	26366	1.7e200

Table 3.2: GLMM results for the RTs from the Experiment 1 transfer session. This table reprinted from "On what it means to automatize a rule" by Kovacs, P., & Ashby, F. G., *Cognition.* Copyright 2022 with permisssion from Elsevier.

particular, the Full model was, by far, the best-fitting model, and a comparison of the Bayes factors for the Full and CondBlock models suggests the evidence for an interaction was extreme.[1] Figure 3.6 suggests that the main effect of condition is driven by the faster RTs in the RDT condition and the main effect of block is largely due to the faster RTs during the control block. The difference between the control and button-switch RTs is approximately the same in the two conditions, so the highly significant interaction is driven by the much larger difference between the control and dual-task RTs in the RDT condition than in the IDT condition.

I also assessed all pairwise differences in Figure 3.6 for significance via standard t-tests. These largely confirmed the GLMM analyses. In the IDT condition, the difference between control and dual-task accuracy was not significant [$t(14) = 0.70$, $p = 0.49$], nor was the RT difference [$t(14) = 1.73$, $p = 0.11$]. However, the differences between control and button-switch performance were significant – both for accuracy [$t(14) =$ -6.35, p ¡ .005] and RT [$t(14) = 4.88$, p ¡ .005]. In the RDT condition, the

[1] The Bayes factors in Tables 3.4 and 3.2 estimate the likelihood of the model relative to the likelihood of the null model. The ratio of the Bayes factors for the Full and CondBlock models estimates the likelihood of the Full model relative to the CondBlock model.

difference between control and dual-task accuracy was not significant [$t(13) = 1.25$, $p = .23$], but the RT difference was significant [$t(13) = 4.46$, p ¡ .005]. Finally, the control versus button-switch differences were both significant in the RDT condition [accuracy: $t(13) =$ -5.19, p ¡ .005; RT: $t(13) = 6.35$, p ¡ .005].

The t-tests suggest that both conditions exhibited a button-switch interference that was characterized by a decrease in accuracy and an increase in RT (relative to control) when the response buttons switched locations. On the other hand, these tests also suggest no effect on accuracy of the dual task in either condition, but a significant increase in RT in the RDT condition only. To examine this RT difference more closely, Figure 3.7 shows the median RTs (averaged across participants) during each 40-trial block of the dual-task trials. Also shown for comparison are the mean RTs during the categorization-only trials. Note that in both conditions, responding is slower in block 1 than in any subsequent blocks – presumably because there was a settling-in period as participants adjusted to the sudden demand to perform two tasks at once. Furthermore, RT dropped about equally from blocks 1 to 2 in both conditions. Therefore, this figure suggests that the most appropriate comparison is between performance on blocks 2 – 5. When dual-task RTs are compared to control RTs over these blocks, t-tests indicate that the effect of the dual task on RT was not significant in the IDT condition [$t(14) = 1.30, p = 0.22$], and highly significant in the RDT condition [$t(13) = 4.17, p = 0.001$].

Decision-Bound Modeling Analysis

Before attempting to interpret these results, it is important to assess the type of decision strategy that participants were using. This is because a variety of different strategies could lead to approximately equal accuracies, and one group could have higher accuracy than another, not because they were more likely to use a strategy of the optimal type, but for some other reason (e.g., better criterial learning; less criterial noise). To examine this issue, I fit a variety of different decision-bound models (Ashby & Valentin, 2018; Maddox & Ashby, 1993) to the responses of individual

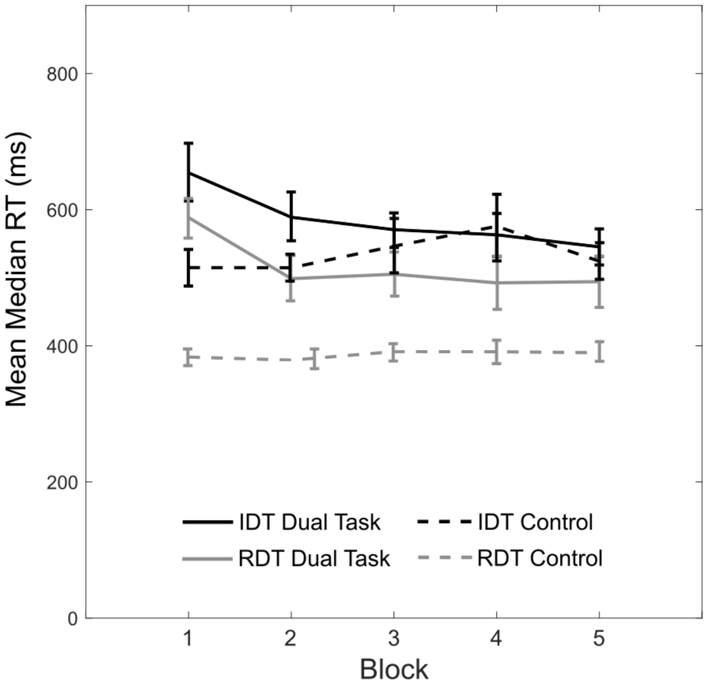

Figure 3.7: The mean of all participants median RTs for each 40-trial block during the Experiment 1 transfer-session dual-task trials. The dotted lines show the mean RTs from the categorization-only trials of the transfer session. The error bars denote standard errors. This table reprinted from "On what it means to automatize a rule" by Kovacs, P., & Ashby, F. G., *Cognition.* Copyright 2022 with permisssion from Elsevier.

participants separately during each of their 15 experimental sessions. The models assumed a procedural strategy, a rule-based strategy, or random guessing. These models are described in the Appendix, but briefly, the rule-based models assumed a single vertical or horizontal decision bound. The procedural-strategy model assumed that the decision bound was a single line of arbitrary slope and intercept, and the guessing models assumed that participants guessed randomly on each trial. The procedural and rule-based models all included a noise variance parameter, and either one (in the case of the rule models), or two (in the case of the procedural model) free parameters that described the decision bound. For every participant, each of these different models was fit separately to responses from each of the 14 training sessions, and to each of the three 200-trial blocks of the transfer session and in each case, the best-fitting model was recorded (i.e., the model with the lowest value of the BIC goodness-of-fit statistic).

During the first session, 86% of the participants responses were best accounted for by a model of the optimal type – that is, a model that assumed a vertical line decision bound. During the other training sessions, this percentage ranged from 72% to 100%. In all cases that a vertical-bound rule model did not fit best, the best fit was provided by a model that assumed a procedural strategy. However, in all cases, visual examination of the decision bounds predicted by these models indicated a bound that was nearly vertical – suggesting that there were only a few trials in these data sets that included responses that were inconsistent with a vertical-bound rule. Overall, this analysis suggests that participants clearly learned the optimal categorization strategy early in training and used this strategy consistently throughout the 13 training sessions.

The results for the transfer session are shown in Table 3.3. Note that in both conditions, use of the optimal strategy was high in all three blocks. Therefore, the appearance of novel stimuli did not cause participants to switch strategies, nor did the presence of a dual task. Even the button switch had only a minor effect on strategy – confusing a few RDT participants enough to cause them to resort to guessing.

Block	IDT	RDT
Single-Task Control		
Optimal 1D Rule	13 (87%)	13 (93%)
Procedural Strategy	2 (13%)	0 (0%)
Guessing	0 (0%)	1 (7%)
Dual Task		
Optimal 1D Rule	15 (100%)	13 (93%)
Procedural Strategy	0 (0%)	0 (0%)
Guessing	0 (0%)	1 (7%)
Button Switch		
Optimal 1D Rule	14 (93%)	9 (64%)
Procedural Strategy	1 (7%)	0 (0%)
Guessing	0 (0%)	5 (36%)

Table 3.3: Decision-bound modeling results of the Experiment 1 transfer data. Number and percentage (in parentheses) of participants whose responses were best accounted for by each type of decision bound model. This table reprinted from "On what it means to automatize a rule" by Kovacs, P., & Ashby, F. G., *Cognition.* Copyright 2022 with permisssion from Elsevier.

3.5 Discussion

Twenty-nine participants each completed 8,400 categorization training trials distributed over 14 experimental sessions. During this time they repeatedly practiced a simple one-dimensional categorization rule. Previous research suggests that after this amount of training, their responses were automatic. The participants were then divided into two groups and both groups completed one final session of 600 trials. During this last session, all participants saw new stimuli that could be categorized using the same rule that they had automatized during training. In the IDT condition, the new stimuli had identical values as the training stimuli on the relevant dimension and unique values on the irrelevant dimension. In the RDT group, the opposite occurred – that is, the new stimuli had novel values on the relevant dimension, but the values on the irrelevant dimension were the same as in training. I then assessed whether automaticity persisted for these novel stimuli by examining performance in the presence of a dual task, and following a switch of the response buttons.

Accuracy was universally high in both conditions, suggesting that participants had no trouble transferring the rule they had been practicing to the novel stimuli. Similar results have been reported after only one session of training (Casale et al., 2012), so this result is not unexpected.

The more interesting results concern my tests of whether automaticity transferred to the novel stimuli that participants categorized during the transfer session. First, consider the IDT condition. My results strongly suggest that automaticity transferred in this condition. In particular, there was no effect of the dual task on either accuracy or RT, whereas switching the locations of the response buttons decreased accuracy and increased RT. Both of these results are classic criteria of automatic responding (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977).

Next consider the RDT condition. Switching the response buttons decreased accuracy and increased RT, which is symptomatic of automatic responding. However, the dual-task results suggest a contradictory conclusion. Although the dual task had no effect on accuracy, it did significantly increase RT – by more than 100 ms. At first glance, it might seem that this interference could have been caused by a surprise effect – that is, that the surprise of seeing stimuli with novel values on the relevant dimension caused participants to respond more slowly. However, closer examination makes this hypothesis easy to reject. Most critically, Nosofsky (1991) reported that surprise effects of this type disappear after only two stimulus presentations. In Nosofsky's experiment, participants learned a one-dimensional categorization rule similar to the one used here. The stimuli were circles that varied in size and the orientation of a radial line. The single relevant dimension was size. After a training period, participants completed several transfer blocks in which a few trials included stimuli that were much larger than any seen during training. On the first two such trials, RT was significantly greater than on trials when the largest training stimuli were presented. But on the third and fourth such trials, responding was faster to these novel transfer stimuli than to any other stimuli. Therefore, the surprise effect persisted for only two trials. The RDT dual-task block included 200 trials, and Figure 3.7 shows that the dual-task interference persisted for all 200 trials – far longer than any documented surprise effect. Figure 3.7 does show that the dual-task interference was largest during the first 40 trials, and the Nosofsky (1991) results suggest that surprise might have contributed to this effect. Even so, Figure 3.7 shows that after 180 trials of practice and long after there was any possibility that participants were still surprised by the stimuli, there was still a dual-task interference in the RDT condition of around 100 ms.

The classical interpretation of the dual-task interference that I observed in the RDT condition is that categorization was dependent on working memory and executive attention during the RDT dual-task trials, and therefore was no longer automatic (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977). In fact, there is direct evidence linking dual-task interference to the "overloaded recruitment" of PFC working memory units (Watanabe & Funahashi, 2014).

On the other hand, the conclusion that automaticity did not transfer in the RDT condition requires more careful analysis because Hélie, Waldschmidt, and Ashby

(2010) concluded that the same qualitative pattern of results supported automaticity. Specifically, they reported that after 20 sessions of training on essentially the same category structure used here, and with the same stimuli, a similar simultaneous dual task had no effect on accuracy but significantly increased RT. They concluded from this result that, despite the RT interference, responding was automatic. What justifies a different conclusion here?

I believe that a number of results suggest that automaticity did not transfer in my RDT condition. First, if the dual-task interference on RT in the RDT condition occurred despite automatic responding, then the same interference should have been apparent in both conditions. However, I found no effect of the dual task on RT (or accuracy) in the IDT condition. This is especially noteworthy because the RDT categories were more widely separated than the IDT categories (i.e., see Figure 3.1). Because of this greater separation, the stimuli in the RDT categories were objectively easier to categorize than the stimuli in the IDT categories. Despite this difficulty difference, the simultaneous dual task interfered more with the easier RDT categories than with the more difficult IDT categories, which strongly suggests that RDT responding was not automatic.

Second, the absence of a dual-task interference on accuracy can not be taken as evidence of automatic responding. When a dual task is introduced on the very first trial of initial training, it significantly impairs learning, in the sense that accuracy is lower at every point of training than in a single-task control group (Waldron & Ashby, 2001; Zeithamova & Maddox, 2006). However, in the present experiment, there is nothing left to learn during the dual-task transfer blocks. Rather than learn a rule, participants only have to apply a well-learned and highly practiced rule. The Kovacs et al. (2021) model predicts that participants will be able to do this accurately, regardless of whether they respond automatically, or whether they respond by appealing back to the learned rule.

Third, there are a number of reasons that the dual-task RT interference reported by Hélie, Waldschmidt, and Ashby (2010) is more consistent with automaticity than with controlled rule application. First, Hélie, Waldschmidt, and Ashby (2010) gave no RT instructions to their participants, and as a result there is no reason to believe they were responding as quickly as possible. In contrast, in the present experiment, participants were instructed to respond as quickly as possible without sacrificing accuracy. Second, Hélie, Waldschmidt, and Ashby (2010) found an identical RT interference in an information-integration (II) categorization condition that is known to recruit procedural learning and memory, rather than rule learning. This is important because a dual task does not interfere with II category learning, even during the first session of training (Waldron & Ashby, 2001; Zeithamova & Maddox, 2006). Therefore, the RT interference in the II condition is inconsistent with either automatic or controlled responding, and instead suggests that the identical RT interference that Hélie, Waldschmidt, and Ashby (2010) observed in all conditions might have been an artifact caused by some unrelated design feature. One possibility is that participants were given no RT instructions, but another possibility concerns the slightly different timing used in the two studies. In both studies, the Stroop digits were displayed first, followed by a blank screen, followed by the categorization stimulus. Participants then made their categorization response, followed by their dual-task response. In the current experiment, the digits were displayed for 1 sec and the blank screen lasted for 300 ms. Therefore, participants had 1,300 ms to encode the sizes and values of the Stroop digits before responding to the categorization stimulus. In the Hélie, Waldschmidt, and Ashby (2010) experiment, the digits were displayed for 200 ms and the blank screen lasted for 100 ms, so participants only had 300 ms to encode the Stoop digits. Therefore, one hypothesis is that 300 ms was insufficient to complete this encoding and as a result, dual-task encoding persisted after the categorization stimulus was presented, thereby delaying the categorization RT.[2]

In summary, I believe that the best account of my RDT results is that the button-switch results are consistent with automaticity, whereas the dual-task results are consistent with controlled responding, and therefore a loss of automaticity. Interestingly,

[2] I thank Sebastien Hélie (personal communication) for suggesting this account.

this is exactly the pattern of results predicted by the Kovacs et al. (2021) model. Recall that this model predicts that rule-guided behaviors are initially triggered by the application of explicit rules, which are represented primarily in PFC, but after the behaviors become automatic they are initiated by projections from the stimulus representations in visual cortex directly to the relevant motor representations in PMC. Therefore, a change in the values of the relevant stimulus dimension should activate representations in visual cortex that project to untrained synapses in PMC. As a result, automatic responding is lost. Even so, the correct rule representation remains in PFC, so accuracy remains high. The cost though, is that suddenly relying on PFC makes the categorization susceptible to dual-task interference. On the other hand, the model also predicts that the projections from PMC to primary motor cortex are activated anytime a response is triggered, regardless of whether the PMC units are activated by direct projections from visual cortex (after automaticity) or by rule units in PFC (before automaticity and during transfer). Therefore, the model predicts a button-switch interference because of the 8,400 previous button presses that participants made in this task.

The model does not make strong predictions about the results of the IDT condition – primarily because it does not completely describe the nature of the stimulus representations that are used to activate units in PMC. Certainly a change in values on the relevant stimulus dimension would cause the stimulus representations to change. But the model makes no predictions about whether a change in values of the irrelevant dimension will cause the stimulus representations to change. There are two clear alternatives. First, the stimulus representations used to select responses in one-dimensional categorization tasks could be gestalts. In this case, the model makes the same predictions in both conditions, because the stimuli changed between training and transfer in both conditions. The second possibility though, is that selective attention filters out irrelevant stimulus information, in which case the stimulus representations used to select responses depend only on values on the relevant stimulus dimension. In this case, the stimulus representations that were projected to PFC and PMC in the IDT condition were identical during training and transfer, so the model predicts that automatic responding will transfer to the novel stimuli. My results support this latter hypothesis. In the IDT condition, the dual-task and button-switch results were both consistent with automaticity – that is, there was no dual-task interference on either accuracy or RT, and the button-switch interference was significant for both dependent measures.

The sample sizes in the RDT and IDT conditions were relatively modest (14 and 15, respectively), which raises the question of whether Experiment 1 was sufficiently powered. Unfortunately, computing power for the appropriate GLMMs is statistically challenging, not only because of the multiple factors included in the experiment, but also because accurate power estimation requires knowledge of both the within- and between-participant variability. As a result, the standard approach is to estimate power from thousands of simulated data sets (e.g., Kumle, Võ, & Draschkow, 2021), and even then, these estimates are only valid if all the sources of variance are correctly specified. Because I know of no prior literature that could be used to estimate between-participant variability, I did not attempt these simulations. However, there are several reasons why I believe that Experiment 1 was sufficiently powered. First, although the most critical statistical analyses were restricted to data collected during the final transfer session, each participant completed 14 prior sessions that included a total of 8,400 trials. This extensive training strongly decreases within-participant variability in both accuracy and RT (e.g., Hélie, Waldschmidt, & Ashby, 2010), which means that my design should be more powerful than the typical categorization experiment with the same number of participants that excludes the extensive prior training. Second, the Bayes factors show that the evidence supporting the critical RT interaction was extreme, and power analyses are most critical when interpreting nonsignificant effects.[3] Third, Experiment 2 tests a prediction that follows directly from my interpretation of the Experiment 1 results. As we will see, that prediction was strongly confirmed, which increases confidence in my interpretation of the Experiment

[3] If an effect is nonsignificant, then the only possible error is a type 2 error, and power is one minus this probability.

1 results.

Chapter 4

Experiment 2: Are Abstract Rules or SR Associations Automatized?

4.1 Introduction

The results of Experiment 1 suggest that automatic rule-guided behaviors are not initiated by some abstract verbal rule, but rather directly by the visual stimulus – and more specifically, only by the relevant dimension(s) of the visual stimulus. This conclusion seems to conflict with results reported by Roeder and Ashby (2016), who concluded that abstract rules are automatized in RB categorization tasks. The experimental design used by Roeder and Ashby (2016) and a summary of their results are shown in Figure 4.1. Each participant in this experiment completed 21 sessions that included 7 consecutive 3-day cycles. During days 1 and 2 of each cycle, participants practiced on the primary categories shown in panel (a) of Figure 4.1, whereas on the third day of each cycle they practiced the secondary categories. At the beginning of each session, participants were told whether the categories that day were primary or secondary, although they were never given any other instructions about the category structures or about what categorization strategy they should use. Note that the optimal strategy on the primary categories is a logical disjunction: "Respond A if the stimulus has a small value on dimension 1 or if the stimulus has a large value

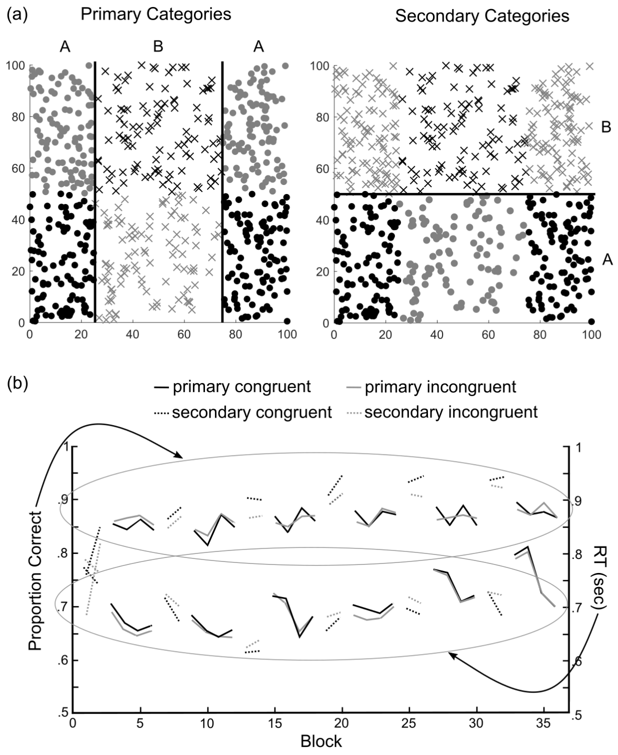

Figure 4.1: (a) Categories used in the rule-based condition of the experiment reported by Roeder and Ashby (2016). Congruent stimuli that maintained their same category assignment on primary and secondary days are shown in black, whereas incongruent stimuli that switched assignments are shown in gray. (b) Proportion corrects and RTs over the first 20 experimental sessions of the experiment. This figure reprinted from "On what it means to automatize a rule" by Kovacs, P., & Ashby, F. G., *Cognition.* Copyright 2022 with permisssion from Elsevier.

on dimension 1; otherwise respond B." In contrast, for the secondary categories the optimal strategy is a simple one-dimensional rule.

An examination of panel (a) of Figure 4.1 shows that half the stimuli changed category membership on days when the secondary categories were practiced and half the stimuli retained their primary category assignments. The stimuli that retained the same category assignment on all days, called congruent stimuli, are denoted in Figure 4.1 by black symbols, whereas stimuli that switched assignments, called incongruent stimuli, are denoted by gray symbols.

The key data-analysis question was whether performance differed on congruent and incongruent stimuli. If an abstract rule is automatized then there should be no difference because the rules on primary and secondary days are different. However, if stimulus-response associations are automatized then performance should be worse on incongruent stimuli, which is exactly what Roeder and Ashby (2016) observed in a separate group of participants who practiced on II categories that are known to recruit procedural learning and memory systems. The RB results are shown in the bottom panel of Figure 4.1. Note that on primary days, there was no difference in accuracy or RT between congruent and incongruent stimuli, and on this basis, Roeder and Ashby (2016) concluded that abstract rules are automatized in RB tasks.

However, on further reflection, the Roeder and Ashby (2016) results do not necessarily conflict with the results of my Experiment 1. The Experiment 1 results suggest a refinement of the Kovacs et al. (2021) Figure 3.3 model in which the projections from visual cortex to PFC and PMC are restricted to visual representations of the relevant stimulus dimension(s) only. The Roeder and Ashby (2016) primary and secondary categories had different relevant dimensions. Therefore, this hypothesis predicts that the visual projections on primary and secondary days will be from different visual units onto different synapses in PFC and PMC and therefore practicing different stimulus-response associations on incongruent stimuli during secondary days will not interfere with associations formed on primary days. My hypothesis is that, from the perspective of PMC, completely different stimuli were used on primary and secondary days and therefore, there were no stimuli in the Roeder and Ashby (2016) study that switched response assignments.

Experiment 2 tests this prediction by replicating the design of Roeder and Ashby (2016), except with category structures for which the revised Kovacs et al. (2021) model predicts that the incongruent stimuli should cause interference. The stimuli and categories I used in Experiment 2 are shown in Figure 4.2. As in Roeder and Ashby (2016), Experiment 2 included seven consecutive 3-day cycles. On the first two days of each cycle, participants practiced the primary categories shown in Figure 4.2. On the third day of each cycle, they practiced the secondary categories. At the beginning of each day, participants were instructed about whether they would be practicing the primary or secondary categories during that session, but they were

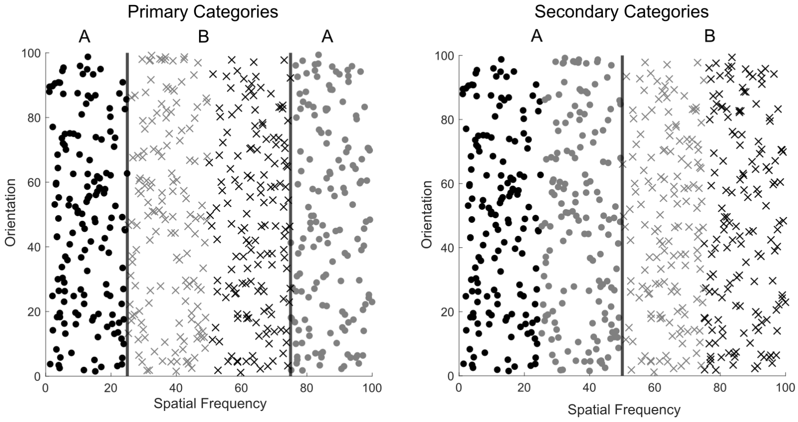

Figure 4.2: Stimuli and category structures used in Experiment 2. Congruent stimuli that maintained their same category assignment on primary and secondary days are shown in black, whereas incongruent stimuli that switched assignments are shown in gray. This figure reprinted from "On what it means to automatize a rule" by Kovacs, P., & Ashby, F. G., *Cognition*. Copyright 2022 with permisssion from Elsevier.

never given any instructions about the nature of the categories.

Note that, as in the Roeder and Ashby (2016) experiment, half the stimuli in Experiment 2 switch their category assignments on primary and secondary days, and half maintain their same assignment on all days. Also note that the primary categories are identical in the two experiments, and in both experiments the secondary categories are separated by a simple one-dimensional rule. However, unlike Roeder and Ashby (2016), the same stimulus dimension is relevant on all days in my Experiment 2. Therefore, the revised Kovacs et al. (2021) model predicts that, in contrast to the results of Roeder and Ashby (2016), performance should be worse on incongruent stimuli than on congruent stimuli.

4.2 Methods

Participants

Thirty-one undergraduate students at the University of California, Santa Barbara participated in this experiment in exchange for course credit.

Stimuli and Apparatus

Due to COVID restrictions, participants performed the experiment at home on their own home computers. As in Experiment 1, all stimuli were circular sine-wave gratings that varied across trials in spatial frequency (i.e., bar width) and bar orientation. Each stimulus was defined by a set of points (x_1, x_2) sampled from a 100×100 stimulus space and converted to a disk using the following equations: spatial frequency $=$ $.1x_1 + 0.25$ cycles per disk and orientation $= .9x_2$ degrees counterclockwise rotation from horizontal.[1]

There were two different kinds of sessions: primary and secondary. The experiment included seven 3-day blocks, during which they practiced the the primary categories on the first two days and the secondary categories on the third day. The secondary session was omitted from the last cycle, so the entire experiment included 20 sessions over 20 nearly consecutive days.

The stimuli were the same as in Experiment 1, as were the events that occurred on each trial, and their timing. The category structures are shown in Figure 4.2. On primary days, the optimal rule was a 1D disjunctive rule. On secondary days, the optimal rule was a simple 1D rule. In both sessions, the single relevant stimulus dimension was spatial frequency.

During primary sessions, stimuli in category A were uniformly distributed (in the 100×100 space) in two distinct intervals [0, 25] and [75, 100] on the spatial frequency dimension and [0, 100] on the orientation dimension. Stimuli in category B were uniformly distributed (in the 100×100 space) in the interval [25,75] on the spatial frequency dimension and [0, 100] on the orientation dimension. During secondary sessions, stimuli in category A were uniformly distributed (in the 100×100 space) in the interval [0, 50] on the spatial frequency dimension and [0, 100] on the orientation dimension. Stimuli in category B were uniformly distributed (in the 100×100 space) in

[1] Note that the transformation to spatial frequency was nonlinear in Experiment 1 and linear in Experiment 2. This is because the Experiment 1 IDT transfer stimuli differed from the training stimuli in spatial frequency, so the range of perceived bar widths was much greater in Experiment 1 than in Experiment 2. In fact, the range was great enough that I felt it important to account for the nonlinear relationship between spatial frequency and perceived bar width (Treutwein, Rentschler, & Caelli, 1989).

the interval [50,100] on the spatial frequency dimension and [0, 100] on the orientation dimension.

Procedure

The trial-by-trial procedures were identical to Experiment 1, except participants were informed that they would be participating in two different kinds of sessions, primary and secondary. They were instructed that the optimal strategy would be different on the secondary days, but they were given no instructions about the nature of the categories or about the type of strategies they should employ. At the beginning of each session, participants were informed about whether they would practice primary or secondary categories on that day.

4.3 Results

Figure 4.3 shows the accuracy results for each 300-trial block and Figure 4.4 shows the means of the median RTs. Data from the first two days are omitted because at this point in the experiment – that is, before the first secondary session – there were no incongruent stimuli. Note that accuracy is considerably higher for congruent stimuli in every session and RT is lower. A comparison back to Figure 4.1 shows that these results are strikingly different from those of Roeder and Ashby (2016).

To test these conclusions statistically, I used the same GLMM analyses as in Experiment 1. I ran these analyses separately for all the data combined, the data only from primary sessions, and the data only from secondary sessions. The results were similar in all cases, but the results from the primary sessions are most important because the number of primary sessions (i.e., 14) was chosen to ensure that responding had become automatic by the end of training (according to results of Hélie, Waldschmidt, & Ashby, 2010). As a result, this section focuses on the results from the primary categories only.

The accuracy analyses are shown in Table 4.1 and the RT analyses are shown in

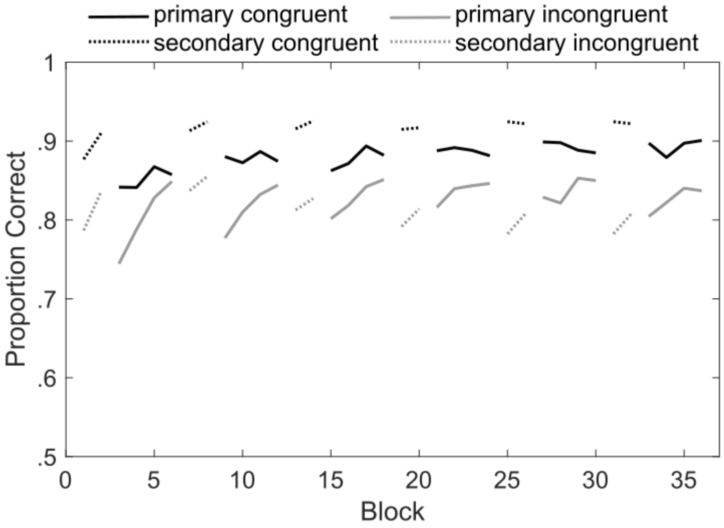

Figure 4.3: Proportion correct in Experiment 2 shown separately for congruent and incongruent stimuli on primary and secondary days. This figure reprinted from "On what it means to automatize a rule" by Kovacs, P., & Ashby, F. G., *Cognition*. Copyright 2022 with permisssion from Elsevier.

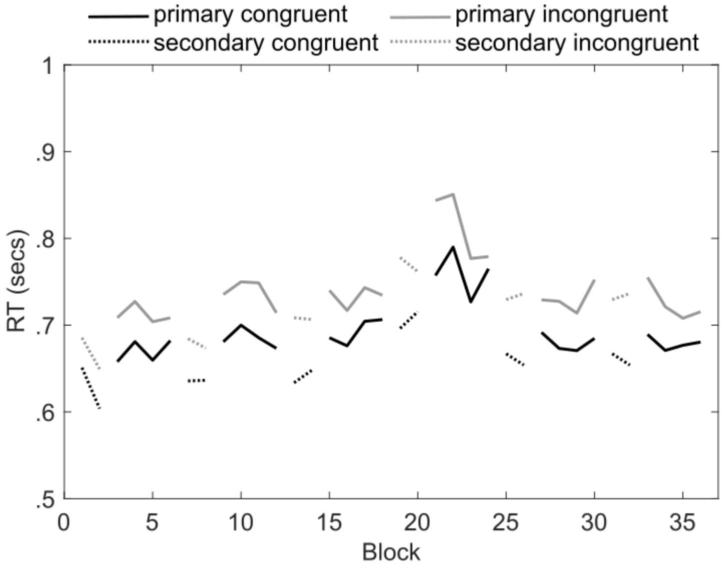

Figure 4.4: Means (across participants) of the median RTs in Experiment 2 shown separately for congruent and incongruent stimuli on primary and secondary days. This figure reprinted from "On what it means to automatize a rule" by Kovacs, P., & Ashby, F. G., *Cognition*. Copyright 2022 with permisssion from Elsevier.

Model	Terms	Log L	BIC	BF
Null	β_0	93366	186745	1
Congruence	$\beta_0 + C$	92676	185376	2.3e297
Session	$\beta_0 + S$	93086	186321	2.0e92
CongSess	$\beta_0 + C + S$	92394	184947	2.8e390
Full	$\beta_0 + C + S + (C \times S)$	92334	184963	9.6e386

Table 4.1: GLMM results for the accuracy data from the Experiment 2 primary sessions. Table reprinted from "On what it means to automatize a rule" by Kovacs, P., & Ashby, F. G., *Cognition*. Copyright 2022 with permisssion from Elsevier.

Table 4.2. In both cases, I tested models that included a main effect of session, a main effect of congruence (congruent stimuli versus incongruent stimuli), and an interaction. As described in the Methods, due to COVID restrictions, all participants performed the experiment at home on their personal computers. As a result, there were more frequent extreme RT outliers than in typical laboratory experiments. Therefore, as a conservative approach, I excluded from the RT analyses all RTs longer than 5 seconds. Figure 4.4 shows that the median RTs were all well below 1 second, so any RT ¿ 5 seconds was almost surely due to some irrelevant distraction. This criterion excluded 1.2% of the RTs from the primary sessions (2638 out of 223,200 total RTs).

Table 4.1 shows that for the accuracy analysis, the best-fitting model (CongSess) included both main effects but no interaction. The Bayes factors (BF) suggest that the evidence for both main effects is extreme, as is the evidence that there is no interaction (Lee & Wagenmakers, 2014). An examination of Figure 4.3 suggests that the main effect of congruency is driven by the higher accuracy for congruent stimuli that was evident in every experimental block. Note that this same difference also occurred with the secondary categories, where it was even more extreme. In fact, the main effect of congruency was highly significant even when I analyzed data from all sessions together and when I analyzed data from the secondary sessions only.

Table 4.2 shows that for the RTs, the best-fitting model again included both main effects but no interaction. And as with the accuracy analysis, the Bayes factors (BF) suggest that the evidence for both main effects is extreme, as is the evidence that there is no interaction. Figure 4.4 shows that the main effect of congruency is driven by the faster RTs for congruent stimuli that was evident in every experimental block,

Model	Terms	Log L	BIC	BF
Null	β_0	1720526	3441076	1
Congruence	$\beta_0 + C$	1720494	3441024	2.5e11
Session	$\beta_0 + S$	1720171	3440503	5.0e124
CongSess	$\beta_0 + C + S$	1720139	3440450	1.4e136
Full	$\beta_0 + C + S + (C \times S)$	1720131	3440569	2.2e110

Table 4.2: GLMM results for the RTs from the Experiment 2 primary sessions. Table reprinted from "On what it means to automatize a rule" by Kovacs, P., & Ashby, F. G., *Cognition.* Copyright 2022 with permisssion from Elsevier.

and that this same effect was seen with both primary and secondary categories.

The accuracy and RT results support the predictions of the revised Kovacs et al. (2021) model only if participants were using the disjunction rule shown in Figure 4.2 on primary days. For example, my labeling of stimuli as congruent or incongruent assumed this strategy. High accuracy and low RT is possible with multiple strategies, so a strategy analysis is needed to supplement my GLMM analyses of accuracy and RT. For this reason, I fit decision-bound models to the responses of each individual participant from each of their 20 experimental sessions. The models, which are described in the Appendix, were the same as the models used in Experiment 1, except the rule-based models also included a model that assumed participants used a disjunction rule.

Each of the 31 participants completed 14 sessions with the primary categories and 6 sessions with the secondary categories. Therefore, I fit all the models to 434 sets of primary session data (31 × 14) and 186 sets of secondary session data (31 × 6). The results are summarized in Table 4.3. The disjunctive classifier assumed a disjunction rule of the type that is optimal on primary days, the "1D: bar width" model assumed a one-dimensional rule of the type that is optimal on secondary days, the procedural strategy model assumed that perceptual information from both dimensions was (predecisionally) integrated, and the guessing models assumed random guessing (see the Appendix for details). Note that on the critical primary days, the participants used a disjunction rule of the optimal type during almost every session. This result greatly increases confidence in my interpretation of the GLMM results.

Several points are worth noting about the results from the secondary sessions.

Model	Number of Sessions	Percentage
Primary Sessions		
Disjunctive Classifier	431	99.3
1D: Bar Width	3	0.7
Procedural Strategy	0	0
Guessing	0	0
Secondary Sessions		
Disjunctive Classifier	60	32.3
1D: Bar Width	122	65.6
Procedural Strategy	4	2.2
Guessing	0	0

Table 4.3: Decision bound modeling results for Experiment 2. Table reprinted from "On what it means to automatize a rule" by Kovacs, P., & Ashby, F. G., *Cognition.* Copyright 2022 with permisssion from Elsevier.

First, participants almost always used a rule-based strategy (i.e., on 97.8% of the sessions). Second, participants used a rule of the optimal type (i.e., a one-dimensional rule on bar width) on most of the sessions (i.e., about two-thirds). Third, the disjunctive classifier that was optimal on primary days provided the best fit on about one-third of the sessions. This is not too surprising since participants had twice as much practice with the disjunction rule, and by the end of training they had automatized this rule. Note though, from Figure 4.2, that if the disjunction rule was used on every trial during secondary sessions, accuracy would be only 50%, whereas Figure 4.3 shows that accuracy on secondary sessions averaged about 85% correct. A closer examination of the secondary sessions for which the disjunctive classifier provided the best fit indicated that in almost every case, only a few responses were incompatible with the optimal one-dimensional rule. These few responses allowed the disjunctive classifier to fit better, even though the great majority of responses were compatible with a one-dimensional rule.[2] Therefore, I believe that my results suggest that virtually all participants used a one-dimensional rule of the optimal type on all but a few trials on each secondary day. However, about a third of the secondary sessions included a few trials in which participants inadvertently applied the more well-practiced

[2] The maximum-likelihood-based goodness-of-fit statistic that I used (i.e., BIC) assigns an extreme penalty to any response that is incompatible with the assumed decision rule (e.g., to any B response in the presumed A response region), and this penalty gets much worse the further the discrepant response is from the decision boundary.

disjunction rule.

4.4 Discussion

Although the design of Experiment 2 was nearly identical to the design used by Roeder and Ashby (2016), the results of the two experiments were strikingly different. Whereas Roeder and Ashby (2016) found no difference on primary days in either accuracy or RT for congruent versus incongruent stimuli, I found that responding was more accurate and faster for congruent than for incongruent stimuli. A comparison of Figures 4.1 and 4.2 shows that the two experiments used identical primary categories, and in both experiments the secondary categories required a simple one-dimensional decision rule. The only difference was that in the Roeder and Ashby (2016) experiment, the relevant dimension on secondary days was irrelevant on primary days, whereas in my Experiment 2, the same stimulus dimension was relevant on all days.

My results are inconsistent with the conclusions of Roeder and Ashby (2016) that participants automatize an abstract rule in RB tasks. In both experiments, the rule on primary and secondary days was different, so if participants had automatized a rule, the two experiments should have yielded identical results. On the other hand, the results of both experiments are predicted by the revised version of the Kovacs et al. (2021) model in which the projections from visual cortex to PFC and PMC are restricted to visual representations of the relevant stimulus dimension only (see Figure 3.3). In the Roeder and Ashby (2016) experiment, the relevant dimension changed from primary to secondary days, and as a result the model predicts that the visual input to PMC was fundamentally different on primary and secondary days. In other words, the model predicts that the effective stimuli were completely different on primary and secondary days, and as a result, the network mediating automaticity did not recognize any stimuli as being incongruent. In contrast, in my Experiment 2, because the same stimulus dimension was relevant on primary and secondary days, the model predicts that the visual projections into PMC were the same on every day, and therefore performance was worse on incongruent stimuli because of the interference that was caused by practicing competing motor responses on primary and secondary days.

Chapter 5

General Discussion

This dissertation proposes a novel theory of how rule-guided behaviors become automatized and describes the results of two extensive experiments that tested novel predictions of that theory. The experiments included a combined total of 633,000 categorization trials. The experiments investigated the nature of what is automatized after lengthy practice with a rule-guided behavior by testing novel predictions of a recent neurocomputational model (Kovacs et al., 2021). The results of both experiments suggest that an abstract rule, if interpreted as a verbal-based strategy, was not automatized during training, but rather the automatization linked a set of stimuli with similar values on one visual dimension to a common motor response.

It is important to note, however, that my results do not suggest that participants no longer had easy access to an abstract rule after automaticity developed. In fact, the Kovacs et al. (2021) model predicts that access to the abstract rule is always available via projections from visual cortex to PFC (see Figure 3.3). However, the model predicts that after automaticity has developed, the behavior is not initiated by this indirect path to PMC, but rather by a faster, direct projection from visual cortex, and that it is only this direct projection that links stimuli with similar values on one visual dimension to a common motor response.

My results clarify a number of puzzling results in the literature. First, categorization tasks, like the one used here, in which the optimal bound is a vertical or horizontal line (and in which the stimulus dimensions are perceptually separable) are known as rule-based (RB) tasks in the literature. These are often compared to information-integration (II) tasks that are identical, except the categories are rotated 45° in stimulus space (so the separating decision bound is now diagonal). One curious, and previously unexplained result is that capuchin and macaque monkeys both learn these one-dimensional RB categories more quickly and to a higher asymptotic accuracy than the rotated II categories (Smith et al., 2010; Smith, Crossley, et al., 2012; Smith et al., 2015). Humans show an even more pronounced RB advantage than macaques, whereas pigeons and rats learn rotated RB and II category structures at exactly the same rate (Ashby et al., 2020; Broschard et al., 2019; Smith, Berg, et al., 2012; Smith et al., 2011). Furthermore, the RB advantage shown by humans (and monkeys) is not because of an inherent difference in task difficulty, but rather because humans learn the two tasks in qualitatively different ways (Ashby et al., 2020).

One leading account of human category learning, called COVIS, proposes that humans learn RB categories by experimenting with simple, explicit rules and that in II tasks they instead rely on procedural learning (Ashby et al., 1998; Ashby & Waldron, 1999). The COVIS acronym stands for COmpetition between Verbal and Implicit Systems because the original proposal was that the learning of rules depends on verbal strategies. However, the superior performance of macaques in RB versus II tasks is strong evidence that verbalization is not a necessary condition for the RB advantage. So why are monkeys better at RB tasks than in rotated II tasks?

The present results offer an answer to this question. Monkeys are better at one-dimensional RB tasks than in rotated II tasks because they can allocate executive attention selectively to the single relevant stimulus dimension in the RB task, and this ability is not language dependent. In fact, the evidence is good that PFC plays a key role in this type of top-down selective attention (e.g., Desimone & Duncan, 1995). Macaque monkeys have a well-developed PFC, and so it is not surprising that there is much neural evidence for feature-based selective attention in monkeys (e.g., Fuster, 1990; Maunsell & Treue, 2006). Therefore, my results suggest that the most fundamental difference between rotated RB and II tasks may not so much be that language facilitates RB learning, but rather that selective visual attention does, whereas this attentional ability provides no benefit in II tasks.

Second, my results offer an alternative interpretation of the many reports of rule-sensitive neurons in PMC (Muhammad et al., 2006; Wallis & Miller, 2003; Vallentin et al., 2012). These studies reported single-unit recordings from neurons in PMC that fired when a monkey applied one of two categorization rules. Furthermore, these neurons did not fire when the alternative rule was applied, and the neural responses were the same regardless of which stimulus was shown and what cue was used as a signal to the animal about which rule to apply. Neurons with similar firing properties have frequently been found in PFC (Asaad et al., 2000; Hoshi et al., 2000; White & Wise, 1999), but finding such neurons in PMC is somewhat surprising, given that the primary function of PMC has long been thought to be the selection of motor actions. My results suggest that rule-sensitive neurons in PMC might not be implementing a categorization rule as it is commonly interpreted, but rather linking a set of stimuli with similar values on one visual dimension to a common motor response.

Third, my results suggest that the automatization of rule-guided behaviors and procedural skills might not be fundamentally different. Ashby et al. (2007) proposed that the automatic execution of procedural skills is mediated entirely within cortex and that the development of automaticity is associated with a gradual transfer of control from the basal ganglia circuits that mediate initial procedural learning to cortical-cortical projections from the relevant sensory areas directly to units in areas of PMC that initiate the behavior. According to this account, a critical function of the basal ganglia is to train purely cortical representations of automatic procedural behaviors (Hélie et al., 2015). The Kovacs et al. (2021) model proposes a similar account of the automatization of rule-guided behaviors, except for two key differences. First, in the case of rule-guided behaviors, the PFC trains the automatic cortical representations, rather than the basal ganglia. And second, the PMC targets are rule-sensitive units, rather than units associated with a specific motor goal. Despite these differences, 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 PMC. The current results reduce the differences between these two theories because they suggest that the PMC targets in the two models are not fundamentally different. For both procedural and rule-guided behaviors, the PMC targets link sensory representations to motor behaviors. My results suggest that the only real difference might be in the nature of the visual representations – gestalts in the case of procedural skills and single stimulus dimensions in the case of rule-guided behaviors.

Finally, at a more speculative level, my results might also be used to reflect on possible developmental origins of rule use. If rules are only abstract sets of verbal instructions, then their learning must necessarily be language dependent. If so, then procedural learning that is mediated by basal ganglia circuits can play at most a minor role in their acquisition. However, my results elevate the role that selective attention might play in this process and, together with the capuchin and macaque results (Smith et al., 2010; Smith, Crossley, et al., 2012; Smith et al., 2015), suggest that rule automatization might not necessarily even require language. Furthermore, the fact that my results reinforce neuroscience theories of automaticity that propose similar accounts for behaviors that are initially rule-guided versus mediated by procedural learning, suggests that rules might develop from an initial period of procedural learning. Together, all of these considerations suggest an intriguing hypothesis that might be worth developing and testing. First, initial rule use begins with a period of procedural learning that is facilitated by dopamine-mediated reinforcement learning in the basal ganglia (as described e.g., by Ashby & Crossley, 2010 and Cantwell et al., 2015). Second, this process simultaneously trains cortical-cortical projections from the visual areas that respond to the stimulus to the relevant PMC targets (as proposed by Ashby et al., 2007). Finally, PFC selective-attention circuits directed at these visual representations begin to filter out irrelevant stimulus information (e.g., J. Feldman, 2021), leading to an end result in which the PMC targets receive input only about the relevant stimulus dimension.

In summary, my results suggest that the common interpretation that rule-guided behavior is mediated by a verbal-based strategy that implements a set of explicit instructions, is valid, at most, only for a period of initial learning. After rule-guided behaviors are practiced long enough to become automatic, they appear to no longer be mediated by anything resembling a rule, but instead to be triggered directly by the visual stimulus. Similar proposals have been made for automatic behaviors that are initially acquired via procedural learning, so my results suggest that behaviors that are acquired via rule or procedural learning, although initially depending on very different neural networks, may be mediated in almost identical ways after they become automatized. The only real difference appears to be that in the case of rule-guided behaviors, top-down selective attention whittles away irrelevant visual information, in the sense that the automatic behavior is triggered by visual representations that depend only on relevant stimulus information.

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. (2010). Interactions between declarative and procedural-learning categorization systems. *Neurobiology of Learning and Memory*, *94*(1), 1-12.

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. doi: http://dx.doi.org/10.1016/j.jmp.2011.04.003

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., Smith, J. D., & Rosedahl, L. A. (2020). Dissociations between rule-based and information-integration categorization are not caused by differences in task difficulty. *Memory & Cognition*, *48*, 541–552.

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* (pp. 307–347). 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.

Broschard, M. B., Kim, J., Love, B. C., Wasserman, E. A., & Freeman, J. H. (2019). Selective attention in rat visual category learning. *Learning & Memory*, *26*(3), 84–92.

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.

Casale, M. B., Roeder, J. L., & Ashby, F. G. (2012). Analogical transfer in perceptual categorization. *Memory & Cognition*, *40*(3), 434-449.

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., 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.

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.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*(1), 193–222.

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.

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.

Feldman, J. (2021). Mutual information and categorical perception. *Psychological Science*, *32*(8), 1298–1310.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2003, jun). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *Journal of Neuroscience*, *23*(12), 5235–46. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/12832548`

Fuster, J. M. (1990). Inferotemporal units in selective visual attention and short-term memory. *Journal of Neurophysiology*, *64*(3), 681–697.

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. Retrieved from `http://doi.apa.org/getdoi.cfm?doi=10.1037/xhp0000063` doi: 10.1037/xhp0000063

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. *The 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. doi: 10.1162/jocn_a_00794

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.

Kovacs, P., Hélie, S., Tran, A. N., & Ashby, F. G. (2021). A neurocomputational theory of how rule-guided behaviors become automatic. *Psychological Review*, *128*(3), 488–508.

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.

Kumle, L., Võ, M. L.-H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*, *53*(6), 2528–2543.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course.* Cambridge University Press.

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. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, *53*(1), 49-70.

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.

Maunsell, J. H., & Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neurosciences*, *29*(6), 317–322.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*(1), 167–202.

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. *The 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. (1991). Typicality in logically defined categories: Exemplar-similarity versus rule instantiation. *Memory & Cognition*, *19*(2), 131–150.

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.

Qadri, M. A., Ashby, F. G., Smith, J. D., & Cook, R. G. (2019). Testing analogical rule transfer in pigeons (Columba livia). *Cognition*, *183*, 256–268.

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. *Cognitive Neuroscience, Journal of*, *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. *Cognitive Neuroscience, Journal of*, *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., Ashby, F. G., Berg, M. E., Murphy, M. S., Spiering, B., Cook, R. G., & Grace, R. C. (2011). Pigeons' categorization may be exclusively nonanalytic. *Psychonomic Bulletin & Review*, *18*(2), 414-421.

Smith, J. D., Beran, M. J., Crossley, M. J., Boomer, J. T., & Ashby, F. G. (2010). Implicit and explicit category learning by macaques (macaca mulatta) and humans (homo sapiens). *Journal of Experimental Psychology: Animal Behavior Processes*, *36*, 54-65.

Smith, J. D., Berg, M. E., Cook, R. G., Murphy, M. S., Crossley, M. J., Boomer, J., … others (2012). Implicit and explicit categorization: A tale of four species. *Neuroscience & Biobehavioral Reviews*, *36*(10), 2355–2369.

Smith, J. D., Crossley, M. J., Boomer, J., Church, B. A., Beran, M. J., & Ashby, F. G. (2012). Implicit and explicit category learning by capuchin monkeys (Cebus apella). *Journal of Comparative Psychology*, *126*(3), 294–304.

Smith, J. D., Zakrzewski, A., Johnston, J. J. R., Roeder, J., Boomer, J., Ashby, F. G., & Church, B. A. (2015). Generalization of category knowledge and dimensional categorization in humans (homo sapiens) and nonhuman primates (macaca mulatta). *Journal of Experimental Psychology: Animal Behavior Processes*, *41*, 322–335.

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. doi: 10.1016/j.neuroimage.2013.01.008

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 *Information processing in motor control and learning* (pp. 117–152). Elsevier.

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.

Treutwein, B., Rentschler, I., & Caelli, T. (1989). Perceptual spatial frequency—orientation surface: Psychophysics and line element theory. *Biological Cybernetics*, *60*(4), 285–295.

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.

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.

Watanabe, K., & Funahashi, S. (2014). Neural mechanisms of dual-task interference and cognitive capacity limitation in the prefrontal cortex. *Nature Neuroscience*, *17*(4), 601–611.

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.