Ashby, F. G., & Wang, Y.-W. (2023). Computational Cognitive Neuroscience Models of Categorization. In R. Sun (Ed.), *The Cambridge Handbook of Computational Cognitive Sciences* (pp. 400-425). Cambridge University Press.

12 Computational Cognitive Neuroscience Models of Categorization

F. Gregory Ashby & Yi-Wen Wang
University of California, Santa Barbara

1. Introduction

Categorization is the process of assigning an object or event to a class or group – typically one that is behaviorally relevant. It is a vitally important skill that is required of all animals, because it allows nutrients and prey to be approached and poisons and predators to be avoided. Interest in how humans categorize dates back at least to Aristotle. For almost all of this long history, theorizing was dominated by purely cognitive approaches. The past few decades however, have seen an explosion of new results that collectively are beginning to paint a detailed picture of the neural mechanisms and pathways that mediate human categorization. These results come from a wide variety of sources, including human behavioral experiments, animal lesion studies, single-cell recordings, neuroimaging experiments, and neuropsychological patient studies. Lagging somewhat behind this avalanche of new data has been the development of new theories that can account for the traditional cognitive results as well as for these newer neuroscience results. Even so, a number of such theories have been proposed. This chapter reviews those theories.

2. Multiple Learning Systems

An enormous literature suggests that humans have multiple learning and memory systems. For example, a Google Scholar search of publications using the terms "memory systems" returns almost a million articles. Since, by definition, learning requires that some trace of previous training episodes must exist, one obvious hypothesis is that there are as many learning systems as there are memory systems (Ashby & O'Brien, 2005). This complicates any review of categorization models because different researchers have proposed models of different category-learning systems. This can be confusing to an outsider because the models might share little in common, including the neural structures and pathways that they claim mediates category learning.

One way to discriminate among models is by attending to what type of category-learning task they focus on, because different types of tasks are thought to recruit different learning systems. And different learning systems are mediated by very different neural networks. Thus, models focusing on different systems will bear little similarity to each other. On the other hand, different neuroscience-based models of the same learning system should be highly similar because all such models are constrained by the same neuroanatomy. For example, an enormous body of evidence implicates the basal ganglia in procedural learning. As a result, any model of procedural-learning-based categorization must assign a prominent role to the basal ganglia, and since the gross neuroanatomy of the basal ganglia is well known, all such models must have a similar architecture. The primary difference among models of the same learning system will likely be that some will include more detail about some neural regions than others. Some of the more popular category-learning tasks are briefly described in the remainder of this section (for more details, see e.g., Ashby & Valentin, 2018).

2.1. Tasks that depend on declarative memory

A number of different category-learning tasks depend on declarative memory. Included in this list are rule-based (RB) tasks in which the optimal strategy is some simple rule that can be described as a Boolean expression of the stimulus values on a few stimulus dimensions. In the simplest example, only one dimension is relevant but in more complex RB tasks, the optimal strategy might be a logical conjunction – for example, the optimal rule might be to give one response if the stimulus is large on two dimensions, and otherwise to give the contrasting response.

The most widely known example of an RB categorization task is the Wisconsin Card Sorting Test (WCST; Heaton, 1981), which is a popular clinical measure that is used to detect frontal dysfunction. The test uses a deck of cards that differ in the shape, number, and color of displayed figures. On each trial, the participant is shown a card and asked to assign it to one of two unknown categories. Feedback is given after each response and the correct categorization strategy is always a simple rule that depends on only one stimulus dimension. After 10 consecutive correct categorizations, the relevant dimension is changed (without telling the participants).

Considerable evidence suggests that RB category learning depends on working memory and selective attention (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Maddox, Filoteo, Hejl, et al., 2004; Waldron & Ashby, 2001; Zeithamova & Maddox, 2006) – skills that are both thought to depend heavily on prefrontal cortex (PFC; e.g., Braver et al., 1997; Curtis & D'Esposito, 2003; Miller & Cohen, 2001). As a result, models of RB category learning will assign a prominent role to the PFC.

Categorization tasks in which the categories have some coherent structure, but in which one or more categories include a small number of exceptions also seem to recruit declarative memory (e.g., Davis, Love, & Preston, 2011).

2.2. Information-integration tasks

Information-integration (II) tasks are those in which accuracy is maximized only if information from two or more incommensurable stimulus dimensions is integrated at some pre-decisional stage (Ashby & Gott, 1988). In II tasks, similar stimuli tend to be in the same category, but the optimal strategy has no Boolean analogue. Evidence suggests that success in II tasks depends on procedural learning that is mediated largely within the striatum (Ashby & Ennis, 2006; Filoteo, Maddox, Salmon, & Song, 2005; Knowlton, Mangels, & Squire, 1996; Nomura et al., 2007; Seger & Miller, 2010). As a result, neuroscientific models of II learning will assign a prominent role to the basal ganglia.

2.3. Prototype-distortion tasks

In prototype-distortion tasks, the exemplars of each category are created by randomly distorting a single category prototype. The most widely known example uses a constellation of 7 or 9 dots as the category prototype, and the other category members are created by randomly perturbing the spatial location of each dot (Posner & Keele, 1968). Sometimes the dots are connected by line segments to create polygon-like images.

Two different types of prototype distortion tasks are common – (A, B) and (A, not A). In (A, B) tasks, two different prototype patterns are distorted to create two coherent categories. In (A, not A) tasks, which are more popular, there is only one prototype pattern that is distorted to create the exemplars of Category A. In contrast, every member of the 'not A' category is generated independently (and randomly). Thus, all Category A exemplars are similar to the prototype, and therefore also to each other, whereas the 'not A' stimuli have no coherent structure. A variety of evidence supports the hypothesis that learning in (A, not A) prototype-distortion tasks is mediated primarily within visual cortex, via the perceptual representation memory system (e.g., Aizenstein et al., 2000; Casale & Ashby, 2008; Reber & Squire, 1999; Reber, Stark, & Squire, 1998). The idea is that accurate responding can be based solely on a feeling of visual familiarity, which should be high on A trials and low on not-A trials.

2.4. Category Learning versus Automatic Categorization

Most categorization decisions made 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. And there is now considerable evidence that categorization decisions are mediated differently during initial learning and automaticity (Ashby & Crossley, 2012). To note just one example, categorization decisions that depend on working memory and executive attention during early learning are immune to dual-task interference after extended practice (Hélie, Waldschmidt, & Ashby, 2010; Schneider & Shiffrin, 1977). For this reason, different models and theories are needed to account for category learning and automatic categorization behaviors.

3. Neuroscience-Based Models of Category Learning

Currently, there are no neuroscience-based theories or models that attempt to account simultaneously for all types of categorization. In fact, the majority of models are designed to account for categorization in only one type of task. Even so, there are a few exceptions. One is provided by the COVIS theory of category learning (Ashby et al., 1998; Ashby & Crossley, 2011; Ashby, Ennis, & Spiering, 2007; Ashby & Waldron, 1999; Cantwell, Crossley, & Ashby, 2015). Briefly, COVIS postulates two systems that compete throughout learning – a frontal-based system that learns explicit rules and depends on declarative memory systems and a basal ganglia-mediated procedural-learning system. The procedural system is phylogenetically older. It can learn a wide variety of category structures, but it learns in a slow incremental fashion and is highly dependent on reliable and immediate feedback. In contrast, the declarative rule-learning system can learn a fairly small set of category structures quickly – specifically, those structures in which the contrasting categories can be separated by simple explicit rules. Thus, COVIS assumes that performance improvements in RB tasks are mediated by an explicit, rule-learning system, whereas performance improvements in II tasks are mediated by a procedural-learning system. In addition, COVIS has been extended to account for automatic categorization behaviors that were acquired initially via procedural learning (Ashby et al., 2007). On the other hand, COVIS is almost certainly incomplete because it ignores all other types of category learning. For example, it provides no account of the kind of perceptual learning thought to mediate performance improvements in (A, not A) prototype-distortion tasks.

Another model that attempts to account for diverse cognitive functions, including categorization, within a single unified framework is called Leabra (O'Reilly, Hazy, & Herd, 2016). Leabra was designed to account for tasks under executive control, so it provides accounts of RB learning, and also perhaps, prototype-distortion learning. But it makes no attempt to account for procedural learning of the type thought to dominate in II tasks. Leabra uses the same set of computational features, including recurrent connections, error-driven Hebbian learning, within-layer inhibitory competition, and sparse distributed representations, for modeling activation within different cortical regions, including visual cortex (O'Reilly, Wyatte, Herd, Mingus, & Jilk, 2013), the medial temporal lobes (Norman & O'Reilly, 2003), and PFC (Rougier & O'Reilly, 2002; O'Reilly, Noelle, Braver, & Cohen, 2002). Among the multiple tasks simulated by Leabra, the most relevant for this review are the WCST and visual object categorization, which will be discussed in later sections.

3.1. Declarative-memory-based models of categorization

3.1.1. COVIS.

As mentioned earlier, COVIS assumes that performance in RB tasks is dominated by a rule-learning system that uses declarative memory. The idea is that this system generates and tests alternative categorization rules until satisfactory performance is achieved, or until the participant gives up and decides that no acceptable rule exists. For example, the initial rule may be to "respond A if the object is large, and B if it is small." This candidate rule is then held in working memory while it is being tested. If feedback signals that this rule is incorrect, then an alternative rule is selected, and executive attention must be switched from the old to the new rule.

Figure 1 shows the neural structures that mediate performance in the COVIS rule-learning system during a trial of an RB task. The key structures in the model are the anterior cingulate (ACC), the prefrontal cortex (PFC), the head of the caudate nucleus, the medial dorsal nucleus of the thalamus (MDN), and the hippocampus. There are three separate subnetworks in this model – one that maintains candidate rules in working memory,

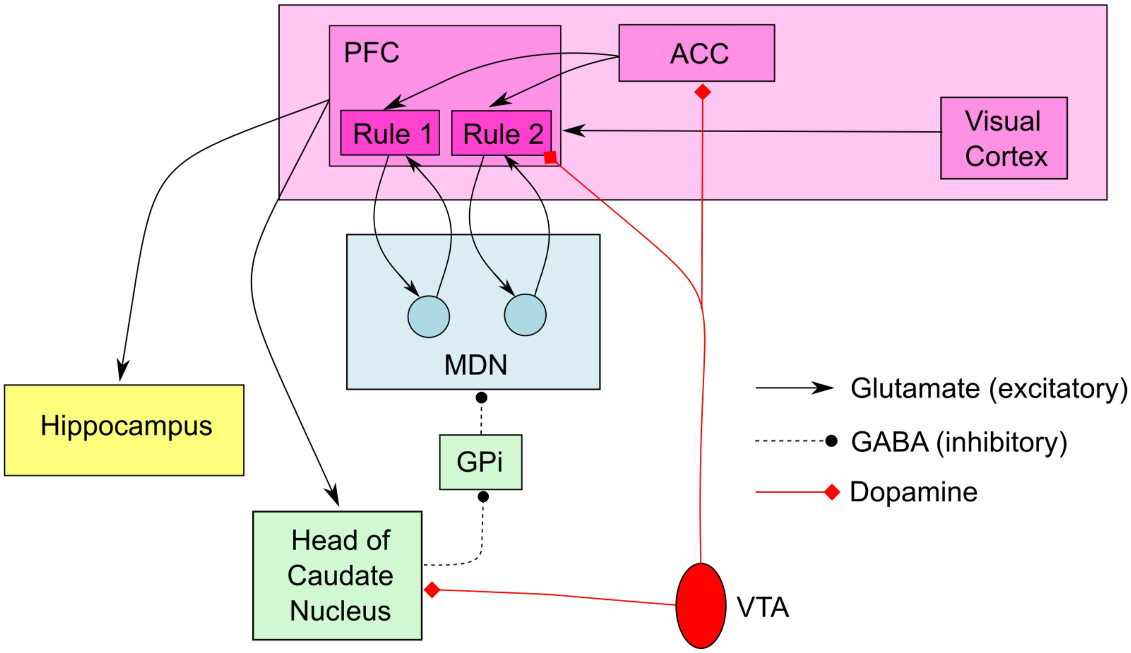

Figure 1. The COVIS declarative system (black lines = excitatory projections, green lines = inhibitory projections, teal line = dopminergic projection, VTA = ventral tegmental area, SN = substantia nigra pars compacta, CD = caudate nucleus, GP = internal segment of the globus pallidus, MDN = medial dorsal nucleus of the thalamus, PFC = prefrontal cortex, ACC = anterior cingulate cortex, HC = hippocampus).

tests those rules, and mediates the switch from one rule to another, one that generates or selects new candidate rules, and a third that consolidates memories of this selection and testing process in a long-term store. Currently, there is no computational model of the entire network. There is a biologically detailed computational model of the working memory maintenance and rule-switching network that was built from spiking neuron units like those described in Equations 2 – 5 below (Ashby, Ell, Valentin, & Casale, 2005). In contrast, the model of rule selection and rule implementation is more abstract (Ashby, Paul, & Maddox, 2011), whereas currently there is no computational model of the consolidation process.

The working memory maintenance and attentional switching network includes all structures in Figure 1, except the ACC and hippocampus. The idea is that the long-term representation of each possible salient rule is encoded in some neural network in sensory association cortex. These cortical units send excitatory signals to working memory units in lateral PFC, which send recurrent excitatory signals back to the same cortical units, thereby forming a reverberating loop. At the same time, the PFC is part of a second excitatory reverberating loop through the MDN (Alexander, DeLong, & Strick, 1986). These double reverberating loops maintain activation in the PFC working memory units during the rule testing procedure. However, the high spontaneous activity that is characteristic of the GABAergic neurons in the globus pallidus tonically inhibit the MDN, which prevents the closing of this cortical-thalamic loop, leading to the loss of information from working memory. To counteract this inhibition, the PFC excites medium spiny neurons in the head of the caudate nucleus (Bennett & Wilson, 2000), which in turn inhibit the pallidal neurons (since medium spiny neurons are GABAergic) that are inhibiting the thalamus. Reducing the pallidal inhibition of the thalamus allows reverberation in cortical-thalamic loops, and thereby facilitates working memory maintenance. The computational version of this model successfully accounts for many behavioral and single-neuron working memory-related phenomena (Ashby et al., 2005).

When feedback convinces the learner that the current categorization rule is incorrect, then a new rule must be selected and executive attention must be switched from the old rule to the new rule. In COVIS, these selection and switching operations are mediated by separate neural processes. In the computational version of the model, the ACC selects among alternative rules by enhancing the activity of the specific PFC working memory unit that represents a particular rule (Ashby et al., 2011). This is accomplished via the following algorithm.

Denote the set of all possible explicit rules by $\mathbf{R} = \{R_1, R_2, ..., R_m\}$. Suppose rule R_i is used on trial n. If the response on trial n was correct, then rule R_i is used again on trial $n+1$ with probability 1. If the response on trial n was incorrect, then the probability of selecting rule R_k from the set $\mathbf{R}$ for use on trial $n+1$ equals

$$P_{n+1}(R_k) = \frac{Y_n(R_k)}{\sum\limits_{i=1}^{m} Y_n(R_m)}, \tag{1}$$

where $Y_n(R_k)$ represents the current weight of rule R_k, which depends on its initial salience, its reinforcement history, and whether or not is was used on trial n. The decision criteria associated with each rule are learned via gradient descent. The model has 6 free parameters: 1) the variance of perceptual and criterial noise associated with rule application; 2) a parameter γ that increases with the tendency of the learner to perseverate following negative feedback (i.e., and continue with the same unsuccessful rule); 3) a parameter λ that measures creativity during rule selection (i.e., as λ increases, lower salience rules are more likely to be selected); 4) Δ_{C} equals the increase in weight of the current rule following positive feedback; 5) Δ_{E} equals the decrease in weight of the current rule following negative feedback; and 6) the gradient-descent learning rate (i.e., δ). As brain dopamine levels rise, the perseveration parameter γ is assumed to decrease and the selection parameter λ is assumed to increase. This model has successfully accounted for learning in RB tasks, under a variety of experimental conditions, including for example, with and without a simultaneous dual task (Ashby et al., 2011)), and under normal or positive affect (Hélie, Paul, & Ashby, 2012b), and also in a variety of different neuropsychological patient populations, including Parkinson's disease (Hélie, Paul, & Ashby, 2012a) and anorexia nervosa (Filoteo et al., 2014). For a complete description of the model, see Ashby et al. (2011).

To perform well in RB tasks, participants must remember which rules they have already tested and rejected, in order to avoid revisiting these failed rules. As in many other models, COVIS assumes that the consolidation from working memory to long-term declarative memory representations is mediated by projections from the PFC to the hippocampus (e.g., Eichenbaum & Cohen, 2001). If the task is simple enough, then working memory might be sufficient to avoid these errors. Thus, COVIS predicts normal learning by medial temporal lobe amnesiacs in simple RB tasks in which the correct rule can be discovered before the list of rejected hypotheses is lost from working memory. In more difficult RB tasks (e.g., with many alternative rules), the search for the correct rule will exceed working memory capacity, so COVIS predicts that in these cases medial temporal lobe amnesiacs will be impaired. Much evidence supports the former prediction (Janowsky, Shimamura, Kritchevsky, & Squire, 1989; Leng & Parkin, 1988), but to our knowledge the latter prediction has not been rigorously tested. Even so, several studies have reported normal performance by amnesiacs on the first 50 trials of a difficult task, but impaired performance later on (Hopkins, Myers, Shohamy, Grossman, & Gluck, 2004; Knowlton et al., 1996). Temporal cortex has also been shown to interact with PFC when rules are retrieved from long-term storage (for a review, see Bunge, 2004).

In conclusion, the COVIS declarative system includes multiple subprocesses, such as selecting a rule, focusing attention of the selected rule, storing the rule in long-term memory, switching between rules, and adjusting the salience of rules depending on the nature of the feedback. Neuroimaging and neuropsychological results have provided evidence for such multiple, distinct processes in RB category learning, (Kehagia, Cools, Barker, & Robbins, 2009; Monchi, Petrides, Petre, Worsley, & Dagher, 2001; Price, Filoteo, & Maddox, 2009; Tachibana et al., 2009). Furthermore, it is known that dopamine influences many of these subprocesses (Ashby & Casale, 2003; Cools, 2006; Cools, Lewis, Clark, Barker, & Robbins, 2007; Frank & O'Reilly, 2006; Monchi et al., 2004; Moustafa & Gluck, 2011; Price et al., 2009; Seamans & Yang, 2004).

3.1.2. Models of the WCST. A number of models have been developed to account for results of experiments with the WCST. Within this set, the more neurobiologically detailed models were developed specifically to account for the impaired WCST performance of a number of different special neuropsychological patient groups – including schizophrenics, Parkinson's disease patients, and patients with Huntington's disease. In general, these models are similar to the rule-learning submodel of COVIS, except typically with more biological detail in certain brain regions.

Monchi, Taylor, and Dagher (2000) proposed a COVIS-like model that includes an extra reward-processing circuit in which reward-related signals from the amygdala project to the nucleus accumbens (NAcc). The goal of this work was to explain how dopamine imbalences cause suboptimal WCST performance in Parkinson's patients and schizophrenics. Monchi et al. (2000) simulated impaired performance in schizophrenic patients by reducing the gains in the NAcc, which caused rule-selection deficits within an ACC/basal ganglia circuit, which in turn reduced PFC activation. In contrast, the suboptimal performance of Parkinson's patients was simulated by reducing the synaptic strengths between PFC and the caudate nucleus, and between the caudate and the internal segment of the globus pallidus. These decreases reduced the cortical activity and impaired the encoding of features in working memory.

Amos (2000) attempted to explain how perseverative and random errors in the WCST might be caused by dopamine imbalances in the PFC and basal ganglia of Parkinson's, schizophrenic, and Huntington's disease patients. His model included a reward/punishment unit (presumably in the ventral tegmental area) that projected to inhibitory units in PFC, which were reciprocally connected to the PFC rule units. By changing the simulated gains in PFC and basal ganglia, Amos (2000) inferred that perseverative errors were more likely to be PFC dependent, whereas random errors were more likely basal ganglia dependent.

Moustafa and Gluck (2011) developed a similar model with the goal of accounting for on- and off-medication performance of Parkinson's patients in a task that was similar to the WCST, in the sense that it also required attentional switches to a new stimulus dimension after a rule is learned. In this model, dopamine neurons in the substantia nigra pars compacta and ventral tegmental area (i.e., the critic) influenced activity in the PFC and striatum by altering two types of dopamine input: tonic dopamine, which affected the gain on activity, and phasic dopamine, which dynamically affected changes in connection weights. They assumed that Parkinson's disease reduces phasic and tonic dopamine levels in PFC and the basal ganglia, and that the primary effect of medication is to increase tonic dopamine levels, but that this increase actually reduces the phasic dopamine signal.

All models considered so far assume that a representation of the stimulus is compared to a representation of the current rule in PFC. In contrast to this, Leabra assumes that the relevant perceptual representations are maintained in posterior cortex, and that these representations are modulated by PFC (Rougier & O'Reilly, 2002; O'Reilly et al., 2002). This view of PFC function is supported by some recent studies suggesting that the PFC plays a mostly modulatory role in working memory maintenance (see e.g., Sreenivasan, Curtis, & D'Esposito, 2014 for a review). In Leabra, the mapping from stimulus to response is mediated directly via weight-based associations between posterior cortex and response output units, which receive top-down bias from PFC along the selected dimension. The ventral tegmental area acts as a critic by sending reward-prediction-error signals to the PFC, which have the effect of stabilizing or destabilizing current PFC activity patterns.

3.2. Models of categorization in II tasks

3.2.1. COVIS. The COVIS procedural-learning system incrementally learns arbitrary stimulus-response associations via dopamine-mediated reinforcement learning. Procedural learning is typically associated with motor learning (e.g., Willingham, 1998; Willingham, Nissen, & Bullemer, 1989), and accordingly, the COVIS procedural system assumes that II learning includes a strong motor component.

Figure 2 shows the architecture of the COVIS procedural-learning system (Ashby et al., 1998; Ashby & Waldron, 1999; Ashby & Crossley, 2011; Cantwell et al., 2015). The key structure is the striatum, a major input region within the basal ganglia that includes the caudate nucleus and the putamen. In primates, all of extrastriate visual cortex projects directly to the striatum, with a cortical-striatal convergence ratio of approximately 10,000 to 1 (e.g., Wilson, 1995). The model assumes that, through a procedural learning process, each striatal medium spiny neuron (MSN) associates an abstract motor program with a large group of visual cortical neurons (i.e., all that project to it). Much evidence supports the hypothesis that procedural learning is mediated within the basal ganglia, and especially at cortical-striatal synapses, where synaptic plasticity is thought to follow reinforcement learning rules (Ashby & Ennis, 2006; Houk, Adams, & Barto, 1995; Mishkin, Malamut, & Bachevalier, 1984; Willingham, 1998). The COVIS procedural-learning system is a formal instantiation of these ideas.

Note that the model includes two loops through the basal ganglia (Cantwell et al., 2015). One loop projects from visual cortex through the body and tail of the caudate nucleus and terminates in preSMA, and the second loop projects from preSMA through the putamen and terminates in SMA. Because this second loop terminates in premotor cortex, COVIS

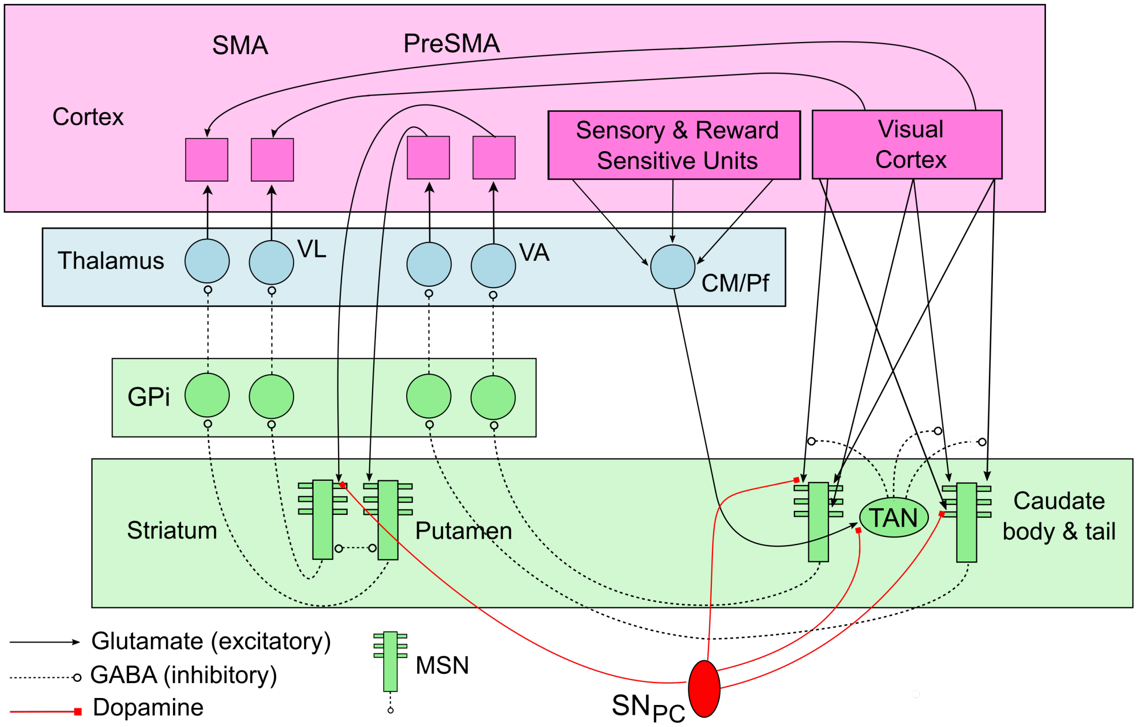

Figure 2. The neural architecture of the COVIS procedural category-learning system (SMA = supplementary motor area, PreSMA = presupplementary motor area, VL = ventral lateral nucleus of the thalamus, VA = ventral anterior nucleus of the thalamus, CMPf = centromedian and parafascicular nuclei of the thalamus, GPi = internal segment of the globus pallidus, TAN = tonically active neuron, SN_{PC} = substantia nigra pars compacta, MSN = medium spiny neuron of the striatum).

predicts that the associations that are learned are between stimuli and abstract motor goals (e.g., press the button on the left). Both loops rely on reinforcement learning at cortical-striatal synapses. The first loop learns which stimuli are associated with the same response and the second loop learns what motor response is associated with each of these stimulus clusters. With novel categories, both types of learning are required. However, note that if we train subjects to make accurate categorization responses and then switch the responses associated with the two categories, then the category structures remain unchanged – only the response mappings must be relearned. So COVIS predicts that reversing the locations of the response keys will interfere with II performance, but that recovery from such a reversal should be easier than learning novel categories – a prediction that has been supported in several studies (Cantwell et al., 2015; Kruschke, 1996; Maddox, Glass, O'Brien, Filoteo, & Ashby, 2010; Sanders, 1971; Wills, Noury, Moberly, & Newport, 2006).[1]

The units in the COVIS procedural-learning model are based on the Izhikevich (2003) spiking-neuron model. Let $V_i(t)$ and $V_j(t)$ denote the intracellular voltages of a pre- and

[1] In contrast, COVIS also predicts that such reversals should not impair initial RB performance, since the COVIS declarative system does not assign a prominent role to any premotor or motor regions of cortex (see Figure 1). Many of these same studies also supported this prediction.

postsynaptic neuron, respectively, at time t. Then the Izhikevich (2003) model assumes that the intracellular voltage of the postsynaptic neuron on trial n is described by the following differential equations:

$$\frac{dV_j(t)}{dt} = w_{ij}(n)f\left[V_i(t)\right] + \beta + \gamma\left[V_j(t) - V_r\right]\left[V_j(t) - V_t\right] - \theta U_j(t),$$
$$\frac{dU_j(t)}{dt} = \lambda\left[V_j(t) - V_r\right] - \omega U_j(t), \tag{2}$$

where β, γ, V_r, V_t, θ, λ, and ω are constants that are adjusted to produce dynamical behavior that matches the neural population being modeled. $U_j(t)$ is an abstract regulatory term that is meant to describe slow recovery in the postsynaptic neuron after an action potential is generated. Equation 2 produces the upstroke of an action potential via its own dynamics. To produce the downstroke, $V_j(t)$ is reset to V_{reset} when it reaches V_{peak}, and at the same time, $U_j(t)$ is reset to $U_j(t) + U_{\text{reset}}$, where V_{reset}, V_{peak}, and U_{reset} are free parameters.

The model has many free parameters and therefore can fit a wide variety of dynamical behavior. Izhikevich (2003) identified different sets of parameter values that allow the model to mimic the spiking behavior of approximately 20 different types of neurons. For example, one set of parameter values allows the model to mimic the firing properties of the striatal medium spiny neurons shown in Figure 2 (including, e.g., their up and down states), and another set of values allows the model to mimic the regular spiking neurons that are common in cortex. Furthermore, Ashby and Crossley (2011) modified the Izhikevich model to account for the unusual dynamics of the striatal cholinergic interneurons known as TANs (which produce a pronounced pause in their high tonic firing rate following excitatory input). In all these cases, the parameters are fixed by fitting the model to single-unit recording data from the neural population being modeled. Once set, the parameter values that define the models of each individual neuron type then remain fixed throughout all applications. Therefore, when testing the model against behavioral or neuroimaging data, the models of each neuron type have zero free parameters.

The function $f[V_i(t)]$ in Eq. 2 models the input from the presynaptic neuron i. In particular, it uses a simple model called the alpha function to mimic the temporal delays of spike propagation and the temporal smearing that occurs at the synapse (Rall, 1967). Specifically, the alpha function assumes that every time the presynaptic neuron spikes, the following input is delivered to the postsynaptic neuron (with spiking time $t = 0$):

$$\alpha(t) = \frac{t}{\delta}\exp\left(\frac{\delta - t}{\delta}\right), \tag{3}$$

where δ is a constant. This function has a maximum value of 1.0 and it decays to .01 at $t = 7.64\delta$. Thus, δ can be chosen to model any desired temporal delay. Suppose the presynaptic neuron i produces N spikes that occur at times $t_1, t_2, ..., t_N$. Then the function f in Eq. 2 equals

$$f\left[V_i(t)\right] = \sum_{k=1}^{N}\left[\alpha(t - t_k)\right]^+, \tag{4}$$

where

$$\left[\alpha(t - t_k)\right]^+ = \begin{cases} \alpha(t - t_k) & \text{if } t > t_k; \\ 0 & \text{if } t \leq t_k. \end{cases} \tag{5}$$

COVIS assumes that the procedural learning in the striatum is facilitated by a dopamine-mediated reward signal from the substantia nigra pars compacta (SNpc). There is a large literature linking dopamine and reward, and many researchers have argued that a primary function of dopamine is to serve as the reward signal in reward-mediated learning (e.g., Houk et al., 1995; Wickens, 1993). The well-accepted theory is that positive feedback that follows successful behaviors increases phasic dopamine levels in the striatum, which has the effect of strengthening recently active synapses, whereas negative feedback causes dopamine levels to fall below baseline, which has the effect of weakening recently active synapses (e.g., Arbuthnott, Ingham, & Wickens, 2000; Calabresi, Pisani, Mercuri, & Bernardi, 1996; Reynolds & Wickens, 2002). In this way, the dopamine response to feedback serves as a teaching signal that allows successful behaviors to increase in probability and unsuccessful behaviors to decrease in probability. These learning-related effects are modeled by the $w_{ij}(n)$ multiplier on $f\left[V_i(t)\right]$ in Eq. 2. The value of this term is adjusted trial-by-trial according to standard models of dopamine-mediated synaptic plasticity in the striatum. For a complete description of this type of mathematical modeling, called computational cognitive neuroscience, see Ashby (2018).

According to this account, synaptic plasticity requires that the visual trace of the stimulus and the post-synaptic effects of dopamine overlap in time. More specifically, synaptic plasticity in the striatum is strongest when the intracellular signaling cascades driven by NMDA receptor activation and dopamine D1 receptor activation coincide (Lisman, Schulman, & Cline, 2002; Rudy, 2014). The further apart in time these two cascades peak, the less effect dopamine will have on synaptic plasticity. For example, Yagishita et al. (2014) reported that synaptic plasticity was best (i.e., greatest increase in spine volume on striatal MSNs) when dopamine neurons were stimulated 600 ms after MSNs. When the dopamine neurons were stimulated before the MSNs or 5 s after the MSNs, then no evidence of any plasticity was observed. Similar results have been reported in II category learning. First, Worthy, Markman, and Maddox (2013) reported that II learning is best with feedback delays of 500ms and slightly worse with delays of 0 or 1000ms. Second, several studies have reported that feedback delays of 2.5 s or longer impair II learning, whereas delays as long as 10 secs have no effect on RB category learning (Dunn, Newell, & Kalish, 2012; Maddox, Ashby, & Bohil, 2003; Maddox & Ing, 2005). Valentin, Maddox, and Ashby (2014) showed that the COVIS procedural-learning system can accurately account for the effects of all these feedback delays.

Ashby and Crossley (2011) proposed that the striatal cholinergic interneurons known as TANs (for tonically active neurons) serve as a context-sensitive gate between cortex and the striatum (see also Crossley, Ashby, & Maddox, 2013, 2014; Crossley, Horvitz, Balsam, & Ashby, 2016). The idea, which is supported by a wide variety of neuroscience evidence, is that the TANs tonically inhibit cortical input to striatal output neurons (e.g., Apicella, Legallet, & Trouche, 1997; Matsumoto, Minamimoto, Graybiel, & Kimura, 2001; Pakhotin & Bracci, 2007; Smith, Raju, Pare, & Sidibe, 2004). The TANs are driven by neurons in the centremedian–parafascicular (CM-Pf) nuclei of the thalamus, which in turn are broadly tuned to features of the environment. In rewarding environments, the TANs learn to pause to stimuli that predict reward, which releases the cortical input to the striatum from inhibition. This allows striatal output neurons to respond to excitatory cortical input, thereby facilitating cortical-striatal plasticity. In this way, TAN pauses facilitate the learning and expression of striatal-dependent behaviors. When rewards are no longer available, the TANs cease to pause, which prevents striatal-dependent responding and protects striatal learning from decay.

Extending the COVIS procedural-learning system to include TANs allows the model to account for many new phenomena – some of which have posed difficult challenges for previous learning theories. One of these is that the reacquisition of an instrumental behavior after it has been extinguished is considerably faster than during original acquisition (Ashby & Crossley, 2011). The model accounts for this ubiquitous phenomenon because the withholding of rewards during the extinction period causes the TANs to stop pausing to sensory cues in the conditioning environment (since they are no longer associated with reward). This closes the gate between cortex and the striatum, which prevents further weakening of the cortical-striatal synapses. When the rewards are reintroduced, the TANs relearn to pause, and the behavior immediately reappears because of the preserved synaptic strengths.

3.2.2. Exemplar models. Exemplar theory assumes that categorization is a process of learning about the exemplars that belong to the category (Estes, 1986; Medin & Schaffer, 1978; Nosofsky, 1986). When an unfamiliar stimulus is encountered, its similarity is computed to the memory representation of every previously seen exemplar from each potentially relevant category. Exemplar theory has been the most prominent cognitive theory of categorization for more than 30 years, but despite this popularity, until recently, it has never had a detailed neurobiological interpretation. Ashby and Rosedahl (2017) showed that the exemplar model is mathematically equivalent to a simplified version of the COVIS procedural-learning model (e.g., with only one loop through the basal ganglia). In this neural version of exemplar theory, category learning is mediated by synaptic plasticity at cortical-striatal synapses. The neural version makes identical quantitative predictions to the cognitive version of exemplar theory, yet it can account for many empirical phenomena that are either incompatible with or outside the scope of the cognitive version. The neural version also reinterprets the psychological assumptions associated with exemplar theory. The cognitive version assumes that for every categorization decision, people retrieve memory representations of every previously seen category exemplar and that they compute the similarity of the presented stimulus to all these stored memories. Categorization decisions are based on the sum of all these similarities. In contrast, the neural version assumes that no memory representations are ever retrieved. Instead, the summed similarities are encoded in the strength of synapses between sensory cortex and the striatum.

3.3. Models of categorization in prototype-distortion tasks

The prototype-distortion task was originally designed to study category learning (Posner & Keele, 1968), but the idea that the brain abstracts a wide variety of perceptual information soon became a key component of many object recognition theories (e.g., see Logothetis & Sheinberg, 1996). Therefore, models that assume categorization depends on the representation of prototypes are often tested with more complex stimuli, such as abstract objects (Riesenhuber & Poggio, 1999), artificial creatures (Riesenhuber & Poggio, 2002; Love & Gureckis, 2007), and real-world scenes (Serre, Oliva, & Poggio, 2007). Prototype-based models assume that categorization decisions are based on the distances between the representations of the stimulus and the prototypes of each category, and that categorization probability is inversely related to these distances. Thus, the stimulus is most likely to be assigned to the category with the nearest prototype. If distance is measured using the Euclidean metric, then this decision strategy always produces piece-wise linear bounds and is equivalent to template matching (Ashby & Gott, 1988). The models differ in how the prototypes are formed and represented in the network.

One way to approach this question is to start from neuroscientific observations. For example, based on single-unit recording results in primates, Riesenhuber and Poggio (1999) proposed a model called HMAX that describes visual processing in the ventral visual stream from V1 up through inferotemporal cortex (see also, Serre et al., 2007). At each stage, the level of abstraction is increased. This is done by converging the projections of many units that respond to similar stimuli onto the same unit at the next higher level, and assuming that the response of each unit equals the maximum activation of all input units. In this way, each level of abstraction can be viewed as a kind of prototype. At a final stage, the object-tuned neurons in inferotemporal cortex project to classification units in PFC, where the output of each unit equals a linear combination of its inputs, with the coefficients adjusted via a supervised learning process to maximize categorization accuracy (Serre et al., 2007). The model is strictly feedforward, and has included as many as 10 million units. Parameters of the units are set to match physiological data – for example, to create units that match the physiological responses of simple and complex cells. Thus, in tests against behavioral data, the model has no free parameters. The model has successfully accounted for single-unit recording results in primates using categories constructed of abstract images (Riesenhuber & Poggio, 1999) and creature-like images (Riesenhuber & Poggio, 2002), and also for the performance of human observers classifying natural scenes (Serre et al., 2007).

Although Leabra was not proposed as a model of learning in prototype-distortion tasks, its visual layers (V1 to inferotemporal cortex) can be viewed as a simplified version of HMAX. Specifically, like HMAX, Leabra also includes feedforward convergent projections in which the response of each unit equals the maximum activation of all its input units (O'Reilly et al., 2013). However, unlike HMAX, which is purely feedforward, Leabra also includes recurrent projections from higher cortical regions, which help shape the response of lower layers. Wyatte, Herd, Mingus, and O'Reilly (2012) argued that this property, along with competitive inhibition, is especially important for forming robust representations for ambiguous images, such as occluded objects. Despite these differences Leabra and HMAX offer similar interpretations of prototype-distortion learning.

Another approach is to develop the model from behavioral observations, and then map components of the model to brain regions. For example, SUSTAIN assumes that each category is represented as collection of stimulus clusters (Love & Gureckis, 2007). Each cluster begins initially as a single stimulus that was unexpected, either because it was dissimilar to previously seen stimuli or because it was associated with a response that feedback indicates was incorrect. New stimuli are added to an existing cluster if similarity is high, or else they form a new cluster if they are unexpected. SUSTAIN is equivalent to a prototype model if each category is defined by a single cluster, and to a multiple prototype model if categories are defined by more than one cluster.

3.4. Neuroscience-Based Models of Automatic Categorization

The COVIS procedural-learning model also accounts for how behaviors that are learned procedurally can eventually come to be executed automatically (Ashby et al., 2007).[2] A key role in this transition is played by the direct projections (shown in Figure 2) from visual cortex to SMA. Ashby et al. (2007) proposed that, by themselves, these projections are incapable of category learning because synaptic plasticity in cortex follows Hebbian, rather than reinforcement learning rules (Feldman, 2009). Although premotor cortex is a target of midbrain dopamine neurons, unlike the basal ganglia, concentrations of dopamine active transporter (DAT) are negligible in cortex (e.g., Varrone & Halldin, 2014). For this reason, dopamine remains in cortical synapses much longer than in striatal synapses. As a result, cortical dopamine levels are likely to remain above baseline during an entire training session. This means that all active synapses in cortex will get strengthened, even those leading to incorrect responses and negative feedback. Therefore, in the Figure 2 model, the basal ganglia play the critical role of training cortex. The idea is that, via dopamine-mediated reinforcement learning, the basal ganglia learn to activate the correct post-synaptic targets in SMA, which allows the appropriate cortical-cortical synapses to be strengthened via Hebbian learning. Once the cortical-cortical synapses have been built, the basal ganglia are no longer required to produce the automatic behavior.

This feature of the COVIS procedural system accounts for behavioral changes that occur as automaticity develops (i.e., improvements in both accuracy and response time), but it also accounts for a variety of neuroscience results that are problematic for other theories of automaticity (Ashby et al., 2007). For example, it correctly predicts that inactivation of the globus pallidus (which essentially prevents the basal ganglia from influencing the cortical motor and premotor areas) does not disrupt the ability of monkeys to fluidly produce an over-learned motor sequence (Desmurget & Turner, 2010), and that Parkinson's disease patients, who have significant striatal dysfunction and are impaired during early learning in some RB and II tasks, are relatively normal in executing automatic behaviors (Asmus, Huber, Gasser, & Schöls, 2008).

The data from many single-unit recording studies that examined neural responses during categorization were collected after the animals were trained on the task for weeks or months, and thus, after it is likely that automaticity had already developed. As a result, the models proposed to account for these data typically focus on cortical activations and do not address the neural changes that might have occurred as automaticity develops. For example, the HMAX model does not specify the neural mechanisms that mediate feedback-based learning in any regions of the model (Serre et al., 2007). As another example, Engel, Chaisangmongkon, Freedman, and Wang (2015) proposed a purely cortical model of how motion categories are learned that included areas MT and LIP. The model assumed that plasticity in this circuit is mediated by a trial-by-trial reward-prediction-error (RPE) signal that is encoded in the phasic activity of dopamine neurons. The low concentrations of DAT in cortex however, suggest that changes in cortical dopamine concentrations are likely to be too sluggish to track trial-by-trial RPEs (Varrone & Halldin, 2014). So one possibility

[2] This might be the only existing neuroscientific model of automatic categorization. On the other hand, there are several, closely related neuroscience-based models of automatic sequence production (e.g., Chersi, Mirolli, Pezzulo, & Baldassarre, 2013; Helie, Roeder, Vucovich, Rünger, & Ashby, 2015).

is that the basal ganglia provide this cortical teaching signal, rather than the dopamine neurons per se (e.g., as described by Ashby et al., 2007).

4. Discussion

Neuroscientific models of categorization are now entering their third decade of existence. They are still in their infancy, however, compared to cognitive models. Each approach has its own advantages. Cognitive models are typically more analytically tractable, and therefore easier to fit to data. In contrast, neuroscientific models are usually analytically intractable, and therefore require extensive computer simulations to test. In addition, cognitive models require less knowledge to formulate. They require specifying the cognitive and perceptual processes that mediate categorization, but they require no knowledge of the underlying neural mechanisms or processes. These advantages guarantee that cognitive models of categorization will continue to make important contributions for the foreseeable future.

Despite their computational complexity, and the greater knowledge they require to build, neuroscientific models have their own advantages. First, of course, they have the potential to account for a wide variety of data. In addition to traditional response accuracy and response time data, neuroscientific models also can be tested against a wide variety of neuroscience data, including single-unit recordings, fMRI BOLD responses, and EEG recordings. In addition, they can make predictions about how transcranial magnetic stimulation, neuropsychological disease, or pharmacological intervention affect behavior.

Second, neuroscientific models are less mathematically flexible than their cognitive counterparts (Ashby, 2018). Flexible models can compensate for incorrect psychological or neurobiological assumptions by sheer mathematical force. The ability to bend a poor model to fit some data makes it more difficult to reject, thereby slowing scientific progress. In contrast, it is more difficult for rigid models that make incorrect assumptions to avoid poor fits to data. As a result, their weaknesses are more quickly exposed, which hastens the model development process. With descriptive models, like linear regression, the sole purpose of adding a new parameter is almost always to fit more data. Therefore, it is almost always true that every new parameter greatly increases mathematical flexibility. However, in neuroscientific models, the goal is generally to model the hardware that mediates categorization decisions. In this case, new parameters are added to model structure – not new data.

Furthermore, mathematical inflexibility is built into neuroscientific models via the architectural and process constraints supplied by the relevant neuroscience literature. For example, consider a model that includes cortical and striatal units. The equations describing each unit will be characterized by a number of free parameters and there will be other parameters that describe the strength of the cortical-striatal synapses. But because the projection from cortex to striatum is excitatory and one way, changing the values of any of these parameters can only have a very limited effect on the behavior of the model – namely, any condition that causes cortical units to increase their firing rate will also cause striatal units to increase their firing rate. This is the only data profile that the model can produce, regardless of how many free parameters it contains, and regardless of the numerical values of those parameters. In other words, this is a parameter-free *a priori* prediction of such models – that is, that for all parameter values, increasing cortical activation can never reduce striatal activation. In fact, neuroscientific models typically make many such *a priori* predictions, whereas more flexible models rarely make any. As one other example (among many), Section 3.2.1 described how COVIS predicts that feedback delays of more than a few seconds can never impair RB category learning more than II learning.

Another advantage of neuroscientific models is that many experiments were run to test their *a priori* predictions. And because many such experiments were successful, they enriched the field by bringing to light a large number of new behavioral phenomena that characterize human category learning. For example, primarily because of *a priori* predictions of neuroscientific models, we now know that feedback delays interfere with II learning more than with RB learning, that a dual-task that recruits working memory interferes with RB learning more than II learning, that switching the locations of the response buttons after initial learning interferes with II performance more than RB performance, that limiting time and attention for feedback processing interferes with RB learning more than II learning, that unsupervised II learning is worse than unsupervised RB learning, and that analogical transfer is much worse in II tasks than in RB tasks. All of the experiments that discovered these results were run to test *a priori* predictions of COVIS, and these predictions were all empirically supported in replicated experiments (for a review and a description of many other examples, see Ashby & Valentin, 2017). Furthermore, it is likely that many of these empirical phenomena might not yet be known without the neuroscientific models that inspired these experiments.

Third, neuroscientific models can easily be extended by adding more structure and/or biological detail. As an example, consider the COVIS procedural-learning model that is described in Figure 2. The original version included only one loop through the striatum, rather than the two loops shown in Figure 2, and it lacked cholinergic interneurons in the striatum (i.e., the TANs) and cortical-cortical projections between visual and premotor cortices. These features were all added in later applications. This was a seamless process because each step in model development was true to the underlying neuroanatomy. For this reason, adding new structure did not require changing the older, simpler version of the model in any way. And adding these new structures allowed the model to account for an enormous number of new empirical phenomena.

An obvious extension of this same principle is that if two different neuroscientific models are both faithful to the known neuroanatomy, and the two models focus on different, but overlapping neural networks, then it should be possible to connect them in a straightforward, plug-and-play fashion. Cantwell, Riesenhuber, Roeder, and Ashby (2017) illustrated this principle. The COVIS procedural-learning model had always included a grossly oversimplified model of visual cortex and the HMAX model of Riesenhuber and Poggio (1999, 2002) had always oversimplified early category learning. To overcome both of these limitations, Cantwell et al. (2017) replaced the COVIS model of visual cortex with HMAX. HMAX uses bitmap images of the stimulus as input and outputs a 4,075 × 1 vector that is presumed to model activation in visual area V2 or V4. Cantwell et al. (2017) simply connected each of these outputs to a unique synapse on each striatal MSN of the COVIS procedural-learning model. Except for some simple scaling of these outputs, no other changes were made to either model. The new HMAX/COVIS model provided impressively good fits to human category-learning data from two qualitatively different experiments that used different types of category structures and different types of visual stimuli and it did this using bitmap images of the stimuli as inputs, rather than the abstract stimulus representations used in previous applications of COVIS.

All of these advantages guarantee that the development of neuroscientific models of categorization will only accelerate in the future. Furthermore, the field of neuroscience is growing at an enormous pace, and many of the new discoveries about human brain function will greatly facilitate this development. The result will almost surely be models that are unprecedented in their breadth and predictive ability.

5. Conclusions

Before the 1990s, almost nothing was known about the neural networks and processes that mediate human categorization. As a result, theories of categorization were dominated by purely cognitive descriptions. The cognitive neuroscience revolution ushered in a new era in which many results dramatically increased our understanding of the neural bases of human categorization. As a result, models grounded in neuroscience are becoming increasingly popular. Collectively, these models have already made profound contributions to our understanding of human categorization – by widening the empirical domain of categorization research, and by motivating experiments that might not otherwise have been run. Furthermore, this trend should increase in the future, as methods for studying the functioning human brain improve and the neuroscience database grows.

Acknowledgments

Preparation of this chapter was supported by Public Health Service Grant MH2R01-063760.

References

Aizenstein, H. J., MacDonald, A. W., Stenger, V. A., Nebes, R. D., Larson, J. K., Ursu, S., & Carter, C. S. (2000). Complementary category learning systems identified using event-related functional mri. *Journal of Cognitive Neuroscience*, *12*(6), 977–987.

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, *9*(1), 357–381.

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

Apicella, P., Legallet, E., & Trouche, E. (1997). Responses of tonically discharging neurons in the monkey striatum to primary rewards delivered during different behavioral states. *Experimental Brain Research*, *116*(3), 456–466.

Arbuthnott, G., Ingham, C., & Wickens, J. (2000). Dopamine and synaptic plasticity in the neostriatum. *Journal of Anatomy*, *196*(04), 587–596.

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., 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., & Casale, M. B. (2003). The cognitive neuroscience of implicit category learning. In L. Jiménez (Ed.), *Attention and implicit learning* (Vol. 48, pp. 109–142). Amsterdam: John Benjamins Publishing Company.

Ashby, F. G., & Crossley, M. J. (2011). A computational model of how cholinergic interneurons protect striatal-dependent learning. *Journal of Cognitive Neuroscience*, *23*(6), 1549-1566.

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., & 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., & O'Brien, J. B. (2005). Category learning and multiple memory systems. *TRENDS in Cognitive Science*, *2*, 83–89.

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., & Rosedahl, L. (2017). A neural interpretation of exemplar theory. *Psychological Review*, *124*(4), 472–482.

Ashby, F. G., & Valentin, V. V. (2017). Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science, Second edition* (pp. 157–188). Amsterdam: 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). New York: Wiley.

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

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

Bennett, B. D., & Wilson, C. J. (2000). Synaptology and physiology of neostriatal neurones. In R. Miller & J. R. Wickens (Eds.), *Brain dynamics and the striatal complex* (pp. 111–140). Amsterdam: Harwood Academic Publishers.

Braver, T. S., Cohen, J. D., Nystrom, L. E., Jonides, J., Smith, E. E., & Noll, D. C. (1997). A parametric study of prefrontal cortex involvement in human working memory. *Neuroimage*, *5*(1), 49–62.

Bunge, S. A. (2004). How we use rules to select actions: A review of evidence from cognitive neuroscience. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(4), 564–579.

Calabresi, P., Pisani, A., Mercuri, N. B., & Bernardi, G. (1996). The corticostriatal projection: From synaptic plasticity to dysfunctions of the basal ganglia. *Trends in Neurosciences*, *19*(1), 19–24.

Cantwell, G., Crossley, M. J., & Ashby, F. G. (2015). Multiple stages of learning in perceptual categorization: Evidence and neurocomputational theory. *Psychonomic Bulletin & Review*, *22*(6), 1598–1613.

Cantwell, G., Riesenhuber, M., Roeder, J. L., & Ashby, F. G. (2017). Perceptual category learning and visual processing: An exercise in computational cognitive neuroscience. *Neural Networks*, *89*, 31–38.

Casale, M. B., & Ashby, F. G. (2008). A role for the perceptual representation memory system in category learning. *Perception & Psychophysics*, *70*(6), 983–999.

Chersi, F., Mirolli, M., Pezzulo, G., & Baldassarre, G. (2013). A spiking neuron model of the cortico-basal ganglia circuits for goal-directed and habitual action learning. *Neural Networks*, *41*, 212–224.

Cools, R. (2006). Dopaminergic modulation of cognitive function-implications for l-dopa treatment in parkinson's disease. *Neuroscience & Biobehavioral Reviews*, *30*(1), 1–23.

Cools, R., Lewis, S. J., Clark, L., Barker, R. A., & Robbins, T. W. (2007). L-dopa disrupts activity in the nucleus accumbens during reversal learning in parkinson's disease. *Neuropsychopharmacology*, *32*(1), 180–189.

Crossley, M. J., Ashby, F. G., & Maddox, W. T. (2013). Erasing the engram: The unlearning of procedural skills. *Journal of Experimental Psychology: General*, *142*(3), 710–741.

Crossley, M. J., Ashby, F. G., & Maddox, W. T. (2014). Context-dependent savings in procedural category learning. *Brain & Cognition*, *92*, 1-10.

Crossley, M. J., Horvitz, J. C., Balsam, P. D., & Ashby, F. G. (2016). Expanding the role of striatal cholinergic interneurons and the midbrain dopamine system in appetitive instrumental conditioning. *Journal of Neurophysiology*, *115*, 240-254.

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences*, *7*(9), 415–423.

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

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

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

Eichenbaum, H., & Cohen, N. J. (2001). *From conditioning to conscious recollection: Memory systems of the brain.* Oxford University Press.

Engel, T. A., Chaisangmongkon, W., Freedman, D. J., & Wang, X.-J. (2015). Choice-correlated activity fluctuations underlie learning of neuronal category representation. *Nature Communications*, *6*, 6454.

Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology*, *18*(4), 500-549.

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

Filoteo, J. V., Maddox, W. T., Salmon, D. P., & Song, D. D. (2005). Information-integration category learning in patients with striatal dysfunction. *Neuropsychology*, *19*(2), 212-222.

Filoteo, J. V., Paul, E. J., Ashby, F. G., Frank, G. K., Helie, S., Rockwell, R., … Kaye, W. H. (2014). Simulating category learning and set shifting deficits in patients weight-restored from anorexia nervosa. *Neuropsychology*, *28*(5), 741.

Frank, M. J., & O'Reilly, R. C. (2006). A mechanistic account of striatal dopamine function in human cognition: Psychopharmacological studies with cabergoline and haloperidol. *Behavioral neuroscience*, *120*(3), 497–517.

Heaton, R. K. (1981). *Wisconsin card sorting test manual.* Odessa, FL: Psychological Assessment Resources.

Hélie, S., Paul, E. J., & Ashby, F. G. (2012a). A neurocomputational account of cognitive deficits in parkinson's disease. *Neuropsychologia*, *50*(9), 2290–2302.

Hélie, S., Paul, E. J., & Ashby, F. G. (2012b). Simulating the effects of dopamine imbalance on cognition: From positive affect to parkinson's disease. *Neural Networks*, *32*, 74-85.

Helie, S., Roeder, J. L., Vucovich, L., Rünger, D., & Ashby, F. G. (2015). A neurocomputational model of automatic sequence production. *Journal of Cognitive Neuroscience*, *27*(7), 1456–1469.

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

Hopkins, R. O., Myers, C. E., Shohamy, D., Grossman, S., & Gluck, M. (2004). Impaired probabilistic category learning in hypoxic subjects with hippocampal damage. *Neuropsychologia*, *42*(4), 524–535.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (p. 249-270). Cambridge, MA: MIT Press.

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

Janowsky, J. S., Shimamura, A. P., Kritchevsky, M., & Squire, L. R. (1989). Cognitive impairment following frontal lobe damage and its relevance to human amnesia. *Behavioral Neuroscience*, *103*(3), 548–560.

Kehagia, A. A., Cools, R., Barker, R. A., & Robbins, T. W. (2009). Switching between abstract rules reflects disease severity but not dopaminergic status in parkinson's disease. *Neuropsychologia*, *47*(4), 1117–1127.

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science*, *273*(5280), 1399-1402.

Kruschke, J. K. (1996). Dimensional relevance shifts in category learning. *Connection Science*, *8*(2), 225-247.

Leng, N. R., & Parkin, A. J. (1988). Double dissociation of frontal dysfunction in organic amnesia. *British Journal of Clinical Psychology*, *27*(4), 359–362.

Lisman, J., Schulman, H., & Cline, H. (2002). The molecular basis of camkii function in synaptic and behavioural memory. *Nature Reviews Neuroscience*, *3*(3), 175-190.

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience*, *19*(1), 577–621.

Love, B. C., & Gureckis, T. M. (2007). Models in search of a brain. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(2), 90–108.

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., Filoteo, J. V., Hejl, K. D., et al. (2004). Category number impacts rule-based but not information-integration category learning: Further evidence for dissociable category-learning systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 227–235.

Maddox, W. T., Glass, B. D., O'Brien, J. B., Filoteo, J. V., & Ashby, F. G. (2010). Category label and response location shifts in category learning. *Psychological Research*, *74*(2), 219-236.

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.

Matsumoto, N., Minamimoto, T., Graybiel, A. M., & Kimura, M. (2001). Neurons in the thalamic cm-pf complex supply striatal neurons with information about behaviorally significant sensory events. *Journal of Neurophysiology*, *85*(2), 960–976.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207-238.

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

Mishkin, M., Malamut, B., & Bachevalier, J. (1984). Memories and habits: Two neural systems. In G. Lynch, J. L. McGaugh, & N. M. Weinberger (Eds.), *Neurobiology of human learning and memory* (p. 65-77). New York: Guilford Press.

Monchi, O., Petrides, M., Doyon, J., Postuma, R. B., Worsley, K., & Dagher, A. (2004). Neural bases of set-shifting deficits in parkinson's disease. *The Journal of Neuroscience*, *24*(3), 702–710.

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.

Monchi, O., Taylor, J. G., & Dagher, A. (2000). A neural model of working memory processes in normal subjects, parkinson's disease and schizophrenia for fMRI design and predictions. *Neural Networks*, *13*(8-9), 953–973.

Moustafa, A. A., & Gluck, M. A. (2011). A neurocomputational model of dopamine and prefrontal–striatal interactions during multicue category learning by parkinson patients. *Journal of Cognitive Neuroscience*, *23*(1), 151–167.

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.

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

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

O'Reilly, R. C., Hazy, T. E., & Herd, S. A. (2016). The Leabra cognitive architecture: How to play 20 principles with nature. *The Oxford handbook of cognitive science*, *91*, 91–116.

O'Reilly, R. C., Noelle, D. C., Braver, T. S., & Cohen, J. D. (2002). Prefrontal cortex and dynamic categorization tasks: Representational organization and neuromodulatory control. *Cerebral Cortex*, *12*(3), 246–257.

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

Pakhotin, P., & Bracci, E. (2007). Cholinergic interneurons control the excitatory input to the striatum. *The Journal of Neuroscience*, *27*(2), 391–400.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3p1), 353–363.

Price, A., Filoteo, J. V., & Maddox, W. T. (2009). Rule-based category learning in patients with parkinson's disease. *Neuropsychologia*, *47*(5), 1213–1226.

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., & Squire, L. R. (1999). Intact learning of artificial grammars and intact category learning by patients with parkinson's disease. *Behavioral Neuroscience*, *113*(2), 235–242.

Reber, P. J., Stark, C. E., & Squire, L. R. (1998). Contrasting cortical activity associated with category memory and recognition memory. *Learning & Memory*, *5*(6), 420–428.

Reynolds, J. N., & Wickens, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks*, *15*(4), 507–521.

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

Riesenhuber, M., & Poggio, T. (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology*, *12*(2), 162–168.

Rougier, N. P., & O'Reilly, R. C. (2002). Learning representations in a gated prefrontal cortex model of dynamic task switching. *Cognitive Science*, *26*(4), 503–520.

Rudy, J. W. (2014). *The neurobiology of learning and memory.* Sunderland, MA: Sinauer.

Sanders, B. (1971). Factors affecting reversal and nonreversal shifts in rats and children. *Journal of Comparative and Physiological Psychology*, *74*, 192-202.

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

Seamans, J. K., & Yang, C. R. (2004). The principal features and mechanisms of dopamine modulation in the prefrontal cortex. *Progress in Neurobiology*, *74*(1), 1–58.

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

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, *104*(15), 6424–6429.

Smith, Y., Raju, D. V., Pare, J.-F., & Sidibe, M. (2004). The thalamostriatal system: A highly specific network of the basal ganglia circuitry. *Trends in Neurosciences*, *27*(9), 520–527.

Sreenivasan, K. K., Curtis, C. E., & D'Esposito, M. (2014). Revisiting the role of persistent neural activity during working memory. *Trends in Cognitive Sciences*, *18*(2), 82–89.

Tachibana, K., Suzuki, K., Mori, E., Miura, N., Kawashima, R., Horie, K., ... Mushiake, H. (2009). Neural activity in the human brain signals logical rule identification. *Journal of Neurophysiology*, *102*(3), 1526–1537.

Valentin, V. V., Maddox, W. T., & Ashby, F. G. (2014). A computational model of the temporal dynamics of plasticity in procedural learning: Sensitivity to feedback timing. *Frontiers in Psychology*, *5*(643).

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.

Wickens, J. (1993). *A theory of the striatum.* Oxford: Pergamon Press.

Willingham, D. B. (1998). A neuropsychological theory of motor skill learning. *Psychological Review*, *105*, 558–584.

Willingham, D. B., Nissen, M. J., & Bullemer, P. (1989). On the development of procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(6), 1047-1060.

Wills, A., Noury, M., Moberly, N. J., & Newport, M. (2006). Formation of category representations. *Memory & Cognition*, *34*(1), 17-27.

Wilson, C. J. (1995). The contribution of cortical neurons to the firing pattern of striatal spiny neurons. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (p. 29-50). Cambridge, MA: MIT Press.

Worthy, D. A., Markman, A. B., & Maddox, W. T. (2013). Feedback and stimulus-offset timing effects in perceptual category learning. *Brain and Cognition*, *81*(2), 283-293.

Wyatte, D., Herd, S., Mingus, B., & O'Reilly, R. (2012). The role of competitive inhibition and top-down feedback in binding during object recognition. *Frontiers in Psychology*, *3*, 182.

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