Published in final edited form as:

Psychol Rev. 2017 July ; 124(4): 472–482. doi:10.1037/rev0000064.

A Neural Interpretation of Exemplar Theory

F. Gregory Ashby and **Luke Rosedahl**

University of California, Santa Barbara

Abstract

Exemplar theory assumes that people categorize a novel object by comparing its similarity to the memory representations of all previous exemplars from each relevant category. Exemplar theory has been the most prominent cognitive theory of categorization for more than 30 years. Despite its considerable success in providing good quantitative fits to a wide variety of accuracy data, it has never had a detailed neurobiological interpretation. This article proposes a neural interpretation of exemplar theory in which category learning is mediated by synaptic plasticity at cortical-striatal synapses. In this model, categorization training does not create new memory representations, rather it alters connectivity between striatal neurons and neurons in sensory association cortex. The new model makes identical quantitative predictions as exemplar theory, yet it can account for many empirical phenomena that are either incompatible with or outside the scope of the cognitive version of exemplar theory.

Keywords

Categorization; Exemplar Theory; COVIS; Striatum

Introduction

Exemplar theory assumes that categorization is a process of learning about the exemplars that belong to the category. When an unfamiliar stimulus is encountered, its similarity is computed to the memory representation of every previously seen exemplar from each potentially relevant category. The probability that the stimulus is assigned to each category increases with the sum of these similarities (Brooks, 1978; Estes, 1986, 1994; Kruschke, 1992; Lamberts, 2000; Medin & Schaffer, 1978; Nosofsky, 1986).

Exemplar theory has been the most prominent cognitive theory of categorization for more than 30 years. Despite its considerable success in providing good quantitative fits to accuracy data from a wide variety of experiments, it has never had a detailed neurobiological interpretation. This article corrects that shortcoming of the literature. We propose a neural version of exemplar theory in which category learning is mediated by synaptic plasticity at cortical-striatal synapses. The neural version makes identical quantitative predictions to the exemplar model, yet it can account for many empirical phenomena that are either incompatible with or outside the scope of the cognitive version of exemplar theory.

One challenge to interpreting exemplar theory at the neural level is to separate the quantitative predictions of the theory from the cognitive interpretations that are used to justify those predictions. At the cognitive level, exemplar theory assumes that people access the memory representations of all previously seen exemplars. The standard interpretation is that these are detailed replicas of each exemplar (filtered by attentional processes) that do not typically include contextual information (e.g., details about the experimental room). The closest match in the memory literature to such representations is probably provided by semantic memory.

This cognitive version of exemplar theory has been a challenge to interpret at the neural level. First, the memory processes postulated by exemplar theory appear qualitatively different from all of the memory systems that have been identified by memory researchers. For example, within the memory systems literature, semantic memory is assumed to be declarative. In contrast, exemplar theorists are careful to assume that people do not have conscious awareness that they are accessing exemplar memories when making categorization decisions. Thus, the cognitive version of exemplar theory appears to postulate a unique memory system that has not yet been discovered by memory researchers.

It is important to note, however, that other instance-based theories postulate more traditional memory systems. For example, RULEX (Nosofsky, Palmeri, & McKinley, 1994) assumes people use explicit rules during categorization but they memorize exceptions. Presumably, people are aware of these exceptions, so this form of memory seems identical to semantic memory.

Previous attempts at providing a neural interpretation of the cognitive processes postulated by exemplar theory have assigned key roles to the hippocampus and surrounding medial temporal lobe structures (e.g., Pickering, 1997; Sakamoto & Love, 2004). The problem with these attempts is that the evidence supporting a major role for medial temporal lobe structures in category learning is weak.

First, medial temporal lobe interpretations of exemplar theory predict that patients with damage to medial temporal lobe structures should be impaired in category learning. We know of three studies that reported category-learning deficits in amnesiacs (Hopkins, Myers, Shohamy, Grossman, & Gluck, 2004; Kolodny, 1994; Zaki, Nosofsky, Jessup, & Unversagt, 2003), one that reported normal performance on the first 50 trials but impaired performance later on (Knowlton, Squire, & Gluck, 1994), and one that reported normal categorization by amnesiacs when the stimuli were faces, but impaired performance when the stimuli were virtual reality scenes (Graham et al., 2006). On the other hand, many more studies have reported intact category-learning performance in patients with amnesia (e.g. Bayley, Frascino, & Squire, 2005; Filoteo, Maddox, & Davis, 2001b; Janowsky, Shimamura, Kritchevsky, & Squire, 1989; Knowlton & Squire, 1993; Kolodny, 1994; Leng & Parkin, 1988; Squire & Knowlton, 1995; Zaki et al., 2003). For example, Filoteo et al. (2001b) reported normal performance by amnesiacs in a difficult information-integration categorization task with nonlinearly separable categories that required hundreds of training trials. In fact, in the Filoteo et al. (2001b) study, one (medial temporal lobe) amnesiac and one control participant completed a second day of testing. Despite lacking an explicit

memory of the previous session, the patient with amnesia performed slightly better than the control on the first block of day 2. This result suggests that amnesiacs do not necessarily rely on working memory to perform normally in category-learning tasks (because working memory cannot be used to retain category knowledge across days).

A second set of problematic results come from neuroimaging studies of unstructured category-learning tasks. Unstructured categories are those in which the stimuli are assigned to each contrasting category randomly, and thus there is no rule- or similarity-based strategy for determining category membership. Introspection seems to suggest that the only way arbitrary categories of this type could be learned is via explicit memorization, so if medial temporal lobe structures play a critical role in any categorization task, then unstructured categorization tasks seem like a good candidate. Even so, fMRI studies of unstructured-category learning have found task-related activity in the striatum, but typically not in the hippocampus or other medial temporal lobe structures (Lopez-Paniagua & Seger, 2011; Seger & Cincotta, 2005; Seger, Peterson, Cincotta, Lopez-Paniagua, & Anderson, 2010). In addition, unstructured category learning includes a motor component that is more typical of striatal-mediated procedural learning than hippocampal-mediated declarative learning (Crossley, Madsen, & Ashby, 2012).

Third, a number of behavioral dissociations have been reported that seem more consistent with a striatal locus for the learning of similarity-based categories than a medial temporal lobe locus (for a review, see Ashby & Valentin, 2016). This work has contrasted rule-based (RB) and information-integration (II) category-learning tasks. In RB tasks, the categories can be learned via some explicit reasoning process (Ashby, Alfonso-Reese, Turken, & Waldron, 1998). In the most common applications, only one stimulus dimension is relevant, and the participant's task is to discover this relevant dimension and then to map the different dimensional values to the relevant categories. A variety of evidence suggests that success in RB tasks depends on declarative memory and especially on working memory and executive attention (Ashby et al., 1998; Maddox, Ashby, Ing, & Pickering, 2004; Waldron & Ashby, 2001; Zeithamova & Maddox, 2006). In II category-learning tasks, accuracy is maximized with a similarity-based strategy in which information from two or more incommensurable stimulus components is integrated at some predecisional stage (Ashby & Gott, 1988; Ashby et al., 1998). Evidence suggests that success in II tasks depends on procedural memory (Ashby, Ell, & Waldron, 2003; Ashby & Ennis, 2006; Filoteo, Maddox, Salmon, & Song, 2005; Knowlton, Mangels, & Squire, 1996; Maddox, Bohil, & Ing, 2004). Exemplar theory assumes a similarity-based categorization strategy so it seems especially tenable for II tasks. Furthermore, many researchers have argued that learning in RB tasks is mediated by an explicit, rule-learning process that is incompatible with exemplar theory (e.g., Ashby et al., 1998; Erickson & Kruschke, 1998; McDaniel, Cahill, Robbins, & Wiener, 2014; Nosofsky et al., 1994; Rouder & Ratcliff, 2006).

Currently, at least 25 separate empirical dissociations between RB and II category learning have been reported (Ashby & Valentin, 2016). Two of these are especially important for the present purposes. First, II but not RB tasks are extremely sensitive to feedback timing (Maddox, Ashby, & Bohil, 2003; Maddox & Ing, 2005; Worthy, Markman, & Maddox, 2013). In particular, II learning is better when the feedback is delivered 500 ms after the

response than when the feedback is immediate or delivered after a 1 sec delay (Worthy et al., 2013), whereas feedback delays of 2.5 sec or longer completely abolish almost all II learning (Maddox et al., 2003; Maddox & Ing, 2005). In contrast, delays of up to 10 sec have no effect on RB learning. The II results are consistent with the effects on cortical-striatal synaptic plasticity of delays between dopamine (DA) release and Ca^{2+} influx into the spines of medium spiny neurons (Yagishita et al., 2014), and inconsistent with the hypothesis that II learning is mediated by medial temporal lobe structures. Another important dissociation is that switching the locations of the response keys interferes with performance of II tasks but not with performance of one-dimensional RB tasks (Ashby, Ell, & Waldron, 2003; Maddox, Bohil, & Ing, 2004; Maddox, Glass, O'Brien, Filoteo, & Ashby, 2010; Spiering & Ashby, 2008). This is relevant because a similar result has been reported for the most widely studied procedural-learning task – namely the serial reaction time task (Willingham, Wells, Farrell, & Stemwedel, 2000), and procedural learning is thought to depend on the striatum – not on the hippocampus.

This article describes a neural interpretation of exemplar theory that easily accounts for all of these problematic results. First though, we define the exemplar model more explicitly.

The Exemplar Model

Many researchers have tested exemplar models of categorization (e.g., Estes, 1986, 1994; Medin & Schaffer, 1978; Nosofsky, 1986, 2011). These various versions are all highly similar, but to establish mathematical equivalence it is necessary to focus on one specific version of exemplar theory. The model has evolved somewhat from its initial description, so our focus will be on the exemplar model known as the generalized context model (GCM; Nosofsky, 1986, 2011).

The equivalence result established below holds for any number of categories, but to keep the notation simpler we will assume a standard categorization experiment with two categories A and B. Under these conditions, the exemplar model assumes that the probability that a participant assigns stimulus k to category A equals

$$P(\mathrm{A}|k)=\frac{\beta_A \sum_{i\in C_A} V_{iA}\eta_{ik}}{\beta_A \sum_{i\in C_A} V_{iA}\eta_{ik}+\beta_B \sum_{i\in C_B} V_{iB}\eta_{ik}}, \quad (1)$$

where C_A and C_B are sets containing the stimuli in categories A and B, respectively, η_{ik} is the similarity between stimuli i and k, β_A and β_B are constants reflecting the participant's bias toward responding A and B, respectively, and V_{iJ} represents the memory strength of stimulus i with respect to category J. The memory strengths V_{iJ} are not free parameters, but instead are usually "set equal to the relative frequency with which each exemplar i is provided with Category J feedback during the classification training phase" (Nosofsky, 2011). Most categorization experiments present all stimuli equally often, and in this case note that all V_{iJ} are equal and therefore can be canceled from Eq. (1).

Similarity is assumed to be inversely related to the distance between the perceptual representations of the stimuli. More specifically, the distance between the perceptual representations of stimuli i and k, denoted δ_{ik}, is computed from the weighted Minkowski metric:

$$\delta_{ik}=\left(\sum_{j=1}^{m} w_j|x_{ij}-x_{kj}|^r\right)^{1/r}, \quad (2)$$

where m is the number of perceptual dimensions, w_j is the proportion of attention allocated to dimension j, x_{ij} is the coordinate value of stimulus i on the jth perceptual dimension, and r determines the nature of the distance metric. In particular, $r = 1$ produces city-block distance and $r = 2$ produces Euclidean distance. Similarity is inversely related to distance via:

$$\eta_{ik}=\exp(-c\delta_{ik}^{\omega}) \quad (3)$$

where c and ω are constants. The parameter c is a measure of stimulus sensitivity, which increases with the overall discriminability of the stimuli. The constant ω defines the nature of the similarity function. In virtually all applications $\omega = 1$ or 2. A value of $\omega = 1$, which produces the exponential similarity function (Shepard, 1987), is typically combined with a city-block distance metric, whereas a value of $\omega = 2$, which produces the Gaussian similarity function, is typically combined with a Euclidean distance metric.

The Neural Model

Although the evidence does not favor medial temporal lobe structures as the most important regions for category learning, it does favor a critical role for the basal ganglia, and especially the striatum – a major input region within the basal ganglia that includes the caudate nucleus and the putamen (for reviews, see e.g., Ashby & Ennis, 2006; Seger, 2008; Seger & Miller, 2010). A complete review is beyond the scope of this article, but briefly, in addition to the dissociations between RB and II categorization mentioned above, patients with striatal dysfunction are highly impaired in category learning (e.g., Ashby, Noble, Filoteo, Waldron, & Ell, 2003; Filoteo, Maddox, & Davis, 2001a; Filoteo et al., 2005; Knowlton et al., 1996; Sage et al., 2003; Shohamy, Myers, Onlaor, & Gluck, 2004; Witt, Nuhsman, & Deuschl, 2002), and almost all fMRI studies of category learning have reported significant task-related activity in the striatum (e.g., Nomura et al., 2007; Poldrack et al., 2001; Seger & Cincotta, 2002, 2005; Waldschmidt & Ashby, 2011).

So the evidence is good that the striatum plays a key role in category learning – at least in the learning of II and unstructured categories (i.e., rather than RB categories). The problem for exemplar theory is that basal ganglia neuroanatomy does not favor traditional interpretations of exemplar representations. Virtually all of cortex (except V1) sends excitatory (glutamatergic) projections to the striatum (Reiner, 2010). These cortical inputs, which synapse on medium spiny neurons (MSNs), are massively convergent, with estimates that somewhere between 50,000 and 350,000 cortical neurons converge onto a single striatal

MSN (Bolam et al., 2006; Kincaid, Zheng, & Wilson, 1998; C. J. Wilson, 1995). In contrast, each of these cortical neurons might synapse onto as few as 10–100 MSNs (Wickens & Arbuthnott, 2010). Thus, it appears that the resolution of the striatal MSNs is not dense enough to allocate one MSN for every exemplar in any arbitrary category, especially when exemplar similarity is high.

As a result, we propose a reconceptualization of 'exemplar representation'. To our knowledge, there are three existing process-level interpretations of exemplar theory – ALCOVE (Kruschke, 1992), the EBRW (Nosofsky & Palmeri, 1997), and the information-accumulation model (Lamberts, 2000). In each of these, the presentation of a new stimulus adds a new node to the network, and this node serves as the exemplar representation of that stimulus on all future trials. We propose instead that the presentation of a stimulus either changes the synaptic strength at an existing cortical-striatal synapse or creates a new synapse. As a result, categorization training does not recruit any new nodes, rather it alters connectivity between MSNs and units in sensory association cortex.

In the remainder of this section, we show that under appropriate conditions, such a model is mathematically equivalent to the exemplar model described by Eqs. (1), (2), and (3).

Model Architecture

The architecture of the proposed model is shown in Figure 1. Essentially this just reproduces the well-known direct pathway through the basal ganglia, and is identical to the simplest version of the procedural-learning system of the COVIS model of category learning (Ashby et al., 1998; Ashby & Waldron, 1999; Ashby, Paul, & Maddox, 2011; Ashby & Crossley, 2011).

Sensory association cortex is modeled in the same way as in Ashby, Ennis, and Spiering (2007). Briefly, this means we assume an ordered array of many units in sensory cortex, each tuned to a different stimulus. Neurons in sensory cortex respond not only to a preferred stimulus, but also more weakly to similar stimuli. In perceptual neuroscience this phenomenon is described by the neuron's tuning curve. We model the tuning curve of each sensory cortical unit in the Figure 1 model via radial basis functions – that is, we assume that each unit responds maximally when its preferred stimulus is presented and that its response decreases with the distance in perceptual space between the stimulus preferred by that unit and the presented stimulus. The equivalence to exemplar theory holds for any number of these units, so long as every perceptually distinct stimulus maximally excites a different unit. Given the known cortical-striatal convergence, with the two MSNs shown in Figure 1 we would expect the relevant regions of sensory cortex to include somewhere between 10,000 and 60,000 neurons.

The equivalence result established below assumes a standard categorization experiment in which the stimulus is presented at constant intensity and without a mask until the participant responds. Under these conditions, activation in each sensory unit is 0 during periods when no stimulus is displayed and is either 0 or equal to some positive constant value during the duration of stimulus presentation. More specifically, consider a trial when stimulus k is

presented. Then the response of sensory unit i during the time when the stimulus is present will equal:

$$I_{i|k}=\exp(-\delta_{ik}^{\omega}/\gamma), \quad (4)$$

where δ_{ik} is the distance in perceptual space between the representations of stimuli i and k as defined by Eq. 2. When $\omega = 2$, Eq. 4 describes a Gaussian radial basis function and when $\omega = 1$, Eq. 4 describes a Laplacian radial basis function. Either way, Eq. 4 is a popular method for modeling the receptive fields of sensory units, both in models of categorization (e.g., Ashby & Crossley, 2011; Kruschke, 1992) and in other tasks (e.g., Er, Wu, Lu, & Toh, 2002; Oglesby & Mason, 1991; Riesenhuber & Poggio, 1999; Rosenblum, Yacoob, & Davis, 1996).

The key insight that links exemplar theory to cortical-striatal synaptic plasticity comes from Ashby and Alfonso-Reese (1995), who showed that exemplar models are equivalent to a classifier that estimates each category distribution via a Parzen (1962) kernel density estimator. Parzen kernels are known today as radial basis functions, so the key to the Ashby and Alfonso-Reese (1995) result was to recognize that the similarity function proposed by exemplar theory (i.e., Eq. 3) could be interpreted as a radial basis function. This reinterpretation of exemplar similarity provides a natural link between exemplar theory and models such as COVIS that use radial basis functions to model the response properties of sensory cortical units.

The second key feature of exemplar theory is that similarities are summed to determine category membership (i.e., see Eq. 1). Radial basis functions are summed in two different ways in neural models such as COVIS. First, each MSN responds to a weighted sum of the radial basis functions that model activation in sensory cortex[1], and second, the rules that govern synaptic plasticity (i.e., learning) dictate that the current synaptic strength is a weighted sum of previous activations. This latter sum is especially important when establishing equivalence to exemplar theory because it depends on all previously seen exemplars, whereas the former sum does not. So both exemplar theory and COVIS assign a key role to weighted sums of radial basis functions of all previously seen exemplars, and we believe it is this common feature of both theories that cause such similar behavior in the two conceptually different accounts of categorization. The mathematical equivalence established below only requires the latter of these two neural summing operations. Later however, we show that if both neural summing operations are allowed and if many of the ancillary assumptions required for exact mathematical equivalence are relaxed, numerical equivalence is barely affected. We believe this is because in both theories, behavior depends fundamentally on sums of radial basis functions that are activated by stimulus presentation.

At the outset of a new experimental session, we assume that the sensory cortical units are fully interconnected with the two striatal MSNs. Specifically, every unit in sensory cortex

[1]Nonlinearities are introduced downstream of this summing process.

synapses with a dedicated spine on each of the two MSNs. Thus, if there are 40,000 units in sensory cortex there are 40,000 spines on each MSN.

Activation in each MSN is modeled using a firing-rate model (Dayan & Abbott, 2001; Ermentrout & Terman, 2010; H. R. Wilson & Cowan, 1972, 1973). Thus, the two MSNs shown in Figure 1 could be viewed as two identical populations of MSNs, or the activations produced could be viewed as average firing rates over many repetitions of the stimulus conditions. Firing-rate models are described by two equations. The first, typically written as a differential equation, describes how presynaptic input generates postsynaptic activation. This equation typically describes within-trial temporal dynamics and it integrates all sensory input[2], but in the present application, only a very simple model is needed. In particular, there is no need to model the temporal dynamics of within-trial activation or the summing of sensory inputs. Specifically, it suffices to simply assume that the postsynaptic activation in each MSN equals the presynaptic activation at the most active spine weighted by the relevant synaptic strength. The most active spine will be the one synapsing with the cortical unit that responds most strongly to the stimulus. Let $A_J(n)$ denote the activation in striatal unit J (where J = A or B) on trial n. Then if stimulus k is presented on trial n

$$\begin{aligned} A_J(n) &= w_{Jk}(n) I_{k|k} \\ &= w_{Jk}(n), \end{aligned} \quad (5)$$

where $w_{Jk}(n)$ is the strength of the synapse between sensory unit k and MSN J on trial n. The latter equality holds because no matter how distance is defined, the distance from a point to itself must be zero. As a result $\delta_{kk} = 0$, and so $I_{k/k} = 1$ (i.e., for an ω and any γ).

The second firing-rate equation converts the postsynaptic activation into postsynaptic firing rate. This equation models nonlinearities in the neural response (e.g., response compression). Following the standard approach, we assume that the postsynaptic firing rate in unit J, denoted by $R_J(n)$ equals

$$R_J(n) = F[A_J(n)], \quad (6)$$

where F is a monotonically increasing function known as the activation function.

Equivalence to exemplar theory requires that F behaves as a natural log. Thus[3],

[2]This integration leads to the weighted sum of all sensory cortical radial basis functions activated by stimulus presentation.

[3]The natural log has two problematic properties that make it an unusual choice for an activation function. First, firing rates are commonly restricted to the range [0,1] and of course, the natural log is unbounded. Second, the natural log can be negative, whereas activations and firing rates are usually restricted to nonnegative values. Mathematical equivalence to the exemplar model holds whether $R_J(n)$ is positive or negative, but of course the model is more biologically realistic if $R_J(n) \geq 0$. In practice, this can almost always be arranged by replacing Eq. (4) with $I_{i|k} = \theta \exp(-\delta_{ik}^{\omega}/\gamma)$, for some sufficiently large value of θ (when the stimulus is present).

$$R_J(n) = \ln[A_J(n)] = \ln[w_{Jk}(n)]. \quad (7)$$

Figure 1 shows the major neuroanatomical projections from the MSNs to regions in premotor cortex that control the motor response. However, the equivalence result established here treats these as simple relays that convey the signals generated within the striatum to motor output units. Therefore, there is no need to model activity in each of these regions separately. Instead, we simply assume that the MSN firing rates are unaltered as they pass through these relays. However, mathematical equivalence requires that two important processes must occur within premotor cortex. First, independent noise is added to each output unit. We assume that the most active output unit controls the response on each trial, so noise is needed to account for probabilistic responding. Second, a response bias term is added to each MSN output. For equivalence to the exemplar model the additive bias must equal $B = \ln \beta_J$ (see Figure 1). So if we denote the activation in output unit J on trial n as $Y_J(n)$ then

$$Y_J(n) = R_J(n) + \ln \beta_J + \varepsilon_J. \quad (8)$$

We assume that response A is made on trial n if $Y_A(n) > Y_B(n)$ and that response B is made if the opposite ordering occurs.

Learning

DA is known to have pronounced effects on cortical-striatal synaptic plasticity (e.g., Centonze, Picconi, Gubellini, Bernardi, & Calabresi, 2001; Shen, Flajolet, Greengard, & Surmeier, 2008). Much evidence suggests that strengthening of cortical-striatal synapses (and long-term potentiation) requires strong pre- and postsynaptic activation and DA levels above baseline. More specifically, the postsynaptic activation must be strong enough to activate NMDA receptors (a high-threshold glutamate receptor). In contrast, cortical-striatal synapses are weakened (and long-term depression occurs) if pre- and postsynaptic activation are strong and DA is below baseline (e.g., Arbuthnott, Ingham, & Wickens, 2000; Ronesi & Lovinger, 2005). DA levels rise above baseline following unexpected rewards and fall below baseline following the failure to receive an expected reward (Hollerman & Schultz, 1998; Mirenowicz & Schultz, 1994; Schultz, 1998). As a result, a number of researchers have proposed that synaptic plasticity at cortical-striatal synapses follows reinforcement-learning rules (Doya, 2000; Houk, Adams, & Barto, 1995).

Suppose every sensory-cortical neuron has one synapse onto each striatal MSN. A biologically motivated form of reinforcement learning assumes that if stimulus S_k is presented on trial n, then the strength of the synapse between striatal unit J (for J = A or B) and sensory-cortical unit k on trial $n + 1$ equals (e.g., Ashby & Crossley, 2011):

$$w_{Jk}(n+1)=w_{Jk}(n) +\alpha_w I_{k|S_n}[R_J(n)-\theta_{NMDA}]^+[D(n)-D_{base}]^+ -\beta_w I_{k|S_n}[R_J(n)-\theta_{NMDA}]^+[D_{base}-D(n)]^+, \quad (9)$$

where α_w and β_w are the constant learning and unlearning rates, respectively, $[f(n)]^+ = f(n)$ if $f(n) > 0$ and 0 if $f(n) \leq 0$, θ_{NMDA} is a constant specifying the threshold for NMDA receptor activation, $D(n)$ is the amount of DA released on trial n, and D_{base} is the baseline DA level. Note that this model assumes that the amount of synaptic strengthening (i.e., the plus term) is proportional to the product of 1) presynaptic activation, 2) the amount that the postsynaptic activation is above the NMDA threshold, and 3) the amount that DA is above baseline. In contrast, the amount of synaptic weakening (i.e., the minus term) is proportional to the product of 1) presynaptic activation, 2) the amount that the postsynaptic activation is above the NMDA threshold, and 3) the amount that DA is below baseline.

Solving Eq. (9) iteratively produces

$$w_{Jk}(n+1)=w_{Jk}(0) +\alpha_w \sum_{i=1}^{n} I_{k|S_i}[R_J(i)-\theta_{NMDA}]^+[D(i)-D_{base}]^+ -\beta_w \sum_{i=1}^{n} I_{k|S_i}[R_J(i)-\theta_{NMDA}]^+[D_{base}-D(i)]^+. \quad (10)$$

Note that this solution includes a sum of radial basis functions activated by all previously seen exemplars. Since the exemplar model includes a similar such sum, this is the key feature of the neural model that allows it to mimic the exemplar model.

Assumptions

Proving that this model is equivalent to the exemplar model requires the following extra assumptions.

A1. Error trials do not change synaptic strengths—This assumption implies that $\beta_w = 0$ in Eq. (9). Although this assumption is biologically implausible (e.g., Calabresi, Maj, Pisani, Mercuri, & Bernardi, 1992), it is not necessarily imcompatible with exemplar theory. As noted above, exemplar theory assumes that the memory strength of exemplar i in category J is strengthened when a categorization response to exemplar i is provided with category J feedback during classification training (Nosofsky, 2011). On trials when the response is correct this is clear. Positive feedback unambiguously signals that the stimulus belongs to the category associated with the participant's response. However, on error trials this is not so clear. Negative feedback typically only signals that the stimulus does not belong to the responded category. In the two-category case an inference can be made about the correct category membership of the stimulus, but when there are more than two categories, then no such inference is possible. As a result, exemplar theory also seems to predict no learning on error trials – at least in experiments with more than two categories.

A2. Only the spines on the striatal unit associated with the most active motor unit are eligible for synaptic plasticity—Assumptions A1 and A2 are clearly over-simplifications. However, note that they tend to offset each other – at least to a certain extent. Together they tend to ensure that active synapses on the MSN associated with the incorrect response (call this the incorrect MSN) never change strength. In the absence of these two assumptions, those synapses would get strengthened on trials when the correct MSN controls the response and weakened on trials when the incorrect MSN controls the response. Another interpretation of Assumptions A1 and A2, therefore, is that this strengthening and weakening cancel each other out.

A3. All initial synaptic strengths are negligible – just large enough to allow postsynaptic activation to the first stimulus presentation, but small enough so that mathematically we can assume that $w_{Ai}(0) = w_{Bi}(0) = 0$, for all i—A justification for this assumption is that the presentation of a novel stimulus that has no previous reward association causes DA release (e.g., Horvitz, Stewart, & Jacobs, 1997; Wickelgren, 1997), and increased DA levels potentiate the postsynaptic effects of glutamate (Ashby & Casale, 2003). Thus, available evidence suggests that the first presentation of a stimulus during an experimental session is likely to cause an uncharacteristically large striatal response.

It is also important to note that even without this assumption the contribution of $w_{Ai}(0)$ and $w_{Bi}(0)$ to $w_{Ai}(n)$ and $w_{Bi}(n)$ decreases as n increases. In other words, the effects of the initial weights on the asymptotic performance of the model are negligible. This is important because the exemplar model is a model of asymptotic performance – it was never proposed as a model of initial learning. So although Assumption A3 is necessary for strict mathematical equivalence, at the practical level this assumption is not critical.

A4. In whichever unit controls the response, $[R_J(n) - \theta_{NMDA}]^+[D(n) - D_{base}]^+ = K$ for all n, where K is a constant—In general, we expect $[R_J(n) - \theta_{NMDA}]^+$ to increase with n because the synaptic strength associated with its sensory input should increase as a result of training. In contrast, virtually all current models predict that $[D(n) - D_{base}]^+$ will decrease with n because the rewards become more predictable as training progresses. The evidence is good that DA neurons respond to the reward prediction error (Schultz, 2002; Schultz, Dayan, & Montague, 1997) – defined as the value of the obtained reward minus the value of the predicted reward. Thus, as learning progresses, accuracy rises and so does the ability to predict the feedback valence. As a result, the amount by which DA levels rise following positive feedback should decrease. Thus, an alternative interpretation of assumption A4 is that $[R_J(n) - \theta_{NMDA}]^+$ increases with n at the same rate that $[D(n) - D_{base}]^+$ decreases with n.

A5. The noise terms ε_J in Eq. (8) are independent random samples from identical double exponential distributions—A double exponential distribution is required for equivalence because probabilities associated with the maximum of double exponentially distributed random variables satisfy the relative goodness rule of Eq. (1) (Yellott, 1977). In particular, suppose $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are a set of independent random variables with identical double exponential distributions. Then (Yellott, 1977)

$$P\left(u_k+\varepsilon_k=\max_{i=1}^{n}\{u_i+\varepsilon_i\}\right)=\frac{e^{u_k}}{\sum_{i=1}^{n}e^{u_i}}. \quad (11)$$

Equivalence to the Exemplar Model

In this section, we show that the model sketched in Figure 1 makes identical quantitative predictions as the exemplar model described by Eqs. (1), (2), and (3), given the reinforcement-learning model described earlier and under the assumptions outlined in the previous section. To keep the notation simple, we will demonstrate the equivalence for a two-category task, but the equivalence holds for any number of categories.

Consider a two-category task in which category A contains the M_A exemplars $C_A = \{a_1, a_2, \ldots, aM_A\}$ and category B contains the MB exemplars $C_B = \{b_1, b_2, \ldots, bM_B\}$. Note that there are four kinds of trials: correct A response trials, correct B response trials, incorrect A response trials, and incorrect B response trials. By assumption 1, neither type of incorrect response trial will change any synaptic weights. So first consider correct A response trials – that is, trials when the stimulus belongs to the set C_A and motor unit A is more active than motor unit B. By assumption 2, no synaptic strengths on unit B will change. So correct A response trials will only cause changes in the strength of synapses on unit A. Similarly, correct B response trials will only cause changes in the strength of cortical-striatal synapses on unit B.

By assumptions A1 – A4, the reinforcement-learning model described by Eq. (10) predicts that the strength of the synapse between sensory unit k and striatal unit A reduces to

$$w_{Ak}(n_A+1)=\alpha_w K\sum_{i=1}^{n_A} I_{k|S_i}, \quad (12)$$

where $i = 1, \ldots, n_A$ denotes the first n_A correct A response trials. Suppose that of these n_A trials, stimulus a_i was presented n_{aj} times (so $\sum_{i=1}^{M_A} n_{a_i}=n_A$, where MA is the number of exemplars in category A). Note that $I_{k|S_i}$ equals the response of sensory unit k on a trial when stimulus S_i is presented. By Eq. (4) this equation becomes

$$\begin{aligned} w_{Ak}(n_A+1)&=\alpha_w K\sum_{i=1}^{n_A}\exp(-\delta^{\omega}_{kS_i}/\gamma)\\ &=\alpha_w K\sum_{j=1}^{M_A} n_{a_j}\exp(-\delta^{\omega}_{kj}/\gamma)\\ &=\alpha_w K\sum_{j\in C_A} n_{a_j}\exp(-\delta^{\omega}_{kj}/\gamma). \end{aligned} \quad (13)$$

Of course, by identical reasoning, a similar equation will hold for the synapses on striatal unit B.

Now consider the very next trial after all this training. Suppose some stimulus k is presented. Then by Eq. (7) the firing rate in striatal unit A will equal

$$\begin{aligned} R_A &= \ln[w_{Ak}] \\ &= \ln\left[\alpha_w K \sum_{j \in C_A} n_{a_j} \exp(-\delta_{kj}^{\omega}/\gamma)\right]. \end{aligned} \tag{14}$$

Similarly,

$$R_B = \ln\left[\alpha_w K \sum_{j \in C_B} n_{b_j} \exp(-\delta_{kj}^{\omega}/\gamma)\right]. \tag{15}$$

By Eq. (8), activation in the output units will equal

$$\begin{aligned} Y_A &= R_A + \ln\beta_A + \varepsilon_A \\ &= \ln\left[\alpha_w K \sum_{j \in C_A} n_{a_j} \exp(-\delta_{kj}^{\omega}/\gamma)\right] + \ln\beta_A + \varepsilon_A \\ &= \ln\left[\alpha_w K \beta_A \sum_{j \in C_A} n_{a_j} \exp(-\delta_{kj}^{\omega}/\gamma)\right] + \varepsilon_A, \end{aligned} \tag{16}$$

and

$$Y_B = \ln\left[\alpha_w K \beta_B \sum_{j \in C_B} n_{b_j} \exp(-\delta_{kj}^{\omega}/\gamma)\right] + \varepsilon_B. \tag{17}$$

Finally, by assumption A5

$$P(\mathrm{A}|k)=\frac{\alpha_w K\beta_A \sum_{j\in C_A} n_{a_j}\exp(-\delta^{\omega}_{kj}/\gamma)}{\alpha_w K\beta_A \sum_{j\in C_A} n_{a_j}\exp(-\delta^{\omega}_{kj}/\gamma)+\alpha_w K\beta_B \sum_{j\in C_B} nb_j\exp(-\delta^{\omega}_{kj}/\gamma)}$$
$$=\frac{\beta_A \sum_{j\in C_A} n_{a_j}\exp(-\delta^{\omega}_{kj}/\gamma)}{\beta_A \sum_{j\in C_A} n_{a_j}\exp(-\delta^{\omega}_{kj}/\gamma)+\beta_B \sum_{j\in C_B} n_{b_j}\exp(-\delta^{\omega}_{kj}/\gamma)}, \quad (18)$$

which is clearly equivalent to the the exemplar model described by Eqs. (1), (2), and (3), with $n_{a_j} = V_{jA}$, $n_{b_j} = V_{jB}$, and $\gamma = 1/c$.

Effects of Relaxing Assumptions

As is generally always the case, proving exact mathematical equivalence requires strong assumptions. In particular, assumptions A1, A2, and A4 seem biologically unreasonable, and assumption A5 seems arbitrary. A natural question to ask therefore is to what extent equivalence depends on these assumptions. As mentioned earlier, we believe that the most important condition for equivalence is that both models assume that categorization responses are largely determined by weighted sums of radial basis functions of all previously seen exemplars. According to this view, relaxing the ancillary assumptions needed for equivalence might affect the initial learning trajectory of the neural model, but is unlikely to profoundly affect the model's predictions about asymptotic performance. This section tests that prediction.

To examine the importance of assumptions A1, A2, A4, and A5 we conducted a number of simulations of the neural model in the absence of those assumptions. We call the model that satisfies all assumptions and is mathematically equivalent to exemplar theory the exemplar-equivalent model, and the more biologically realistic version that satisfies all assumptions of the exemplar-equivalent model except assumptions A1, A2, A4, and A5 the biologically-plausible model. Specifically, the biologically-plausible model was essentially identical to the procedural-learning component of the COVIS model described by Ashby et al. (2011). It differs from the exemplar-equivalent model in the following ways:

1. The biologically-plausible model changed synaptic strengths on all trials. So this version of the model was not constrained by assumption A1. The amount of synaptic weakening on error trials was set equal to the amount of synaptic strengthening on correct trials (i.e., β was set equal to α in Eq. 9).
2. All cortical-striatal spines were eligible for synaptic plasticity on every trial. In particular, synaptic strengths on both MSNs were modified on every trial. So this version was not constrained by assumption A2.
3. No assumptions were made about the relationship between the striatal and DA responses. So this version was not constrained by assumption A4. In particular,

DA release was set to the same piecewise linear function of reward prediction error as in Eqs. 11 – 13 of Ashby et al. (2011).

4. In the exemplar-equivalent model, double-exponentially distributed noise was added to the activations of the premotor units. In the biologicaly-plausible model, no noise was added to the premotor units and normally distributed noise was added to the MSN activations.

Our general goal was to compare asymptotic performance of the exemplar-equivalent and biologically-plausible models, and if these differed, to ask whether the altered structure of the biologically-plausible model was consistent or inconsistent with the predictions of exemplar theory. To begin, it was important to set parameter values in the exemplar-equivalent model to values that are typical of previous applications of exemplar theory. To satisfy this constraint, we did a crude search to find parameter values that allowed the exemplar-equivalent model to reproduce the 16 × 2 confusion matrices predicted by the GCM in the Dimensional, Criss-Cross, Interior-Exterior, and Diagonal transfer conditions reported by Nosofsky (1986)[4].

The critical question now is how much these predictions will change when we switch to the biologically more plausible version of the model. It turns out that the predictions change only a small amount. The r^2 between the predictions of the exemplar-equivalent and biologically-plausible models was .999, .998, .981, and .955, for the Dimensional, Criss-Cross, Interior-Exterior, and Diagonal conditions respectively. Thus, the only condition where the predictions changed by a non-negligible amount was the Diagonal condition. There are two possibilities here. Either the biologically-plausible model fundamentally changed the structure of the data in a way this is incompatible with the exemplar model, or it changed the predictions in a way that is consistent with exemplar theory but best accounted for by some slight change in one or more parameter values. To answer this question, we fit the GCM to the confusion matrices predicted by the biologically-plausible model. If the biologically-plausible model changed the structure of the data in a way that is incompatible with exemplar theory, then the GCM will fit poorly. In fact, the GCM provided excellent fits in all four conditions. The r^2 between the predictions of the biologically-plausible model and the GCM was .999, .999, .981, and .990, for the Dimensional, Criss-Cross, Interior-Exterior, and Diagonal conditions, respectively.

In summary, although exact mathematical equivalence requires all assumptions described above, assumptions A1, A2, A4, and A5 have little numerical effect on the asymptotic predictions of the model.

[4]The model was run through the same sequence of trials as the Nosofsky (1986) participants – that is, 1200 training and 3500 transfer trials for each category structure. We fit the model to the average of the Subject 1 and Subject 2 GCM predicted confusion matrices, and we used the coordinates of each stimulus in physical, rather than perceptual space. This is because our goal was not to reproduce exactly what Nosofsky (1986) did, but rather to produce predictions of the exemplar-equivalent model that were representative of a typical exemplar model application. In this sense we were successful because when we fit the exemplar model (i.e., the GCM) to the confusion matrices that resulted from this process, r^2 was .996, .996, .994, and .993 for the Dimensional, Criss-Cross, Interior-Exterior, and Diagonal conditions, respectively.

Discussion

The neural model described in this article makes very different psychological assumptions than are usually associated with exemplar theory, yet it is mathematically equivalent to the exemplar model. The cognitive version of exemplar theory assumes people retrieve memory representations of previously seen exemplars and that they compare the presented stimulus to these stored memories. Every new stimulus creates a new memory representation. In contrast, the neural version assumes that no memory representations are retrieved. For example, the neural model assumes that stimulus presentation always activates two striatal MSNs (in two-category tasks), regardless of the number of exemplars in each category and regardless of the number of training trials. Instead of adding a new exemplar memory representation, the neural version of exemplar theory assumes that presentation of a new stimulus either modifies the strength of an existing cortical-striatal synapse or creates a new synapse (e.g., by adding a new spine to an existing MSN). Thus, the exemplar representation in the neural version of exemplar theory is either the strength of a cortical-striatal synapse or perhaps it could be interpreted as a dedicated spine on an MSN. The more important point however, is that stimulation of that MSN (or that dendritic spine) would not retrieve an exemplar memory, but rather create an urge to respond either A or B, and in this sense the neural version makes different psychological assumptions than classical accounts of exemplar theory[5].

Mathematical equivalence to the (GCM) exemplar model requires a number of strong assumptions that in some cases are biologically implausible. Even so, we showed via simulation that these assumptions have little effect on the asymptotic predictions of the model. Thus, the neural model described here mimics the numerical predictions of exemplar theory when the biologically implausible assumptions are replaced by assumptions that are more biologically realistic. We believe that the predictions of the neural model are robust with respect to violations of these ancillary assumptions because in both models the probability that a stimulus is assigned to a particular category increases with the weighted sum of radial basis functions activated by all previously seen exemplars from that category. The ancillary assumptions do not change this fundamental property of the neural model, and since this same property is shared by exemplar theory, the two psychologically different models make similar (or even identical) quantitative predictions.

Because the two models make identical or nearly identical quantitative predictions, any previous evidence in favor of the exemplar model that is based on its success in providing good quantitative fits to categorization data is also evidence in favor of the neural model. But in addition, the neural model is able to account for many empirical phenomena that are either incompatible with or outside the scope of the cognitive version of exemplar theory. This includes all the problematic data reviewed earlier in this article.

First, because the neural model assumes that the critical site of category learning is at cortical-striatal synapses, it easily accounts for the many reports that patients with striatal

[5]Note however, that many parameters have the same meaning in both accounts, including the response biases β_J, the attention weights w_j, and overall discriminability c.

dysfunction are impaired in category learning. Second, for the same reason, it accounts for the common fMRI finding of categorization-related activation in the striatum (Lopez-Paniagua & Seger, 2011; Nomura et al., 2007; Poldrack et al., 2001; Seger & Cincotta, 2002, 2005; Seger et al., 2010; Waldschmidt & Ashby, 2011). Third, it accounts for the many empirical dissociations that have been reported between RB and II categorization tasks. For example, it accounts for the sensitivity of II category learning to feedback delays because of the known sensitivity of cortical-striatal plasticity to delays between DA release and activity at cortical-striatal synapses (Valentin, Maddox, & Ashby, 2014; Yagishita et al., 2014). It also accounts for the interference that occurs in II and unstructured categorization performance when the location of the response buttons is switched – because the terminal nodes in the Figure 1 model are in premotor cortex, rather than prefrontal cortex. Fourth, it accounts for deficits by patients with Parkinson's disease in II and unstructured category learning (Hélie, Paul, & Ashby, 2012).

If more anatomical details are added into the model of the striatum (as in Ashby & Crossley, 2011; Cantwell, Crossley, & Ashby, 2015; Crossley, Ashby, & Maddox, 2013), then the neural version of exemplar theory can also account for a wide variety of other phenomena including i) single-unit recordings from MSNs and TANs during instrumental conditioning (Ashby & Crossley, 2011); ii) many behavioral phenomena from instrumental conditioning experiments, including fast reacquisition after extinction, the partial reinforcement extinction effect, spontaneous recovery, and renewal (Crossley, Horvitz, Balsam, & Ashby, 2016); iii) the result that recovery from a full reversal is quicker than learning new categories constructed from the same stimuli (Cantwell et al., 2015); and iv) unlearning and failures of unlearning (e.g., renewal) during II categorization (Crossley et al., 2013; Crossley, Ashby, & Maddox, 2014).

David Marr (1982) famously described three levels of mathematical modeling, which he referred to as computational, algorithmic, and implementational. Computational models (often called descriptive models in psychology) make quantitative predictions, but do not describe the algorithms that produce those predictions. Algorithmic models (often called process models in psychology) describe the algorithms, but not the architecture that implements those algorithms. At the lowest level, implementational models describe the architecture that implements the algorithms that produce the quantitative predictions. The original versions of exemplar theory – for example, the version described by Eqs. (1), (2), and (3) – were computational-level descriptions of asymptotic categorization behavior (i.e., the theory made no attempt to account for learning). Cognitive process was used to motivate the quantitative predictions, but no attempt was made to model those processes. To see this, note for example, that Eq. (1) makes no predictions about what response a participant will make on any single trial of a categorization experiment. Rather it simply describes the relative proportion of times the participant will give every possible response under the ideal conditions in which the same categorization trial is repeated an infinite number of times.

As one moves down Marr's hierarchy, it is generally true that more than one interpretation is possible. For example, we know of three different algorithmic-level versions of exemplar theory (Kruschke, 1992; Lamberts, 2000; Nosofsky & Palmeri, 1997), which all make slightly different assumptions. The neural model proposed here is the first known

implementational-level version, and it shows that some qualitatively different psychological assumptions are compatible with the quantitative predictions of computational-level versions of exemplar theory. One of the main reasons for developing lower-level versions of any theory is to account for more data. The algorithmic-level versions of exemplar theory can account for response time and learning data that are outside the scope of the computational-level version. And the neural version proposed here can account for many phenomena that are outside the scope of the algorithmic-level versions. Although many other neural versions may be theoretically possible, the many extra constraints provided by all these neuroscience-related phenomena would seem to narrow the set of candidate neural interpretations to a model that assigns a major role to the striatum. If so, then any alternative neural interpretation that is also consistent with the available neuroscience-related data should look qualitatively similar to the version proposed here (e.g., since both models would be constrained by the same basal ganglia neuroanatomy).

The neural version of exemplar theory is essentially a simplified version of the procedural-learning component of the COVIS model of category learning[6] (Ashby & Crossley, 2011; Ashby et al., 2007; Cantwell et al., 2015). The COVIS model includes much more biological detail (e.g., spiking MSNs, striatal cholinergic interneurons), other brain areas (e.g., interlaminar thalamic nuclei), a more sophisticated reinforcement-learning model (e.g., with rate-limiting terms that constrain synaptic strengths to a fixed interval $[0, w_{max}]$), and it does not make any of assumptions A1 – A5. Even so, it is important to note that the equivalence established here greatly benefits both theories. Of course it provides exemplar theory access to a huge range of neuroscience-related phenomena that generally are outside the scope of any purely cognitive theory. But it also greatly benefits COVIS. First, it means that successes of the exemplar model in providing good quantitative fits to categorization data imply that COVIS should be equally successful at accounting for those data. Second, it provides a method to quickly fit COVIS to asymptotic response proportions collected in the course of a categorization experiment. Fitting the COVIS model currently requires time-consuming Monte Carlo simulations and therefore, exploring its predictions is a challenging computational process. In contrast, the exemplar model is simple enough that it can quickly be fit to data in a straightforward manner using standard optimization algorithms. Therefore, because of the equivalence established here, asymptotic behavioral predictions of COVIS can be quickly tested by fitting the exemplar model to the empirical confusion matrices.

Finally, this article makes one more equally important contribution. The equivalence of exemplar theory and the procedural-learning component of COVIS means that it is probably fruitless to attempt to test between these two models by comparing goodness-of-fits in any categorization experiment. Instead, the two historically disparate approaches to categorization modeling should be used together to create a more powerful armamentarium that can be used to improve our overall understanding of categorization behavior.

[6]COVIS assumes this procedural-learning component dominates in unstructured and II category-learning tasks, but that an explicit-learning system controls behavior in RB tasks. Exemplar theory was proposed before there was any evidence or inclination that humans might have multiple category-learning systems. In the interim, some exemplar theorists have proposed a second rule-learning system (e.g., Erickson & Kruschke, 1998; Nosofsky et al., 1994). Of these, perhaps most similar to COVIS is ATRIUM (Erickson & Kruschke, 1998), which includes two sub-models – an explicit rule-learning model that would dominate in RB tasks and an exemplar model that would dominate in unstructured and II tasks.

Acknowledgments

This research was supported by NIH grant 2R01MH063760.

References

Arbuthnott G, Ingham C, Wickens J. Dopamine and synaptic plasticity in the neostriatum. Journal of Anatomy. 2000; 196(4):587–596. [PubMed: 10923989]

Ashby FG, Alfonso-Reese LA. Categorization as probability density estimation. Journal of Mathematical Psychology. 1995; 39(2):216–233.

Ashby FG, Alfonso-Reese LA, Turken AU, Waldron EM. A neuropsychological theory of multiple systems in category learning. Psychological Review. 1998; 105(3):442–481. [PubMed: 9697427]

Ashby FG, Casale MB. A model of dopamine modulated cortical activation. Neural Networks. 2003; 16(7):973–984. [PubMed: 14692632]

Ashby FG, Crossley MJ. A computational model of how cholinergic interneurons protect striatal-dependent learning. Journal of Cognitive Neuroscience. 2011; 23(6):1549–1566. [PubMed: 20521851]

Ashby FG, Ell SW, Waldron EM. Procedural learning in perceptual categorization. Memory & Cognition. 2003; 31(7):1114–1125. [PubMed: 14704026]

Ashby FG, Ennis JM. The role of the basal ganglia in category learning. Psychology of Learning and Motivation. 2006; 46:1–36.

Ashby FG, Ennis JM, Spiering BJ. A neurobiological theory of automaticity in perceptual categorization. Psychological Review. 2007; 114(3):632–656. [PubMed: 17638499]

Ashby FG, Gott RE. Decision rules in the perception and categorization of multidimensional stimuli. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1988; 14:33–53.

Ashby FG, Noble S, Filoteo JV, Waldron EM, Ell SW. Category learning deficits in parkinson's disease. Neuropsychology. 2003; 17(1):115–124. [PubMed: 12597080]

Ashby, FG., Paul, EJ., Maddox, WT. COVIS. In: Pothos, EM., Wills, A., editors. Formal approaches in categorization. New York: Cambridge University Press; 2011. p. 65-87.

Ashby, FG., Valentin, VV. Multiple systems of perceptual category learning: Theory and cognitive tests. In: Cohen, H., Lefebvre, C., editors. Handbook of categorization in cognitive science, second edition. New York: Elsevier; 2016. in press

Ashby FG, Waldron EM. On the nature of implicit categorization. Psychonomic Bulletin & Review. 1999; 6(3):363–378. [PubMed: 12198775]

Bayley PJ, Frascino JC, Squire LR. Robust habit learning in the absence of awareness and independent of the medial temporal lobe. Nature. 2005; 436(7050):550–553. [PubMed: 16049487]

Bolam, J., Bergman, H., Graybiel, A., Kimura, M., Plenz, D., Seung, H., … Wickens, J. Microcircuits, molecules and motivated behaviour: Microcircuits in the striatum. In: Grillner, S., Graybiel, AM., editors. Microcircuits: The interface between neurons and global brain function. Cambridge, MA: MIT Press; 2006. p. 165-190.

Brooks, LR. Nonanalytic concept formation and memory for instances. In: Rosch, E., Lloyd, BB., editors. Cognition and categorization. Hillsdale, NJ: Erlbaum; 1978. p. 169-211.

Calabresi P, Maj R, Pisani A, Mercuri NB, Bernardi G. Long-term synaptic depression in the striatum: physiological and pharmacological characterization. Journal of Neuroscience. 1992; 12(11):4224–4233. [PubMed: 1359031]

Cantwell G, Crossley MJ, Ashby FG. Multiple stages of learning in perceptual categorization: Evidence and neurocomputational theory. Psychonomic Bulletin & Review. 2015; 22:1598–1613. [PubMed: 25917141]

Centonze D, Picconi B, Gubellini P, Bernardi G, Calabresi P. Dopaminergic control of synaptic plasticity in the dorsal striatum. European Journal of Neuroscience. 2001; 13(6):1071–1077. [PubMed: 11285003]

Crossley MJ, Ashby FG, Maddox WT. Erasing the engram: The unlearning of procedural skills. Journal of Experimental Psychology: General. 2013; 142(3):710–741. [PubMed: 23046090]

Crossley MJ, Ashby FG, Maddox WT. Context-dependent savings in procedural category learning. Brain and Cognition. 2014; 92:1–10.

Crossley MJ, Horvitz JC, Balsam PD, Ashby FG. Expanding the role of striatal cholinergic interneurons and the midbrain dopamine system in appetitive instrumental conditioning. Journal of Neurophysiology. 2016; 115(1):240–254. [PubMed: 26467514]

Crossley MJ, Madsen NR, Ashby FG. Procedural learning of unstructured categories. Psychonomic Bulletin & Review. 2012; 19(6):1202–1209. [PubMed: 22965328]

Dayan, P., Abbott, LF. Theoretical neuroscience: Computational and mathematical modeling of neural systems. Cambridge, MA: MIT Press; 2001.

Doya K. Complementary roles of basal ganglia and cerebellum in learning and motor control. Current Opinion in Neurobiology. 2000; 10(6):732–739. [PubMed: 11240282]

Er MJ, Wu S, Lu J, Toh HL. Face recognition with radial basis function (rbf) neural networks. IEEE Transactions on Neural Networks. 2002; 13(3):697–710. [PubMed: 18244466]

Erickson MA, Kruschke JK. Rules and exemplars in category learning. Journal of Experimental Psychology: General. 1998; 127(2):107–140. [PubMed: 9622910]

Ermentrout, GB., Terman, DH. Mathematical foundations of neuroscience. New York: Springer Science & Business Media; 2010.

Estes WK. Array models for category learning. Cognitive Psychology. 1986; 18(4):500–549. [PubMed: 3769427]

Estes, WK. Classification and cognition. New York: Oxford University Press; 1994.

Filoteo JV, Maddox WT, Davis JD. A possible role of the striatum in linear and nonlinear category learning: Evidence from patients with hungtington's disease. Behavioral Neuroscience. 2001a; 115(4):786–798. [PubMed: 11508718]

Filoteo JV, Maddox WT, Davis JD. Quantitative modeling of category learning in amnesic patients. Journal of the International Neuropsychological Society. 2001b; 7:1–19. [PubMed: 11253835]

Filoteo JV, Maddox WT, Salmon DP, Song DD. Information-integration category learning in patients with striatal dysfunction. Neuropsychology. 2005; 19(2):212–222. [PubMed: 15769205]

Graham KS, Scahill VL, Hornberger M, Barense MD, Lee AC, Bussey TJ, Saksida LM. Abnormal categorization and perceptual learning in patients with hippocampal damage. The Journal of Neuroscience. 2006; 26(29):7547–7554. [PubMed: 16855082]

Hélie S, Paul EJ, Ashby FG. A neurocomputational account of cognitive deficits in parkinson's disease. Neuropsychologia. 2012; 50(9):2290–2302. [PubMed: 22683450]

Hollerman JR, Schultz W. Dopamine neurons report an error in the temporal prediction of reward during learning. Nature Neuroscience. 1998; 1(4):304–309. [PubMed: 10195164]

Hopkins RO, Myers EC, Shohamy D, Grossman S, Gluck M. Impaired probabilistic category learning in hypoxic subjects with hippocampal damage. Neuropsychologia. 2004; 42:524–535. [PubMed: 14728924]

Horvitz JC, Stewart T, Jacobs BL. Burst activity of ventral tegmental dopamine neurons is elicited by sensory stimuli in the awake cat. Brain Research. 1997; 759(2):251–258. [PubMed: 9221945]

Houk, J., Adams, J., Barto, A. A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: Houk, JC.Ldavis, J., Gbeiser, D., editors. Models of information processing in the basal ganglia. Cambridge, MA: MIT Press; 1995. p. 249-270.

Janowsky JS, Shimamura AP, Kritchevsky M, Squire LR. Cognitive impairment following frontal lobe damage and its relevance to human amnesia. Behavioral Neuroscience. 1989; 103:548–560. [PubMed: 2736069]

Kincaid AE, Zheng T, Wilson CJ. Connectivity and convergence of single corticostriatal axons. The Journal of Neuroscience. 1998; 18(12):4722–4731. [PubMed: 9614246]

Knowlton BJ, Mangels JA, Squire LR. A neostriatal habit learning system in humans. Science. 1996; 273(5280):1399–1402. [PubMed: 8703077]

Knowlton BJ, Squire LR. The learning of natural categories: Parallel memory systems for item memory and category-level knowledge. Science. 1993; 262:1747–1749. [PubMed: 8259522]

Knowlton BJ, Squire LR, Gluck MA. Probabilistic classification learning in amnesia. Learning & Memory. 1994; 1(2):106–120. [PubMed: 10467589]

Kolodny JA. Memory processes in classification learning: An investigation of amnesic performance in categorization of dot patterns and artistic styles. Psychological Science. 1994; 5:164–169.

Kruschke JK. Alcove: An exemplar-based connectionist model of category learning. Psychological Review. 1992; 99(1):22–44. [PubMed: 1546117]

Lamberts K. Information-accumulation theory of speeded categorization. Psychological Review. 2000; 107(2):227–260. [PubMed: 10789196]

Leng NR, Parkin AJ. Double dissociation of frontal dysfunction in organic amnesia. British Journal of Clinical Psychology. 1988; 27:359–362. [PubMed: 3214689]

Lopez-Paniagua D, Seger CA. Interactions within and between corticostriatal loops during component processes of category learning. Journal of Cognitive Neuroscience. 2011; 23(10):3068–3083. [PubMed: 21391766]

Maddox WT, Ashby FG, Bohil CJ. Delayed feedback effects on rule-based and information-integration category learning. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2003; 29:650–662.

Maddox WT, Ashby FG, Ing AD, Pickering AD. Disrupting feedback processing interferes with rule-based but not information-integration category learning. Memory & Cognition. 2004; 32(4):582–591. [PubMed: 15478752]

Maddox WT, Bohil CJ, Ing AD. Evidence for a procedural-learning-based system in perceptual category learning. Psychonomic Bulletin & Review. 2004; 11(5):945–952. [PubMed: 15732708]

Maddox WT, Glass BD, O'Brien JB, Filoteo JV, Ashby FG. Category label and response location shifts in category learning. Psychological Research. 2010; 74(2):219–236. [PubMed: 19471959]

Maddox WT, Ing AD. 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. 2005; 31(1):100–107.

Marr, D. Vision: A computational investigation into the human representation and processing of visual information. New York: Freeman; 1982.

McDaniel MA, Cahill MJ, Robbins M, Wiener C. Individual differences in learning and transfer: Stable tendencies for learning exemplars versus abstracting rules. Journal of Experimental Psychology: General. 2014; 143(2):668–693. [PubMed: 23750912]

Medin DL, Schaffer MM. Context theory of classification learning. Psychological Review. 1978; 85(3):207–238.

Mirenowicz J, Schultz W. Importance of unpredictability for reward responses in primate dopamine neurons. Journal of Neurophysiology. 1994; 72(2):1024–1027. [PubMed: 7983508]

Nomura E, Maddox W, Filoteo J, Ing A, Gitelman D, Parrish T, … Reber P. Neural correlates of rule-based and information-integration visual category learning. Cerebral Cortex. 2007; 17(1):37–43. [PubMed: 16436685]

Nosofsky RM. Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General. 1986; 115:39–57. [PubMed: 2937873]

Nosofsky, RM. The generalized context model: An exemplar model of classification. In: Pothos, EM., Wills, A., editors. Formal approaches in categorization. New York: Cambridge University Press; 2011. p. 18-39.

Nosofsky RM, Palmeri TJ. An exemplar-based random walk model of speeded classification. Psychological Review. 1997; 104(2):266–300. [PubMed: 9127583]

Nosofsky RM, Palmeri TJ, McKinley SC. Rule-plus-exception model of classification learning. Psychological Review. 1994; 101(1):53–79. [PubMed: 8121960]

Oglesby, J., Mason, J. Radial basis function networks for speaker recognition. 1991 international conference on acoustics, speech, and signal processing, icassp-91; 1991. p. 393-396.

Parzen E. On estimation of a probability density function and mode. The Annals of Mathematical Statistics. 1962; 33(3):1065–1076.

Pickering AD. New approaches to the study of amnesic patients: What can a neurofunctional philosophy and neural network methods offer? Memory. 1997; 5(1–2):255–300. [PubMed: 9156101]

Poldrack RA, Clark J, Pare-Blagoev E, Shohamy D, Moyano JC, Myers C, Gluck M. Interactive memory systems in the human brain. Nature. 2001; 414(6863):546–550. [PubMed: 11734855]

Reiner, A. Organization of corticostriatal projection neuron types. In: Steiner, H., Tseng, KYKY., editors. Handbook of basal ganglia structure and function. New York: Elsevier; 2010. p. 323-339.

Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex. Nature Neuroscience. 1999; 2(11):1019–1025. [PubMed: 10526343]

Ronesi J, Lovinger DM. Induction of striatal long-term synaptic depression by moderate frequency activation of cortical afferents in rat. The Journal of Physiology. 2005; 562(1):245–256. [PubMed: 15498813]

Rosenblum M, Yacoob Y, Davis LS. Human expression recognition from motion using a radial basis function network architecture. IEEE Transactions on Neural Networks. 1996; 7(5):1121–1138. [PubMed: 18263509]

Rouder JN, Ratcliff R. Comparing exemplar-and rule-based theories of categorization. Current Directions in Psychological Science. 2006; 15(1):9–13.

Sage JR, Anagnostaras SG, Mitchell S, Bronstein JM, De Salles A, Masterman D, Knowlton BJ. Analysis of probabilistic classification learning in patients with parkinson's disease before and after pallidotomy surgery. Learning & Memory. 2003; 10(3):226–236. [PubMed: 12773587]

Sakamoto Y, Love BC. Schematic influences on category learning and recognition memory. Journal of Experimental Psychology: General. 2004; 133(4):534–553. [PubMed: 15584805]

Schultz W. Predictive reward signal of dopamine neurons. Journal of Neurophysiology. 1998; 80(1):1–27. [PubMed: 9658025]

Schultz W. Getting formal with dopamine and reward. Neuron. 2002; 26:241–263.

Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997; 275(5306):1593–1599. [PubMed: 9054347]

Seger CA. How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. Neuroscience & Biobehavioral Reviews. 2008; 32(2):265–278. [PubMed: 17919725]

Seger CA, Cincotta CM. Striatal activity in concept learning. Cognitive, Affective, & Behavioral Neuroscience. 2002; 2(2):149–161.

Seger CA, Cincotta CM. The roles of the caudate nucleus in human classification learning. The Journal of Neuroscience. 2005; 25(11):2941–2951. [PubMed: 15772354]

Seger CA, Miller EK. Category learning in the brain. Annual Review of Neuroscience. 2010; 33:203–219.

Seger CA, Peterson EJ, Cincotta CM, Lopez-Paniagua D, Anderson CW. Dissociating the contributions of independent corticostriatal systems to visual categorization learning through the use of reinforcement learning modeling and granger causality modeling. NeuroImage. 2010; 50(2): 644–656. [PubMed: 19969091]

Shen W, Flajolet M, Greengard P, Surmeier DJ. Dichotomous dopaminergic control of striatal synaptic plasticity. Science. 2008; 321(5890):848–851. [PubMed: 18687967]

Shepard RN. Toward a universal law of generalization for psychological science. Science. 1987; 237(4820):1317–1323. [PubMed: 3629243]

Shohamy D, Myers C, Onlaor S, Gluck M. Role of the basal ganglia in category learning: How do patients with parkinson's disease learn? Behavioral Neuroscience. 2004; 118(4):676–686. [PubMed: 15301595]

Spiering BJ, Ashby FG. Response processes in information–integration category learning. Neurobiology of Learning and Memory. 2008; 90(2):330–338. [PubMed: 18550397]

Squire LR, Knowlton BJ. Learning about categories in the absence of memory. Proceedings of the National Academy of Science USA. 1995; 92:12470–12474.

Valentin VV, Maddox WT, Ashby FG. A computational model of the temporal dynamics of plasticity in procedural learning: Sensitivity to feedback timing. Frontiers in Psychology. 2014; 5(643)doi: 10.3389/fpsyg.2014.00643

Waldron EM, Ashby FG. The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. Psychonomic Bulletin & Review. 2001; 8(1):168–176. [PubMed: 11340863]

Waldschmidt JG, Ashby FG. Cortical and striatal contributions to automaticity in information-integration categorization. Neuroimage. 2011; 56(3):1791–1802. [PubMed: 21316475]

Wickelgren I. Getting the brain's attention. Science. 1997; 278(5335):35–37. [PubMed: 9340756]

Willingham DB, Wells LA, Farrell JM, Stemwedel ME. Implicit motor sequence learning is represented in response locations. Memory & Cognition. 2000; 28(3):366–375. [PubMed: 10881554]

Wilson, CJ. The contribution of cortical neurons to the firing pattern of striatal spiny neurons. In: Houk, JC.Davis, JL., Beiser, DG., editors. Models of information processing in the basal ganglia. Cambridge, MA: MIT Press; 1995. p. 29-50.

Wilson HR, Cowan JD. Excitatory and inhibitory interactions in localized populations of model neurons. Biophysical Journal. 1972; 12:1–24. [PubMed: 4332108]

Wilson HR, Cowan JD. A mathematical theory of the functional dynamics of nervous tissue. Kybernetik. 1973; 13:55–80. [PubMed: 4767470]

Witt K, Nuhsman A, Deuschl G. Dissociation of habit-learning in parkinson's and cerebellar disease. Journal of Cognitive Neuroscience. 2002; 14(3):493–499. [PubMed: 11970808]

Worthy DA, Markman AB, Maddox WT. Feedback and stimulus-offset timing effects in perceptual category learning. Brain and Cognition. 2013; 81(2):283–293. [PubMed: 23313835]

Yagishita S, Hayashi-Takagi A, Ellis-Davies GC, Urakubo H, Ishii S, Kasai H. A critical time window for dopamine actions on the structural plasticity of dendritic spines. Science. 2014; 345(6204): 1616–1620. [PubMed: 25258080]

Yellott JI. The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. Journal of Mathematical Psychology. 1977; 15(2):109–144.

Zaki SR, Nosofsky RM, Jessup NM, Unversagt FW. Categorization and recognition performance of a memory-impaired group: Evidence for single-system models. Journal of the International Neuropsychological Society. 2003; 34:394–406.

Zeithamova D, Maddox WT. Dual-task interference in perceptual category learning. Memory & Cognition. 2006; 34(2):387–398. [PubMed: 16752602]

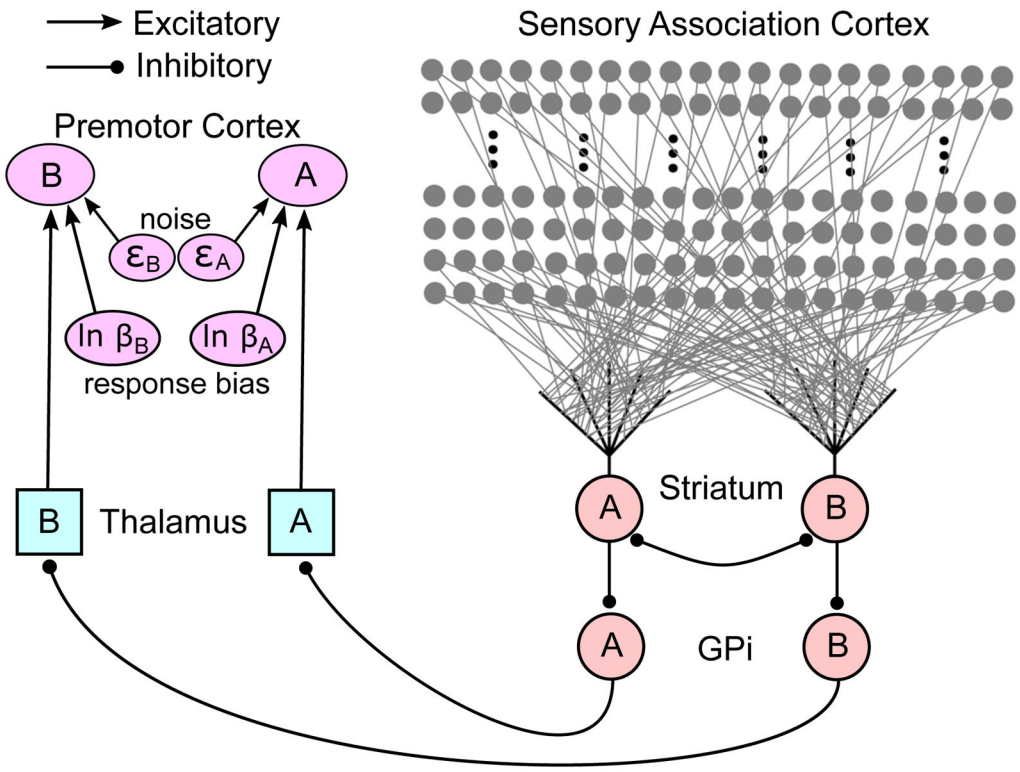

Figure 1.
Architecture of the model that is computationally equivalent to exemplar theory.