CHAPTER

22

Categorization

Bradley C. Love

Abstract

Judging a person as a friend or foe, a mushroom as edible or poisonous, or a sound as an *l* or *r* are examples of categorization problems. This chapter considers the relative merits of four basic types of category learning models: rule-, prototype-, exemplar-, and cluster-based models. The history of model progression is marked by descendant models displaying increasingly sophisticated processing mechanisms that can manifest the behaviors of ancestral models. These four basic model types are related to the computations performed by four candidate learning systems in the human brain, which rely on prefrontal cortex, posterior occipital cortex, the striatum, and the medial temporal lobes. One issue raised is whether the prefrontal cortex and posterior occipital cortex support true learning systems or are better viewed as supporting general cognitive and perceptual abilities. Use of well-specified cognitive models can help answer related theoretical questions, such as how multiple learning systems contribute to categorization behavior.

Key Words: categorization, category learning, classification, memory systems, learning systems

Introduction

The act of categorization is ubiquitous in human behavior. Judging a person as a friend or a foe, a mushroom as edible or poisonous, or a sound as an *l* or *r* are examples of categorization problems. Because people never encounter the same exact stimulus twice, they must develop categorization schemes that capture the useful regularities in their environment. One key research challenge is to determine how humans acquire and represent categories. The focus of this chapter will be on proposed category learning mechanisms and their brain basis. While there are a number of other valuable topics in categorization research, such as how semantic information is organized (Cree & McRae, 2003), the nature of category-specific deficits (Caramazza & Shelton, 1998), and how prior knowledge guides category acquisition (Rehder & Murphy, 2003), this chapter will focus on models and studies that address how people acquire novel categories from observed examples. For a review of how well-learning categories are represented in the brain, see work by Martin (2007).

Category learning is a theory- and model-rich area within cognitive psychology. Models have played a prominent role in shaping our understanding of human category learning. Accordingly, proposed mechanisms are diverse, including rule-, prototype-, and exemplar-based models, as well as clustering models and models that contain multiple systems. One general trend is toward models with increasingly sophisticated processing mechanisms that can mimic the behaviors of existing models, as well as address behaviors outside the scope of previous models.

Cognitive models are beginning to play an important role in cognitive neuroscience research as well, particularly in the area of category learning.

Cognitive models are distinguished from other useful analysis tools, such as multivoxel pattern recognition (see Pereira, Mitchell, & Botvinick, 2009), in that cognitive models are theories of the mental operations that support behavior, rather than simply analysis tools. Operations and components in cognitive models can be linked to brain measures (such as the BOLD signal in fMRI studies; see Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006) to understand the brain basis of interesting behaviors, such as the operations that support categorization (Davis, Love, & Preston, 2012). Model-based analysis can help us understand how human behavior arises from the interaction of numerous brain regions. In addition to aiding data analysis, formal cognitive models make clear predictions that can be evaluated analytically or through simulation. Successful models are formal characterizations of the field's best theories, and unlike verbal theories, formal models can be evaluated quantitatively.

Cognitive models may help overcome common century-old criticisms of cognitive neuroscience research. Franz remarked in his 1912 essay "New Phrenology" that "the individual parts of the brain do not work independently; they work interdependently, and it is because of the possible functional and anatomical connections that certain types or kinds of mental states are more in evidence than others." To Franz, the allure of localizing mental activities in the brain begot overly simplistic and crude theories of mental processes and brain function. Cognitive models may offer a solution to these difficulties (see Love & Gureckis, 2007). Localizing mental function need not be problematic. The issue is what to localize. The value of a theory that localizes mental function lies in both the characterization of the mental process and the bridge theory that links this characterization to the brain. Starting with an ill-specified or folk psychological theory of mental function ultimately limits the value of the overall enterprise and invites comparison to Franz's new phrenology.

For these reasons, this chapter places an emphasis on model mechanisms and their linkage to the brain. One claim is that well-specified, process models of cognitive functions are the appropriate targets for localization. Successful cognitive models, which are quantitatively validated on a broad range of data sets, offer a number of advantages over folk psychological, ad hoc, or traditional psychological theories. In addition to being predictive, behavioral models have mechanisms and dynamics that can be related to brain measures. For example, models of decision processes (a component process in categorization) have been useful for understanding how choice is implemented in the brain (Purcell et al., 2010). Although not naive accounts of mental function, cognitive models are typically idealized and relatively simple. This clarity provides a good starting point for localizing function. Given that debates persist over the basic function of areas as well studied as the hippocampus (Eichenbaum, 1999; Stark, Bayley, & Squire, 2002), starting simple makes sense.

In the course of reviewing a variety of category learning models, I will emphasize what the relative merits of each model reveal about the nature of human learning. After reviewing the basic model types, the relationship between models of category learning and candidate learning systems in the brain will be considered. Finally, a number of challenges for understanding the brain basis of categorization will be discussed.

Models of Category Learning

In this section, I will briefly review several models of human category learning. Presentation order is organized chronologically, from oldest to most recent accounts of category learning. Although more recent models offer some advantages over their ancestors, it would be a mistake to view ancestral models as being supplanted by their descendants. Each model class addresses some key aspects of human category learning and serves an important theoretical role. In fact, many older models have taken on new life as components in recently proposed multiple systems models. One common component in these multiple systems models is a rule-based system, which is the first model class considered here.

Rule-Based Models

The classical view of categories holds that categories are defined by logical rules. This view has a long history, dating back to Aristotle. In Figure 22.1, any item that is a square is a member of category A. This simple rule determines category membership. According to the rule view, our category of category A can be represented by this simple rule. Discovering this rule would involve a rational hypothesis-testing procedure. Through this procedure one attempts to discover a rule that is satisfied by all of the positive examples of a category, but none of the negative examples of the category (i.e., items that are members of other categories). In trying to come up with such a rule for category A, one might first try the rule *if dark, then in category A*. After rejecting this rule (because there are counterexamples), other rules would be tested (starting with simple rules and progressing toward more complex rules) until the correct rule is eventually discovered. For example, in learning about birds, one might first try the rule *if it flies, then it is a bird*. This rule works pretty well, but not perfectly (penguins do not fly and bats do). Another simple rule like *if it has feathers, then it is a bird* would not work either because a pillow filled with feathers is not a bird. Eventually, a more complex rule might be discovered, such as *if it has feathers and wings, then it is a bird*.

For decades psychologists have conducted experiments to characterize the relative difficulty people have in learning various types of rules (Bruner, Goodnow, & Austin, 1956; Shepard, Hovland, & Jenkins, 1961). These studies have provided the primary data used to develop and validate models of hypothesis testing. Some models, such as RULEX (Nosofsky, Palmeri, & McKinley, 1994), embody the hypothesis testing procedure described above. RULEX starts with simple hypotheses and progresses toward more complex hypotheses until a set of rules and exceptions is discovered that properly discriminates between the categories.

The term *rule* has various, somewhat conflicting, interpretations. Here, I focus on rule-based models, like RULEX, that engage in explicit, hypothesis testing. RULEX's mechanistic approach (i.e., algorithmic in the sense of Marr, 1982) contrasts with other approaches that aim to predict how difficult learning should be, based on calculations of how complex the correct hypothesis is (Feldman, 2000). The latter approaches, which are not concerned with the actual process of learning, have more in common with measures of complexity and compression (Pothos & Chater, 2002). Yet other approaches, such as General Recognition Theory (Maddox & Ashby, 1993), aim to assess and compactly describe people's performance rather than characterize the learning process. Unlike these more abstract approaches, mechanistic models of hypothesis testing, such as RULEX, largely implement the strategic and conscious thought processes that we feel (by introspection) that we are carrying out when solving classification problems. These explicit processes are thought to rely on limited working memory capacity (Zeithamova & Maddox, 2006).

Although rules can in principle provide a concise representation of a category, often more elaborate representations would serve us better. Category representation needs to be richer than a simple rule, because we use categories for much more than simply classifying objects we encounter. For instance, we often use categories to support inference (e.g., a child infers that members of the category stove can be dangerously hot). Using categories to make inferences is a very important use of categories (Markman & Ross, 2003). Knowing something is an example of a category tells us a great deal about the item. For example, after classifying a politician from the United States as a Republican, one can readily infer the politician's position on a number of issues. The point is that our representations of categories must include information beyond what is needed to classify items as examples of the category. For example, the rule *if square, then in category A* correctly classifies all members of category A in Figure 22.1, but it doesn't capture the knowledge that all category A members are *dark*. One problem with rule representations of categories is that potentially useful information is discarded. In fact, even when people explicitly use rules to classify item, performance is heavily influenced by rule-irrelevant information (Allen & Brooks, 1991; Lacroix, Giguere, & Larochelle, 2005; Sakamoto & Love, 2004), which is inconsistent with rules serving as the sole basis for category representations.

Perhaps the biggest problem with the rule approach to categories is that most of our everyday categories do not seem to be describable by a tractable rule. To demonstrate this point, Wittgenstein (1953) noted that the category game lacks a defining property. Most games are fun, but Russian roulette is not fun. Most games are competitive, but ring around the roses is not competitive. While most games have characteristics in common, there is not a rule that unifies them all. Rather, we can think of

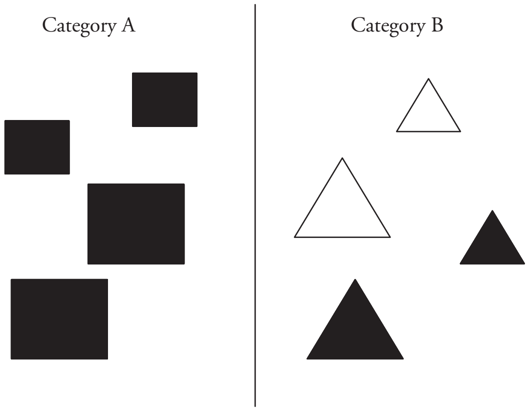

Figure 22.1 Examples of category A and category B. A simple rule on shape discriminates between the two categories.

the members of the category game as being organized around a family resemblance structure (analogous to how members of your family resemble one another). Rosch and colleagues' (Rosch & Mervis, 1975) seminal work demonstrated the psychological reality of many of Wittgenstein's intuitions. Even some paradigmatic examples of rule-based classification reveal a non-rule-based underbelly (see Love, Tomlinson, & Gureckis, 2008, for a review). Hahn and Ramscar (2001) offer one such example. Tigers are defined as having tiger DNA, which is a seemingly rule-based category definition. However, determining whether an animal has tiger DNA amounts to assessing the similarity of the animal's DNA to known examples of tiger DNA.

A related weakness of the rule account of categories is that examples of a category differ in their typicality (Barsalou, 1985; Posner & Keele, 1968; Reed, 1972; Rosch & Mervis, 1975). If all a category consisted of was a rule that determined membership, then all examples should have equal status. According to the rule account, all that should matter is whether an item satisfies the rule. Our categories do not seem to have this definitive flavor. For example, some games are better examples of the category game than others. Basketball is a very typical example of the category of games. Children play basketball in a playground, it is competitive, there are two teams, each team consists of multiple players, you score points, etc. Basketball is a typical example of the category of games because it has many characteristics in common with other games. Russian roulette, by contrast, is not a very typical game—it requires a gun and one of the two players dies. Russian roulette does not have many properties in common with other games. In terms of family resemblance structure, we can think of basketball as having a central position and Russian roulette being a distant cousin to the other family members. These findings extend to categories in which a simple classification rule exists. For example, people judge the number 3 to be a more typical odd number than the number 47, even though membership in the category odd number can be defined by a simple rule (Gleitman, Gleitman, Miller, & Ostrin, 1996).

The fact that category membership follows a gradient as opposed to being all or none affords us flexibility in how we apply our categories. Of course, this flexibility can lead to ambiguity. Consider the category mother (see Lakoff, 1987, for a thorough analysis). It is a category that we are all familiar with that seems straightforward—a mother is a woman who becomes pregnant and gives birth to a child. But what about a woman who adopts a neglected infant and raises it in a nurturing environment? Is the birth mother who neglected the infant a mother? What if a woman is implanted with an embryo from another woman? Court cases over maternity arise because the category of motherhood is ambiguous. The category exhibits greater flexibility and productivity than is even indicated above. For example, is it proper to refer to an architect as the mother of a building? All the above examples of the category mother share a family resemblance structure (i.e., they are organized around some commonalities), but the category is not rule based. Some examples of the category mother are better than others.

I do not want to imply that rule-based approaches do not have their place. For example, rule-based approaches might be viable for some socially defined categories. For example, determining whether currency is legal tender might largely involve applying a series of rules (Hampton, 2001). Also, as we will see later in this chapter, rule-based approaches figure prominently in multiple systems accounts. While rule-based approaches might not provide a sufficient explanation of human learning in isolation, such approaches might prove viable in certain domains or as components of multiple systems models.

Prototype-Based Models

The prototype approach to category learning and representation was developed by Rosch and colleagues to address some of the shortcomings of the rule approach. Prototype models represent information about all the possible properties (i.e., stimulus dimensions), instead of focusing on only a few properties like rule models do. The prototype of a category is a summary of all of its members (Posner & Keele, 1968; Reed, 1972; Smith & Minda, 2001). Mathematically, the prototype is the average or central tendency of all category members. Figure 22.2 displays the prototypes for two categories, simply named categories A and B. Notice that all the items differ in size and luminance (i.e., there are two stimulus dimensions) and that the prototype is located amidst all of its category members. The prototype for each category has the average value on both the stimulus dimensions of size and luminance for the members of its category.

The prototype of a category is used to represent the category. According to the prototype model, a novel item is classified as a member of the category whose prototype it is most similar to. For example, a large bright item would be classified as a member

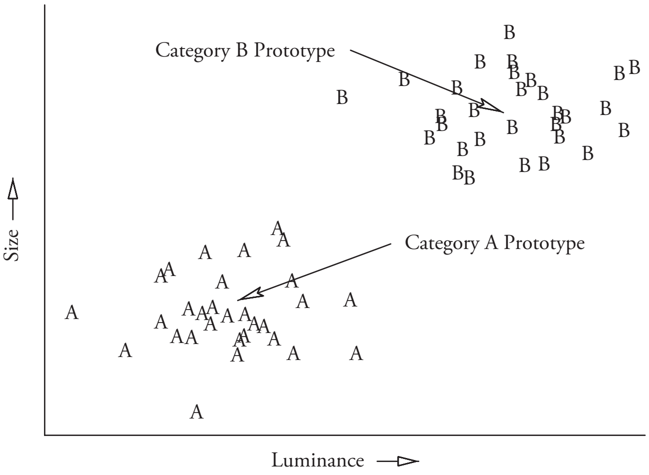

Figure 22.2 Two categories and their prototypes.

of category B because category B's prototype is large and bright (see Figure 22.2). The position of the prototype is updated when new examples of the category are encountered. For example, if one encountered a very small and dark item that is a member of category A, then category A's prototype would move slightly toward the bottom left corner in Figure 22.2. As an outcome of learning, the position of the prototype shifts toward the newest category member in order to take it into account. A prototype can be very useful for determining category membership in domains where there are many stimulus dimensions that each provide information useful for determining category membership, but no dimension is definitive. For example, members of a family may tend to be tall and have large noses, a medium complexion, brown eyes, and good muscle tone, but no family member possesses all of these traits. Matching on some subset of these traits would provide evidence for being a family member.

Notice the economy of the prototype approach. Each cloud of examples in Figure 22.2 can be represented by just the prototype. The prototype is intended to capture the critical structure in the environment without having to encode every detail or example. It is also fairly simple to determine which category a novel item belongs to by determining which category prototype is most similar to the item.

Unlike the rule approach, the prototype model can account for typicality effects. According to the prototype model, the more typical category members should be those members that are most similar to the prototype. In Figure 22.2, similarity can be viewed in geometric terms—the closer together items are in the plot, the more similar they are. Thus, the most typical items for categories A and B are those that are closest to the appropriate prototype. Accordingly, the prototype approach can explain why robins are more typical birds than penguins. The bird prototype represents the average bird: has wings, has feathers, can fly, can sing, lives in trees, lays eggs, etc. Robins share all of these properties with the prototype, whereas penguins differ in a number of ways (e.g., penguins can't fly, but they do swim). Extending this line of reasoning, the best example of a category should be the prototype, even if the actual prototype has never been viewed (or doesn't even exist). Indeed, numerous learning studies support this conjecture. After viewing a series of examples of a category, human participants are more likely to categorize the prototype as a category member (even though they never actually viewed the prototype) than they are to categorize an item they have seen before as a category member (Posner & Keele, 1968).

Because the prototype approach does not represent categories in terms of a logical rule that is either satisfied or not, it can explain how category membership has a graded structure that is not all or none. Some examples of a category are simply better examples than other examples. Also, categories do not need to be defined in terms of logical rules but are rather defined in terms of family resemblance to the prototype. In other words, members of a category need not share a common defining thread, but can have many characteristic threads in common with one another.

The prototype approach, while preferable to the rule approach for the reasons just discussed, does fail to account for important aspects of human category learning. The main problem with the prototype model is that it does not retain enough information about examples encountered in learning. For instance, prototypes do not store any information about the frequency of each category, yet people are sensitive to frequency information. If an item was about equally similar to the prototype of two different categories and one category had 100 times more members than the other, people would be more likely to assign the item to the more common category (under most circumstances, see Kruschke, 1996). Of course, some of these concerns could be addressed by expanding the information that a prototype encodes.

However, other concerns seem fundamental to the prototype approach. Prototypes are not sensitive to the correlations and substructure within a category. For example, a prototype model would not be able to represent that spoons tend to be large and made of wood or small and made of steel. These two subgroups would simply be averaged together into one prototype. This averaging makes some categories unlearnable with a prototype model. One example of such a category structure is shown in Figure 22.3. Each category consists of two subgroups. Members of category A are either *small* and *dark* or they are *large* and *light*, whereas members of category B are either *large* and *dark* or they are *small* and *light*. The prototypes for the two categories are both in the center of the stimulus space (i.e., medium size and medium luminance). Items cannot be classified correctly by which prototype they are most similar to because the prototypes provide little guidance.

In general, prototype models can only be used to learn category structures that are linearly separable. A learning problem involving two categories is linearly separable when a line or plane can be drawn that separates all the members of the two categories. The category structure shown in Figure 22.2 is linearly separable because a diagonal line can be drawn that separates the category A and B members (i.e., the category A members fall on one side of the line and the category B members fall on the other side of the line). Thus, this category structure can be learned with a prototype model. The category structure illustrated in Figure 22.3 is nonlinear—no single line can be drawn to segregate the category A and B members. Mathematically, a category structure is linearly separable when there exists a weighting of the feature dimensions that yields an additive rule that correctly indicates one category when the sum is below a chosen threshold and the other category when the sum is above the threshold.

The inability of the prototype model to learn nonlinear category structures detracts from its worth as a model of human category learning because people are not biased against learning nonlinear category structures. While the extent to which natural categories deviate from linear structures is contested (Murphy, 2002), the general consensus is that people in the laboratory do not show a preference for linear structures in supervised learning (Medin & Schwanenflugel, 1981), though they might in unsupervised learning (Love, 2002). Some nonlinear category structures may actually be easier to acquire than linear category structures. For example, it seems quite natural that small birds sing, whereas large birds do not sing. Many categories have subtypes within them that we naturally pick out. One way for the prototype model to address this learnability problem is to include complex features that represent the presence of multiple simple features (e.g., large and blue). Unfortunately, this approach quickly becomes unwieldy as the number of stimulus dimensions increases (e.g., Gluck & Bower, 1988).

Related to the prototype model's inability to account for substructure within categories is its inadequacy as a model of item recognition. Unlike exemplar models considered in the following section (Medin & Schaffer, 1978; Nosofsky, 1986), prototypes models do not readily account for how people recognize specific items because the category prototype averages away item-distinguishing information that people retain in some situations.

Exemplar-Based Models

Exemplar models store every training example in memory instead of just the prototype (i.e., the summary) of each category. Perhaps surprising upon first consideration, exemplar models can account for findings marshaled in support of prototype models, such as sensitivity to family resemblance structure. At the same time, by retaining all of the information from training, exemplar models address many of the shortcomings of prototype model. Exemplar models are sensitive to the frequency, the variability, and the correlations among items. In this section, I will discuss how exemplar-based models can display these behaviors.

Unlike prototype models, exemplar models can master category structures that contain substructure. For the learning problem illustrated in Figure 22.3,

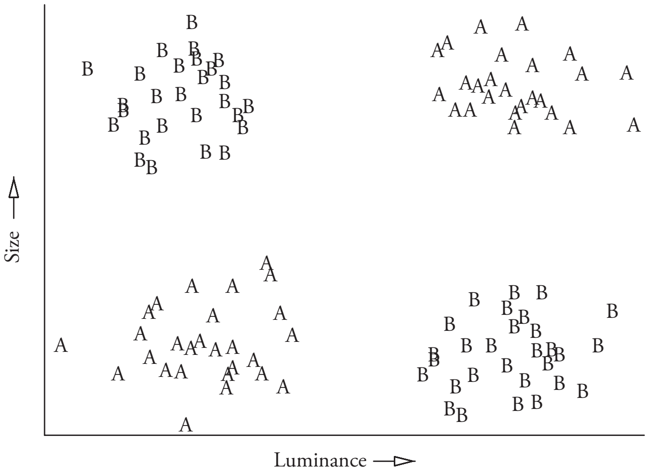

Figure 22.3 Two categories and their prototypes.

an exemplar model would store every training example. New items are classified by how similar they are to all items in memory (not just the prototype). For the category structure illustrated in Figure 22.3, the pairwise similarity of a novel item and every stored item would be calculated. If the novel item tended to be more similar to the category A members (i.e., the item was small and dark) than to the category B members, then the novel item would be classified as a member of category A.

One aspect of exemplar models that seems counterintuitive is their lack of any abstraction in category representation. It seems that humans do learn something more abstract about categories than a list of examples. Surprisingly, exemplar models are capable of displaying abstraction. For instance, exemplar models can correctly predict that humans more strongly endorse the underlying prototype (even if it has not been seen) than an actual item that has been studied (a piece of evidence previously cited in favor of the prototype model). How could this be possible without the prototype actually being stored? It would be impossible if exemplar models simply functioned by retrieving the exemplar in memory that was most similar to the current item and classified the current item in the same category as the retrieved exemplar (this is essentially how processing works in a prototype model, except that a prototype is stored in memory instead of a bunch of exemplars).

Instead, exemplar models engage in more sophisticated processing and calculate the similarity between the current item (the item that is to be classified) and every item in memory. Some exemplars in memory will be very similar to the current item, whereas others will not be very similar. The current item is classified in the category in which the sum of its similarities to all the exemplars is greatest. When a previously unseen prototype is presented to an exemplar model, it can be endorsed as a category member more strongly than a previously seen item. The prototype (which is the central tendency of the category) will tend to be somewhat similar to every item in the category, whereas any given nonprototype item will tend to be very similar to some items (especially itself!) in memory, but not so similar to other items. Overall, the prototypical item can display an advantage over an item that has actually been studied. Abstraction in an exemplar model is indirect and results from processing (i.e., calculating and summing pairwise similarities), whereas abstraction in a prototype model is rather direct (i.e., prototypes are stored).

By and large, exemplar models can mimic all the behaviors of prototype models, but the opposite is not true. There are some subtle behaviors that the prototype model can display that versions of exemplar models cannot. For example, prototype and exemplar models predict slightly different category endorsement gradients (i.e., probability of membership) as one moves toward the center of a category (see Nosofsky & Zaki, 2002; and Smith, 2002, for a recent debate).

Although exemplar models are decent models of recognition, they do have some fundamental shortcomings. Exemplar models calculate recognition strength as the sum of similarity to all items stored in memory. Thus, the pairwise similarity relations among items govern recognition. However, humans often appear to build schema-like structures in memory and store items preferentially that deviate from these structures (see Sakamoto & Love, 2004, for a review). Thus, exemplar models do not correctly predict enhanced recognition for items that violate salient rules or patterns (Palmeri & Nosofsky, 1995). Exemplar models do not capture these results because exception items that violate these patterns are not exceptional in terms of their pairwise similarity relations to other items. Exception items are exceptional in terms of violating a knowledge structure stored in memory (Sakamoto & Love, 2004, 2006).

At a more philosophical level, exemplar models seem to make some questionable assumptions. For example, exemplar models store every training example, which seems excessive. Also, every exemplar is retrieved from memory every time an item is classified (though see Nosofsky & Palmeri, 1997, for an exception). In addition to these assumptions, one worries that the exemplar model does not make strong enough theoretical commitments because it retains all information about training and contains a great deal of flexibility in how it processes information. In support of this conjecture, Sakamoto, Matsuka, and Love (2004) built an exemplar model that effectively built distributed knowledge structures and could account for exception recognition findings (also see Rodrigues & Murre, 2007). While their model did not explicitly build schema or exception representations, the model did learn to selectively tune exemplars (broad tunings for rule-following items and tight tunings for exception items) and properly weight these exemplars to give rise to an exemplar model that functionally contained exception and schema-like knowledge structures. If there are no constraints on how items are processed, then in principle an exemplar model can account for any pattern of results, thereby reducing the exemplar model's theoretical utility. However, in practice, exemplar models often follow previously published formalisms and serve as valuable theoretical tools.

Clustering Models

Prototype and exemplar models can be seen as opposite ends of a continuum of category representation. On one extreme, prototype models store every category member together in memory. At the other extreme, exemplar models store every category member separately in memory. Between these two extremes lie a wealth of possibilities. Categories in the real world contain multiple subtypes and exceptions. For example, the category mammals contains subcategories like cats, dogs, horses, and bats. Ideally, our mental representations would reflect this structure. Both prototype and exemplar models are inflexible in that they treat the structure of each category as predetermined. These models do not let the distribution of category members influence the form category representations take. For example, prototype models assume that categories are always represented by one node (i.e., the prototype) in memory, whereas exemplar models assume that categories are always represented by one node in memory for every category example encountered.

One reasonable intuition is that similar items should cluster together in memory (Anderson, 1991; Love, Medin, & Gureckis, 2004; Vampaemel & Storms, 2008). For example, a person walking down Congress Avenue in Austin in the fall will encounter thousands of seemingly identical grackles. The rationale for storing each of these birds separately in memory is unclear. At the same time, someone walking down the street probably would mentally note unusual or otherwise surprising birds.

Clustering models embody these intuitions about memory. For example, Anderson's (1991) rational model (also see Sanborn, Griffiths, & Navarro, 2006) computes the probability that an item belongs to an existing cluster (a prototype can be thought of as a cluster that encodes all category members). If this probability is sufficiently high, the cluster is updated to reflect its new member. However, if the item is more likely from a new cluster, then a new cluster is created. The overarching goal of Anderson's model is to create clusters that are maximally predictive.

Love et al.'s SUSTAIN model operates along similar lines in that it incrementally adds clusters as it learns, but its recruitment process is somewhat different from the rational model's. In the SUSTAIN model, new clusters are recruited in response to surprising events. What counts as a surprising event depends on the learner's current goals. When the learner's goals are somewhat diffuse, as in unsupervised learning, SUSTAIN's operation is very similar to that of the rational model. In such cases, items that are dissimilar from existing clusters result in a new cluster being recruited to encode the item. However, in supervised learning situations, such as in classification learning (the learner's goal is to properly name the stimulus's category), items are recruited when a surprising error results. For example, upon encountering a bat for the first time and being asked to name it, a child surprised to learn that a bat is not a bird would recruit a new cluster to capture this example. If the child activates this cluster in the future to successfully classify other bats, then the cluster would come to resemble a bat prototype.

Both the rational model and SUSTAIN can be viewed as multiple prototype models in which the number of prototypes is determined by the complexity of the category structure. When categories are very regular, these models will function like prototype models. When categories are very irregular (i.e., there is no discernable pattern linking members to one another), these models will tend to function like exemplar models. SUSTAIN's sensitivity to a learner's goal allows it to capture performance differences across different induction tasks. For example, people learning through inference (e.g., *This is a mammal. Does it have fur?*) tend to focus on the internal structure of categories, whereas people learning through classification (e.g., *This has fur. Is it a mammal?*) tend to focus on information that discriminates between categories (see Markman & Ross, 2003, for a review). These two ways of interacting with stimuli during learning have very different acquisition and retention profiles (Sakamoto & Love, 2010).

Clustering models, like exemplar and prototype models, can be coupled with selective attention mechanisms that can learn to emphasize critical stimulus properties. For example, in learning to classify car makes, SUSTAIN would learn to weight shape more than color because shape reliably indicates model type whereas color varies idiosyncratically. The motivation for selective attention comes from the observation that people can only process a limited number of stimulus properties simultaneously. Selective attention mechanisms have been developed through consideration of human and animal learning data (see Kruschke, 2003, for a review). In tasks that require people to actively sample stimulus dimensions, selective attention mechanisms predict which dimensions are fixated (Rehder & Hoffman, 2005).

Importantly, selective attention mechanisms allow non-rule models to display rule-like behaviors. When a prototype, exemplar, or clustering model places all of its attention on one stimulus dimension, a model's operation is indistinguishable from the application of a simple rule. In terms of accounting for human data, SUSTAIN outperforms RULEX in some respects on learning problems that require acquiring a simple rule and storing exceptions to these rules (Sakamoto & Love, 2004). SUSTAIN creates a small set of clusters to encode items that follow the rules and encodes exceptions in their own clusters. Attention is heavily biased to the rule-relevant dimensions. This allows SUSTAIN to show enhanced recognition for exceptions and rule-like behavior for rule-following items, while maintaining some sensitivity to non-rule-relevant dimensions like human subjects do.

The incorporation of selective attention mechanisms into non-rule models invites a number of theoretical questions. It is not entirely clear whether these selective attention mechanisms should be viewed as an integral part of non-rule models or as rule mechanisms grafted onto non-rule models. One possibility is that people are relying on rule and non-rule systems, thus necessitating the need for selective attention mechanisms in non-rule models.

Multiple Systems Models

Determining the best psychological model can be difficult, as one model may perform well in one situation but be bested by a competing model in a different situation. One possibility is that there is not a single "true" model. In category learning, this line of reasoning has led to the development of models containing multiple learning systems. These more complex models hold that category learning behavior reflects the contributions of different systems organized around discrepant principles that use qualitatively distinct representations. The idea that multiple learning systems support category learning behavior enjoys widespread support among researchers in the cognitive neuroscience of category learning (see Ashby & O'Brien, 2005, for a review and Nosofsky & Zaki, 1998, for a dissenting opinion).

Multiple systems models of category learning detail the relative contributions of the component learning systems. For each categorization decision, some multiple systems models select which individual system governs the response (Ashby, Alfonso-Reese, Turken, & Waldron, 1998). Over time, one system might prove more useful and dominate responding. Alternatively, the modeler can predetermine the timing of the shift from one system to another. This is sensible in cases where there is good evidence for predictable shifts, such as the shift from rule-based to exemplar-based responding in classification learning (Johansen & Palmeri, 2002).

Both of these multiple systems approaches are somewhat inadequate in that they do not allow the current situation to dictate which system is operable. For example, when trying to learn how to operate a new piece of machinery, a person might use a hypothesis (i.e., rule) system, but when riding a bicycle, a more procedural system might govern responding and be updated. In some models, like ATRIUM (Erickson & Kruschke, 1998), the relative contributions of divergent systems can depend on the circumstances (cf., Yang & Lewandowsky, 2004). ATRIUM contains a rule-and-exemplar learning system. The system that is operable is determined by a gating system, allowing different classification procedures to be applied to different parts of the stimulus space. For example, familiar items could be classified by the exemplar system, whereas rules could be applied to unfamiliar items.

Somewhat muddying the waters, ostensibly single-system models have been developed that also manifest this ability. In CLUSTER (Love & Jones, 2006), clusters can tune themselves (i.e., attend) to different stimulus properties and encode categories at various levels of granularity. This allows CLUSTER to apply different procedures to different parts of the stimulus space, like ATRIUM does. For example, clusters would heavily weight color in the domain of clothing and processor type in the domain of laptops. This tuning is accomplished by minimizing an error term that reflects the model's predictive accuracy, a technique commonly used in connectionist modeling. Tunable parameters that encode each cluster's specificity and attentional weighting of different properties are shaped by experience.

Models like CLUSTER are very rich. Consideration of such models leads to the question of what constitutes or defines a system. As previously discussed, one could even construe the selective attention mechanism of various models as being a separate system (see Poldrack & Foerde, 2008, for a related discussion on model parameters). Fortunately, models are mathematically well specified and allow researchers to make predictions and state their theories clearly without having to be overly concerned with the semantics of what constitutes a system. The mathematical specification of models can free researchers from some potentially thorny debates.

The notion of a system perhaps takes on greater significance when considered in the context of the brain (Ashby & Crossley, 2010). Within cognitive neuroscience, it is generally accepted that there is a hypothesis-testing system that relies on frontal circuitry (Ashby et al., 1998), a dopamine-mediated procedural learning system that involves the striatum (Ashby et al., 1998), a repetition priming system that involves early visual areas (Reber, Gitelman, Parrish, & Mesulam, 2003), and a medial temporal lobe (MTL) learning system that maps onto exemplar- or cluster-based learning (Love & Gureckis, 2007). For each system, there are behavioral manipulations that tend to emphasize the one system over the other systems. Lesion, patient, and imaging studies provide compelling evidence for the multiple systems view. The relationship between the models discussed above and proposed learning systems in the brain is discussed in greater detail in the next section.

Brain Basis of Category Learning

In this section, the relationship between the models described above and candidate learning systems in the brain is considered. Successful models that have been developed in light of these learning systems' detailed circuitry (e.g., Becker & Wojtowicz, 2007; Frank, Seeberger, & O'Reilly, 2004; Norman & O'Reilly, 2003) will not be discussed. Instead, the focus will be on linking the basic computational properties of category learning models to learning systems in the brain.

Posterior Occipital Cortex

Forms of implicit learning (i.e., learning without awareness) with visual stimuli are thought to rely on the posterior occipital cortex (see Smith & Grossman, 2008, for a review). The best support for this hypothesis comes from prototype abstraction studies in which subjects view numerous stimuli that are similar to an underlying prototype (e.g., dot pattern tasks). In these tasks, patients with impaired declarative memory, such those with lesions in the MTL (Knowlton & Squire, 1993; Kolodny, 1994; Reed, Squire, Patalano, Smith, & Jonides, 1999) and Alzheimer's disease (Bozoki, Grossman, & Smith, 2006; Eldridge, Masterman, & Knowlton, 2002), retain the ability to extract a single prototype through implicit means.

After exposure to items that coalesce around a prototype, imaging studies find deactivations of posterior occipital cortex (roughly V2) for items that are similar to the prototype (Aizenstein et al., 2000; Koenig et al., 2008; Reber, Stark, & Squire, 1998; Reber et al., 2003). High accuracy in prototype extraction tasks does not appear to require involvement of declarative memory areas, though such areas can be engaged by these learning tasks (Koenig et al., 2008).

Interestingly, this form of implicit learning seems to be very limited in terms of the types of categories that can be learned. Alzheimer's patients and amnesiacs can extract a single prototype but are unable to discriminate two prototypes (Sinha, 1999; Zaki, Nosofsky, Jessup, & Unversagt, 2003). These results suggest that the learning supported by posterior occipital cortex is better viewed as a perceptual priming system than as a general mechanism for acquiring category knowledge. One possibility is that people experience a feeling of fluency (based on deactivations in visual areas) for items similar to the average of recent items and that this feeling of fluency supports categorization performance for tasks in which there is a single prototype. Such a learning system would not be useful for discriminating categories. In terms of the models discussed, a prototype model restricted to a single prototype provides the best characterization of the reviewed findings. The other models all master a greater variety of discriminations than the posterior occipital cortex appears to support. An open issue is whether perceptual priming for prototypical stimuli leads to lasting representations or is short-lived.

Prefrontal Cortex

The prefrontal cortex (PFC) and head of the caudate nucleus are theorized to engage a rule-based category learning system that depends on working memory (WM) to support maintenance of rules and new hypothesis testing (Ashby et al., 1998; Monchi, Petrides, Petre, Worsley, & Dagher, 2001; Seger et al., 2000; Smith, Patalano, & Jonides, 1998). This learning system appears to correspond with explicit hypothesis testing in which learners are aware of applying a rule and can accurately verbally report the hypothesis they are entertaining. Manipulations that disrupt WM or executive attention are particularly detrimental to this form of rule-based learning (DeCaro, Thomas, & Beilock, 2008; Waldron & Ashby, 2001; Zeithamova & Maddox, 2006).

Patient studies indicate that explicit learning of rules does not rely on intact MTL (Janowsky, Shimamura, Kritchevsky, & Squire, 1989; Leng & Parkin, 1988). One possibility is that people solve simple rule-based tasks by entertaining rules in WM. Indeed, executive attention is mediated by structures in the PFC (Posner & Petersen, 1990). Imaging results of rule-based learning corroborate this interpretation (Konishi et al., 1999; Monchi et al., 2001; Smith et al., 1998). Here, the PFC may be better viewed as supporting rule-based reasoning during category learning tasks than as a dedicated category learning system. Like the posterior occipital cortex, it is not clear to what extent learned PFC representations persist over time (though see Asaad, Rainer, & Miller, 1998).

In terms of the models reviewed, the PFC best corresponds to a rule-based model in which there is no permanent store of inferred rules (i.e., inferred rules reside in a WM and are subject to disruption). Like the perceptual priming system, the explicit rule system is extremely limited in the types of categories that it can learn. All the aforementioned studies involve acquiring rules with one or two antecedents (e.g., "If the item is big and bright, then it is a member of Category A"). Multiple prototype, exemplar, and clustering models all provide more general and powerful learning mechanisms. That said, common laboratory and neuropsychology tasks, such as the Wisconsin Card Sorting Task, likely rely on the PFC (Joel, Weiner, & Feldon, 1997). Additionally, subtle category discriminations can involve a rule component supported by the PFC (especially during initial acquisition).

Striatum and Midbrain Dopaminergic Areas

The tail and body of the caudate nucleus are theorized to support a category learning system that involves the strengthening of associations between individual stimuli and category responses, often described as procedural learning (Ashby et al., 1998; Foerde, Knowlton, & Poldrack, 2006; Knowlton, Mangels, & Squire, 1996; Knowlton, Squire, & Gluck, 1994; Poldrack et al., 2001). Unlike the previously discussed learning systems, the procedural learning system appears able to learn arbitrary category discriminations under appropriate conditions.

Necessary conditions for learning include corrective feedback arriving shortly after responding (Shohamy, Myers, Kalanithi, & Gluck, 2008). Following Schultz, Dayan, and Montague (1997), one hypothesis is that delaying feedback disrupts dopamine-mediated learning (Maddox, Ashby, & Bohil, 2003). Likewise, manipulations that disrupt procedural learning in serial reaction time tasks (e.g., Willingham, 1998) also disrupt category learning tasks based on subtle (non-rule-based) discriminations (Ashby, Noble, Filoteo, Waldron, & Ell, 2003). Further supporting the linkage of procedural learning to dopamine-mediated striatal learning, patients with Parkinson's disease have deficits in processing feedback in procedural learning tasks (Shohamy et al., 2004). Neuroimaging studies further support this linkage (Nomura et al., 2007; Poldrack & Foerde, 2008; Shohamy et al., 2008).

In terms of the previously reviewed models, variants of exemplar models are the best computational analog to the procedural learning system. Like human procedural learning, exemplar models can master arbitrary category discriminations and are sensitive to the details of their inputs. The best matching variant is the covering map version of Kruschke's (1992) connectionist exemplar model. This model seeds the space of possible stimuli uniformly with a number of exemplar nodes and uses error-driven learning to associate stimuli with category responses. Such a model corresponds to a standard exemplar model when training examples are uniformed sampled over the space of possible items. The Striatal Pattern Classifier (SPC; Ashby & Waldron, 1999) has a similar operation, though the high-level motivation for this model is quite different. In the SPC, the mechanisms in the model are described as associating regions of stimulus space with motor responses, not as storing experienced exemplars in memory. Nevertheless, at an abstract computational level, these approaches are highly similar.

Medial Temporal Lobe

One neurobiological system that has proven difficult to characterize in terms of its role in category learning is the MTL. The essential role of the MTL for encoding and retrieval of declarative memories, long-term memory for facts and events, is well established (Scoville & Milner, 1957; Squire, 1992). However, the role of the MTL in category learning remains controversial; each of the major fixed representational forms (e.g., rules, prototypes, exemplars) has been ascribed to the function of the MTL by different groups of researchers. For example, many theories suggest that the MTL uses exemplar-based representations (Ashby & Maddox, 2005; Ashby & O'Brien, 2005; Pickering, 1997). However, empirical work has suggested that the MTL may be essential for the storage of category rules (Nomura et al., 2007; Seger & Cincotta, 2006) or representations of category prototypes (Aizenstein et al., 2000; Reber et al., 2003; Zaki et al., 2003; Zeithamova, Maddox, & Schnyer, 2008). In contrast, other therorists question whether the MTL is involved in category learning at all (Ashby et al., 1998; Maddox & Ashby, 2004). Given these difficulties in ascribing a single, fixed representational type to the function of the MTL, one plausible alternative that may integrate these disparate theories is that the MTL builds representations that are appropriate for a specific learning context, like those proposed by clustering models (e.g., Anderson, 1991; Love et al., 2004).

One hypothesis is that the SUSTAIN clustering model corresponds to the operation of MTL and its subregions (Davis et al., 2012; Love & Gureckis, 2007). In terms of declarative memory, the hippocampus is thought to play a critical role in rapidly forming conjunctive representations that bind together different sources of information into a single flexible memory (Brown & Aggleton, 2001; Eichenbaum, Yonelinas, & Ranganath, 2007; Norman & O'Reilly, 2003). Conjunctive representations are thought to be encoded by the hippocampus in response to novelty (Stark & Squire, 2001; Tulving, Markowitsch, Craik, Habib, & Houle, 1996; Yamaguchi, Hale, Desposito, & Knight, 2004) in as little as a single trial (Morris, Garrud, Rawlins, & O'Keefe, 1982; Rutishauser, Mamelak, & Schuman, 2006), as well as code information about the spatiotemporal context in which an item occurred (Staresina & Davachi, 2009; Wallenstein, Eichenbaum, & Hasselmo, 1998). SUSTAIN's clusters resemble hippocampal conjunctive representations in that they can be dynamically recruited in response to novelty on a single trial. They also bind together multiple-item features and category information into a single flexible representation that can promote generalization to novel contexts (Love et al., 2004).

Many real-world categories often appear to be describable by simple representations, such as logical rules, but upon closer inspection are found to be more complex (Wittgenstein, 1953). For example, natural categories such as birds and mammals are often associated with verbalizable rules such as, if it has wings, it is a bird, but also contain violations of these rules, such as bats. People can verbally report descriptions of bats and explicitly relate bats to other mammals, but these descriptions are not rules per se. In order for people to learn that examples as diverse as bats and ponies are all members of the category mammals, people need to build representations of the category mammals that are appropriate for this goal. The SUSTAIN model would predict that people achieve this goal by forming a separate cluster for birds and mammals, and then creating additional specialized clusters for exceptions, like bats, as they are encountered. One possibility is that the MTL acquires declarative knowledge that eclipses the limitations of rule-based models through mechanisms similar to that of the SUSTAIN model.

Conclusion

In this chapter, I have reviewed the relative merits of a variety of category learning models, including rule-, prototype-, and exemplar-based models, as well as clustering models and multiple systems models that combine two or more of these model types. Also considered was how inclusion of selective attention mechanisms can increase the capabilities of these models by endowing them with the ability to manifest rule-following behavior.

To review briefly each model family's merits, rule-based models conform to our intuition that we effortfully search for patterns that we can verbally communicate to others. In contrast to rule models, prototype models successfully reflect the graded nature of category membership. Exemplar models address deficiencies in the prototype model and can capture correlations within categories. Exemplar models also capture aspects of recognition memory performance. Clustering models successfully transition between prototype- and exemplar-like representations, depending on the complexity of the category structure.

All of these models have played a critical role in advancing the theory and design of key experiments. The development of new models is informed by the failings of preceding models. The history of model development is marked by the arrival of models with increasingly sophisticated processing mechanisms that can manifest the behaviors of previous models as well as additional human behaviors beyond the reach of previous models. Of course, the value in models lie more in predicting unanticipated behaviors than in simply accounting for known behaviors. Thus, it is important for models to be somewhat constrained to have theoretical value.

Later in the chapter, these four basic model types were related to four candidate learning systems in the brain: a PFC-supported rule-based system, a perceptual priming system that operates like a restricted prototype model, a procedural learning system that has some characteristic of exemplar models and related variants, and an MTL-supported flexible clustering model. One important question for future research is how these multiple mechanisms interact.

Some researchers may question whether it is even useful to think in terms of multiple learning systems. After all, many behavioral findings thought to indicate the need for multiple systems of representation have subsequently been shown to be consistent with a single-system interpretation (Johansen & Palmeri, 2002; Nosofsky & Johansen, 2000; Nosofsky & Zaki, 1998). At first blush, this position might seem recalcitrant, but given the mounting evidence that many brain areas perform cooperatively in learning tasks (Koenig et al., 2008; Sadeh, Shohamy, Levy, Reggev, & Maril, 2011), one could reasonably argue there is a single system at a functional level as long as it is acknowledged that certain brain areas are best suited to certain learning conditions. For example, secondary task load impairs PFC- and MTL-mediated learning but not procedural learning (Foerde et al., 2006), whereas delayed feedback impairs procedural learning but not rule-based learning (Maddox & Ing, 2005). Our recommendation is to specify model-based mechanisms and relate these mechanisms to brain function, not to argue for or against a particular number of learning systems. We believe that, in practice, the criteria for delineating separate systems is often underspecified and can lead to needless controversy. Indeed, SUSTAIN, which is a single-system model, can act as an exemplar-, prototype-, or rule-based model depending on the nature of the category learning task.

Future Directions

1. Now that many in the field are confident that several learning systems have been identified, basic questions surround how these learning systems interact during learning. Under what conditions do systems cooperate or compete? For a given situation, what determines which learning system guides behavior? Answering these questions will likely require the specifying of model gating mechanisms that determine how the outputs of systems influence behavior.
2. I suggested that two learning systems, the rule-based and perceptual priming systems, may be better viewed as general cognitive and perceptual abilities than as proper learning systems. One question for future research is how processes outside of category learning systems, such as those engaged in analogy and language use, impact categorization behavior.
3. For decades, cognitive psychologists have made theoretical progress by comparing the predictions and fits of models to behavioral data. One fruitful area for future research may be to extend this endeavor to incorporate brain imaging and neuropsychological data.

Acknowledgments

This work was supported by AFOSR grant FA9550-10-1-0268 and NIH grant MH09152.

Further Reading

Ashby, F. G., & Crossley, M. (2010). The neurobiology of categorization. In D. Mareschal, P. Quinn, & S. Lea (Eds.), *The making of human concepts* (p. 75–98). New York: Oxford University Press.

Poldrack, R. A., & Foerde, K. (2008). Category learning and memory systems debate. *Neuroscience and Biobehavioral Reviews*, *32*, 197–205.

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

Smith, E. E., & Grossman, M. (2008). Multiple systems of category learning. *Neuroscience and Biobehavioral Reviews*, *32*, 249–264.

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*, 977–987.

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General*, *120*, 3–19.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409–429.

Asaad, W., Rainer, G., & Miller, E. (1998). Neural activity in the primate prefrontal cortex during associative learning. *Neuron*, *21*, 1399–1407.

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

Ashby, F. G., & Crossley, M. (2010). The neurobiology of categorization. In D. Mareschal, P. Quinn, & S. Lea (Eds.), *The making of human concepts* (pp. 75–98). New York: Oxford University Press.

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

Ashby, F. G., Noble, S., Filoteo, J. V., Waldron, E. M., & Ell, S. W. (2003). Category learning deficits in Parkinson's disease. *Neuropsychology*, *17*, 115–124.

Ashby, F. G., & O'Brien, J. B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Sciences*, *9*, 83–89.

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

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure of categories. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *11*, 629–654.

Becker, S., & Wojtowicz, J. (2007). A model of hippocampal neurogenesis in memory and mood disorders. *Trends in Cognitive Science*, *11*, 70–76.

Bozoki, A., Grossman, M., & Smith, E. E. (2006). Can patients with Alzheimer's disease learn a category implicitly? *Neuropsycholgia*, *44*, 816–827.

Brown, M. W., & Aggleton, J. P. (2001). Recognition memory: What are the roles of the perirhinal cortex and hippocampus? *Nature Reviews Neuroscience*, *2*, 51–61.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.

Caramazza, A., & Shelton, J. R. (1998). Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of Cognitive Neuroscience*, *10*, 1–34.

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, *132*, 63201.

Davis, T., Love, B. C. & Preston, A. R. (2012). Learning the exception to the rule: Model-based fmri reveals specialized representations for surprising category members. *Cerebral Cortex*, *22* (2), 260–273.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

DeCaro, M. S., Thomas, R. D., & Beilock, S. L. (2008). Individual differences in category learning: Sometimes less working memory capacity is better than more. *Cognition*, *107*, 284–294.

Eichenbaum, H. (1999). Conscious awareness, memory, and the hippocampus. *Nature Neuroscience*, *2*, 775–776.

Eichenbaum, H., Yonelinas, A., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience*, *30*, 123–152.

Eldridge, L., Masterman, D., & Knowlton, B. (2002). Intact implicit habit learning in Alzheimer's disease. *Behavioral Neuroscience*, *116*, 722–726.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*, 107–140.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*, 630–633.

Foerde, K., Knowlton, B. J., & Poldrack, R. A. (2006). Modulation of competing memory systems by distraction. *Proceedings of the National Academy of Sciences U S A*, *103*, 11778–11783.

Frank, M., Seeberger, L., & O'Reilly R, C. (2004). By carrot or by stick: Cognitive reinforcement learning in parkinsonism. *Science*, *306*, 1940–1943.

Franz, S. I. (1912). New phrenology. *Science*, *35*, 321–328.

Gleitman, L. R., Gleitman, H., Miller, C., & Ostrin, R. (1996). Similar, and similar concepts. *Cognition*, *58*, 321–376.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*, 225–244.

Hahn, U., & Ramscar, M. (2001). Conclusion: Mere similarity? In U. Hahn & M. Ramscar (Eds.), *Similarity and categorization* (p. 257–272). New York: Oxford University Press.

Hampton, J. A. (2001). The role of similarity in natural categorization. In U. Hahn & M. Ramscar (Eds.), *Similarity and categorization* (p. 13–28). New York: Oxford University Press.

Janowsky, J. S., Shimamura, A. P., Kritchevsky, M., & Squire, L. R. (1989). The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. *Behavioral Neuroscience*, *103*, 548–560.

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

Johansen, M. K., & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive Psychology*, *45*, 482–553.

Knowlton, B., Mangels, J., & Squire, L. (1996). Neostriatal habit learning system in humans. *Science*, *273*, 1399–1402.

Knowlton, B., & Squire, L. R. (1993). The learning of categories: Parallel brain systems for item memory and category knowledge. *Science*, *262*, 1747–1749.

Knowlton, B., Squire, L., & Gluck, M. (1994). Probabilistic classification in amnesia. *Learning and Memory*, *1*, 106–120.

Koenig, P., Smith, E. E., Troiani, V., Antani, S., McCawely, G., Moore, P., et al. (2008). Medial temporal lobe involvement in an implicit memory task: Evidence of collaborating implicit and explicit memory systems from fmri and Alzheimer's disease. *Cerebral Cortex*, *18*, 2831–2843.

Kolodny, J. (1994). Memory processes in classification learning: an investigation of amnesic performance in categorization of dot patterns and artistic styles. *Psychological Science*, *5*, 164–169.

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

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

Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *22*, 3–26.

Kruschke, J. K. (2003). Attention in learning. *Current Directions in Psychological Science*, *12*, 171–175.

Lacroix, G. L., Giguere, G., & Larochelle, S. (2005). The origin of exemplar effects in rule-driven categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *31*, 272–288.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.

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

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review*, *9*, 829–835.

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

Love, B. C., & Jones, M. (2006). The emergence of multiple learning systems. *Proceedings of the Cognitive Science Society*.

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

Love, B. C., Tomlinson, M., & Gureckis, T. (2008). The concrete substrates of abstract rule use. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory*. 49, 167–207.

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, *53*, 49–70.

Maddox, W. T., & Ashby, F. G. (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioral Processes*, *66*, 309–332.

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

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, & Cognition*, *31*, 100–107.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*, 592–613.

Marr, D. (1982). *Vision*. San Francisco: W. H. Freeman.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology*, *58*, 25–45.

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

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning & Memory*, *7*, 355–368.

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

Morris, R., Garrud, P., Rawlins, J., & O'Keefe, J. (1982). Place navigation impaired in rats with hippocampal lesions. *Nature*, *297*, 681–683.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

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

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

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

Nosofsky, R. M., & Johansen, M. (2000). Exemplar-based accounts of multiple-system phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, *7*, 375–402.

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

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994, January). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79.

Nosofsky, R. M., & Zaki, S. F. (1998). Dissociations between categorization and recognition in amnesic and normal individuals. *Psychological Science*, *9*, 247–255.

Nosofsky, R. M., & Zaki, S. F. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus *generalization. Journal of Experimental Psychology: Learning, Memory, & Cognition*, *28*, 924–940.

Palmeri, T. J., & Nosofsky, R. M. (1995). Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *21*, 548–568.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *Neuroimage*, *45*, S199–S209.

Pickering, A. (1997). New approaches to the study of amnesic patients: What can a neurofunctional philosophy and neural network methods offer? *Memory*, *5*, 255–300.

Poldrack, R. A., Clark, J., Pare-Blagoev, E., Shohamy, D., Creso Moyano, J., Myers, C., et al. (2001). Interactive memory systems in the human brain. *Nature*, *414*, 546–550.

Poldrack, R. A., & Foerde, K. (2008). Category learning and memory systems debate. *Neuroscience and Biobehavioral Reviews*, *32*, 197–205.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 241–248.

Posner, M. I., & Petersen, S. E. (1990). Attention systems in the human brain. *Annual Review of Neuroscience*, *13*, 25–42.

Pothos, E. M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science*, *26*, 303–343.

Purcell, B., Heitz, R., Cohen, J., Schall, J., Logan, G., & Palmeri, T. (2010). Neurally constrained modeling of perceptual decision making. *Psychological Review*, *117*, 1113–1143.

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

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

Reed, J., Squire, L., Patalano, A., Smith, E., & Jonides, J. (1999). Learning about categories that are defined by object-like stimuli despite impaired declarative memory. *Behavioral Neuroscience*, *113*, 411–419.

Reed, S. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 382–407.

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*, 1–41.

Rehder, B., & Murphy, G. L. (2003). A knowledge-resonance (KRES) model of category learning. *Psychonomic Bulletin & Review*, *10*, 759–784.

Rodrigues, P. M., & Murre, J. M. J. (2007). Rules-plus-exception tasks: A problem for exemplar models? *Psychonomic Bulletin & Review*, *14*, 640–646.

Rosch, E., & Mervis, C. B. (1975). Family resemblences: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605.

Rutishauser, U., Mamelak, A., & Schuman, E. (2006). Single-trial learning of novel stimuli by individual neurons of the human hippocampus-amygdala complex. *Neuron*, *49*, 805–813.

Sadeh, T., Shohamy, D., Levy, D., Reggev, N., & Maril, A. (2011). Cooperation between the hippocampus and the striatum during episodic encoding. *Journal of Cognitive Neuroscience*, *23* , 1597–1608.

Sakamoto, Y., & Love, B. C. (2004). Schematic influences on category learning and recognition memory. *Journal of Experimental Psychology: General*, *33*, 534–553.

Sakamoto, Y., & Love, B. C. (2006). Vancouver, toronto, montreal, austin: Enhanced oddball memory through differentiation, not isolation. *Psychonomic Bulletin & Review*, *13*, 474–479.

Sakamoto, Y., & Love, B. C. (2010). Learning and retention through predictive inference and classification. *Journal of Experimental Psychology: Applied*, *16*, 361–377.

Sakamoto, Y., Matuska, T., & Love, B. C. (2004). Dimension-wide vs. exemplar-specific attention in category learning and recognition. In M. Lovett, C. Schunn, C. Lebiere, & P. Munro (Eds.), *Proceedings of the international conference of cognitive modeling* (Vol. 27, p. 261–266). Mahwah, NJ: Lawrence Erlbaum.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*, 1144–1167.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Scoville, W., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, & Psychiatry*, *20*, 11–21.

Seger, C. A., & Cincotta, C. (2006). Dynamics of frontal, striatal, and hippocampal systems during rule learning. *Cerebral Cortex*, *16*, 1546–1555.

Seger, C. A., Poldrack, R. A., Prabhakaran, V., Zhao, M., Glover, G., & Gabrieli, J. (2000). Hemispheric asymmetries and individual differences in visual concept learning as measured by functional mri. *Neuropsychologia*, *38*, 1316–1324.

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75*(13, Whole No. 517).

Shohamy, D. S., Myers, C. E., Grossman, S., Sage, J., Gluck, M. A., & Poldrack, R. A. (2004). Cortico-striatal contributions to feedback-based learning: converging data from neuroimaging and neuropsychology. *Brain*, *127*, 851–859.

Shohamy, D. S., Myers, C., Kalanithi, J., & Gluck, M. (2008). Basal ganglia and dopamine contributions to probabilistic category learning. *Biobehavioral Reviews.*, *32*, 219–236.

Sinha, R. R. (1999). Neuropsychological substrates of category learning. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 60 (5-B)*, *2381*, UMI No. AEH9932480.

Smith, E. E., & Grossman, M. (2008). Multiple systems of category learning. *Neuroscience and Biobehavioral Reviews*, *32*, 249–264.

Smith, E. E., Patalano, A. L., & Jonides, J. (1998). Alternative strategies of categorization. *Cognition*, *65*, 167–196.

Smith, J. (2002). Exemplar theory's predicted typicality gradient can be tested and disconfirmed. *Psychological Science*, *13*, 437–442.

Smith, J. D., & Minda, J. P. (2001). Journey to the center of the category: The dissociation in amnesia between categorization and recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *27*, 984–1002.

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, *99*, 195–231.

Staresina, B.P., & Davachi, L. (2009). Mind the gap: Binding experience across space and time in the human hippocampus. *Neuron*, *63*, 267–276.

Stark, C. E. L., Bayley, P. J., & Squire, L. R. (2002). Recognition memory for single items and for associations is similarity impaired following damage to the hippocampal region. *Learning & Cognition*, *9*, 238–242.

Stark, C. E. L., & Squire, L. R. (2001). When zero is not zero: The problem of ambiguous baseline conditions in fMRI. *Proceedings of the National Academy of Sciences U S A*, *98*(22), 12760–12766.

Tulving, E., Markowitsch, H., Craik, F., Habib, R., & Houle, S. (1996). Novelty and familiarity activations in PET studies of memory encoding and retrieval. *Cerebral Cortex*, *6*, 71–79.

Vampaemel, W., & Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review*, *15*, 732–749.

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*, 168–176.

Wallenstein, G. V., Eichenbaum, H., & Hasselmo, M. E. (1998). The hippocampus as an associator of discontiguous events. *Trends in Neuroscience*, *21*(8), 317–323.

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

Wittgenstein, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, trans.). Oxford, England: Blackwell.

Yamaguchi, S., Hale, L., Desposito, M., & Knight, R. (2004). Rapid prefrontal-hippocampal habituation to novel events. *Journal of Neuroscience*, *24*, 5356–5363.

Yang, L. X., & Lewandowsky, S. (2004). Context-gated knowledge partitioning in categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *30*, 1045–1064.

Zaki, S. R., Nosofsky, R. M., Jessup, N. M., & Unversagt, F. W. (2003). Categorization and recognition performance of a memory-impaired group: Evidence for single-system models. *Journal of the International Neuropsychological Society*, *9*, 394–406.

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

Zeithamova, D., Maddox, W. T., & Schnyer, D. M. (2008). Dissociable prototype learning systems: Evidence from brain imaging and behavior. *Journal of Neuroscience*, *28*(49), 13194–13201.