Chapter 4

Concepts: Structure and Acquisition

KIMERY R. LEVERING & KENNETH J. KURTZ

Marist College & Binghamton University

A good way to begin thinking about the psychology of concepts and categories is by making some connections to other familiar and foundational elements of human cognition. Perception provides organized sensory impressions about the physical world. Memory contains a record of experience and a storehouse of what we know about the world. Reasoning is the process of going beyond available information to generate inferences or conclusions. How do concepts and categories fit in? One can convincingly argue that they tie these elements of our cognitive system together.

Perhaps the most fundamental and universal cognitive task is matching our perceptions of the environment around us with our knowledge in memory about the kinds of things that exist and the kinds of meaning that characterize scenes and situations. This knowledge is our set of **concepts**—the tools of thought or mental representations we apply to identify and understand a stimulus. From a memory perspective, it would take a lot of effort and capacity to remember (and treat as distinct) each of the seemingly infinite number of objects, people, places, and ideas in our environment. Instead, our cognitive system has the remarkable ability to organize our experiences in long-term memory, grouping instances together into one common concept despite the many ways they might differ. Every apple you encounter is a little different, but the commonalities shared across the **category** cognitively outweigh their differences enough to warrant grouping them together into a concept of *apple*.

As a result of classifying something we have never encountered before (e.g., recognizing an item on display in a grocery store as an *apple*), we do not need to figure out everything about it from scratch. We can assume that our category knowledge applies to this instance and a number of important consequences follow. We can access other knowledge that is connected to the category (e.g. trees, serpents, gravity, teachers, pies, etc.), we can communicate to others about it (e.g., "Hey, pass me that apple!"), we can reason about and predict characteristics that may not otherwise have been obvious (e.g., it tastes sweet and offers nutrients), and we can use the categorization toward further explanation (e.g., someone who orders an apple instead of fries is trying to be healthy). As Murphy (2002) wrote, concepts are "the glue that holds our mental world together" because of their role in virtually every cognitive experience we have.

Philosophers and other theorists have long reasoned about how people learn, represent, and use concepts, but in the latter half of the 20th century, psychologists began to collect empirical data from carefully controlled laboratory experiments to test theories grounded in the information-processing framework. As in other areas of the field, research has blossomed through the application of interdisciplinary approaches such as computational modeling. In this chapter, we will review theories, models, and behavioral data that have helped us to understand how concepts are acquired and structured.

https://doi.org/10.17885/heiup.470.c6667

4.1 How Concepts Arise from Experience

It is understood that we do not come into the world as infants knowing what concepts like *fork* or *athlete* are. The rich knowledge we achieve about natural concepts comes about at least in part from experiencing examples and organizing them into groups (either on our own or based on what we are told). But what is the organizing basis that causes individuals or cultures to divide up the world as we do? What gives concepts their naturalness, their coherence, and their usefulness?

Most work in the field is consistent with the broad assumption that concepts emerge because the members of a category are like each other and different from other kinds of things. On this view, categories arise because there are regularities and a natural order in the world that can be discovered. It does not take any special work to invent categories—for example, apples are intrinsically like one another and unlike non-apples. The physical properties of objects as experienced through our senses are the grounding basis for categories. This idea of featural **similarity** has been defined in a number of ways, but it often refers to how many properties or features are shared (e.g., Tversky, 1977). For example, you would probably say that a dog is more similar to a wolf than a peacock in part because a dog and a wolf both typically have four legs, paws, fur, etc. while a dog and a peacock share far fewer characteristics. Another foundational approach to similarity is based on the geometric distance between items represented as points in a multidimensional psychological space (Shepard, 1957, 1987). To understand this, consider a cube where each interior point represents a value along each of three spatial dimensions (length, width, and depth). Shepard proposed that examples are represented as points in a multidimensional space corresponding to their values on the set of psychological dimensions along which examples vary (for example, apples may be defined in terms of roundness, redness, crunchiness, size, etc.).

When we experience a set of examples that are importantly alike (or when we are directly told that they belong to the same category), this experience invites a process of building up a general-level understanding that holds across these examples and supports generalization to new cases. This basis for category membership can be a set of features or dimension values that an item must be similar to—or it can be a rule that specifies exactly what features or dimension values are required for membership. There have been various attempts to describe how concepts arise from experience, and evaluating the relative merits of these theories has made up a considerable amount of the work in human category learning.

4.1.1 Concepts as Abstractions from the Data

Many theories of categorization assume that as you encounter examples from a category, you engage in a process of **abstraction**. This means that some detail about an example or collection of examples is lost and only the most important parts make up your concept. To understand abstraction, imagine being asked to draw a picture of your bedroom. Rather than a precise replica of the room, your picture would likely be simpler and contain fewer details. The exact number of dresser drawers, the color of your bedspread, and maybe even the presence of certain items might not be included in your drawing because you have either forgotten those details or don't consider them to be important. This is a gist-like representation of a single instance. To form concepts, the gist is formulated across many examples (other people's bedrooms) or at increased levels of abstraction (different types of rooms, interiors, physical environments, etc.). There are a number of ways that categories can be formed as abstractions, depending on the specific basis for what information to keep or discard.

4.1.1.1 Abstracting Defining Features—Classical View

The first possibility considered was that concepts are formed by abstracting a fundamentally important characteristic or set of characteristics that all examples of a category have in common. For example, you may learn over time that to be a *grandmother*, someone must (1) be female and (2) have grandchildren. As long as someone meets those necessary (they must have these qualities) and sufficient (having just these qualities is enough) conditions for membership, they are a *grandmother*. Because all that is needed is satisfying some criteria, examples are either members of the category or not, and no example is any better or worse than any other. Acquiring a concept then is a process of gradually learning the essential properties that something needs to have in order to be considered a member.

This account of essential or defining properties has been around so long and was so popular in philosophy that it is often called the **classical view** (Smith & Medin, 1981). It wasn't until the mid-20th century that philosophers and psychologists began to take issue with some of its assumptions. First, it was argued that there are no perfect definitions for categories. Wittgenstein (1953) famously argued that the concept "game" cannot be defined by any set of necessary and sufficient properties. He defended against a number of possible attempts to do so (e.g., must a game involve competition? must a game involve winning/losing?) You may expect these kinds of definitions to be easier for taxonomic categories like animal species or chemical compounds, but it has been exceedingly difficult to come up with hard and fast definitions even for these types of categories. If a necessary characteristic of a dog is that it has four legs, does an animal stop being a dog if one of its legs is amputated? Objects not fitting a definition can also sometimes be considered members of a category. For example, Lupyan (2013) found that people were willing to call someone a "grandmother" even if they had no grandchildren. Second, there are many examples that do not seem to fit cleanly into one category or another. Medin (1989) gives an example of rugs, which could be considered members of the category *furniture*, but do not seem to quite belong. Third, we see evidence of **graded structure**, meaning that some examples of a category are seen as better examples of that category than others. If you were asked to rate a list of fruit in terms of how typical they were of the category *fruit*, you would probably rate a banana as more typical than an avocado. This has been found consistently, even for categories thought to be the most well-defined. For example, Armstrong, Gleitman, and Gleitman (1983) found that certain examples of the category *even numbers* (e.g., 4) were considered to be better examples than others (e.g., 34). Such typicality effects are not easily explained by a theory that assumes examples to be simply in a category or not.

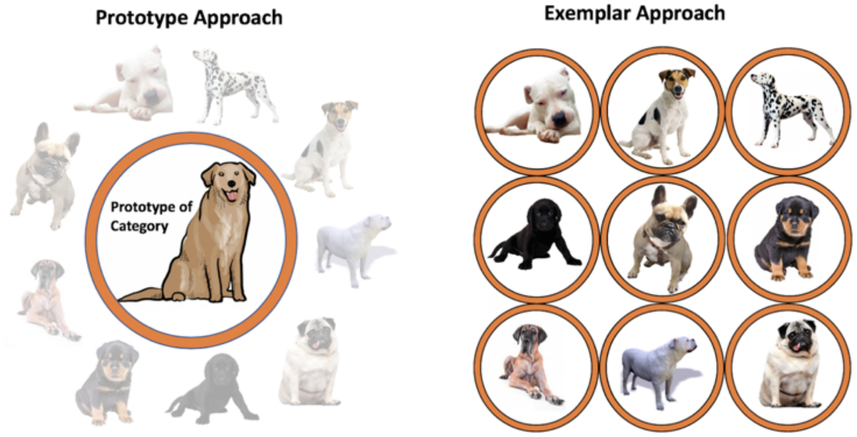

Figure 4.1: Difference between prototype and exemplar approach to the concept of *dog* arising from experiencing nine different dogs. Exemplar theory assumes the concept to be the collection of memories of each instance while prototype theory assumes the concept to be an abstracted example representing an average on relevant features.

4.1.1.2 Abstracting a Set of Common Features—Prototype Approach

In response to criticism of the classical view, a theory arose in philosophy (Wittgenstein, 1953) and later in psychology (Posner & Keel, 1968; Hampton, 1993; Smith & Minda, 2001; Rosch & Mervis, 1975) that while we do abstract the most common or central properties among category members, none of these properties are necessary or sufficient. In this set of views, eventually called the **prototype approach**, an item can be missing some features and still be considered a member of the category. Proponents of this view often think of concepts as boiling down to a single example, a *prototype*, that has the most common characteristics (e.g., has four legs) or the most common values along relevant dimensions (e.g., is 2.5 feet long). In Figure 4.1, the prototype is the average of the nine dogs experienced, even though that average is not exactly like any one of the dogs previously seen. In this view, we develop prototypes for every concept and then a new instance is classified based on which category's prototype it is more similar to. This view is often thought to better describe natural categories as members often share most but not all features, a property called *family resemblance*. This view is also considered more successful at explaining experimental findings such as unclear category membership (rugs just don't have many of the common features of furniture and are far from the category prototype) and typicality effects (items rated as less typical tend to possess fewer common features).

4.1.1.3 Abstracting a Boundary

Rather than developing a conceptual representation that is the center or average of a set of category members, other researchers have proposed that we instead update information about the boundaries of a category (Ashby, 1992; Ashby & Maddox, 1993). If the goal of concepts is to differentiate between types of things, perhaps the most important consideration is the partition line—where one category ends and another begins. For example, rather than seeing how similar a new banana is to your prototypes for the concepts *ripe banana* and *unripe banana*, we may simply use information about the point at which a banana goes from being classified as unripe to ripe along one or more dimensions. Knowledge of these partitions can identify examples of a concept without having to know anything specific about other examples or common/average features.

4.1.2 Concepts as just the Data—Exemplar Approach

More recently, a set of theories has centered on the idea that we do not form abstractions at all but rather store specific information about examples themselves (see Figure 4.1). In other words, your concept of *apple* is made up of some version of a memory of every apple you have encountered (or at least the first or most prominent ones). New apples are recognized because they are highly similar to examples that have been thought of as apples before. In fact, the most successful explanations rely on the assumption that only the examples *most* similar to the new apple have influence on classification.

This **exemplar approach** (Medin & Schaffer, 1978; Nosofsky, 1984, 1986; Kruschke, 1992) can explain prototype effects related to typicality and fuzzy boundaries because examples that are dissimilar to prototypes are also frequently dissimilar to other examples in the category. *Rug* and *ostrich* would be considered poor examples of their respective categories because they are not highly similar to any other piece of furniture or bird. Formal versions of exemplar theory have been highly successful at predicting human performance, particularly in cases where there are not many examples to learn. These draw upon two main design principles. The first is that category representations are labeled exemplars that serve as reference points for similarity comparisons. When a new example is experienced, the model figures out how similar it is to the known examples it has stored, and bases classification on the category associated with the closest match. The second has to do with how similarity is computed. In the process of looking for particularly close matches, some dimensions may be treated as more important than others, a property known as dimensional **selective attention**. If we learn that size is useful when distinguishing between types of dogs, this feature should be given more influence than something less useful like number of legs. Selective attention is typically thought to happen during encoding (meaning the number of legs a dog has does not even register) but could also be applied at the point of making a decision (the number of legs registers but does not contribute to the decision of what type of dog it is). There is plenty of experimental evidence suggesting that we use selective attention when we are learning categories, although this tendency does not seem to be as central to categorization in infants and young children.

4.1.3 Piecing Together Concepts

Much research in the last 50 years has been directed at evaluating whether concepts should be thought of as rules, prototypes, or a collection of exemplars, and evidence has been found in support of each account to differing degrees. Given that learning appears to vary in important ways across people, situations, and content, the category learning system could involve multiple processes or systems that invoke different underlying mechanisms. In line with this, several hybrid models have been developed, each asserting that information from separate systems is either combined, competes, or that a second system takes over when a primary system fails. One class of hybrid models assumes that concepts are acquired through a combination of learning rules for membership and storing individual examples (Erickson & Kruschke, 1998; Nosofsky, Palmeri, & McKinley, 1994). An approach that emphasizes separate neurobiological systems makes a strong distinction between an explicit verbal rule induction system and an implicit, procedural system (Ashby & Maddox, 2005). Similarity-based models have been developed that allow for both abstraction and exemplar-like effects by letting the model determine on the fly whether to represent the category with many clusters (a unique cluster for each item would be the exemplar approach), with one cluster (prototype view), or with an intermediate number of clusters (having a set of sub-prototypes to capture different aspects of the category; Love, Medin, & Gureckis, 2004). Another highly flexible approach is based on learning what configurations of feature values are consistent with each category—this involves no explicit use of rules or reference to specific exemplars or prototypes (Kurtz, 2007, 2015; see Hot Topic).

4.1.4 Explaining the Data

The approaches we have considered up to this point take the data about categories (i.e., the members of a category) as the direct basis for psychological representations of categories. This is most clearly evident in the exemplar view: the representation of a category consists strictly of the stored examples known to belong to the category. Abstractive accounts are based on finding a summary representation that captures the character of the category members without having to store them all. A rule is a representation that only requires storing the features that are necessary and sufficient for determining category membership. Instead of storing every example, the learner stores the information that must be true of each category member. A prototype is a statistical rather than logical form of summarization—instead of trying to summarize what is true of each example, the idea is to keep track of the central tendency among the examples. In this way, the nature of the category is captured by the set of feature values that are most representative of its members (i.e., storing a single canonical example –that could be real or made-up—instead of storing them all).

Are there alternatives to category representations that use the examples or summaries of the examples as building blocks? Why might such alternatives be important? One important consideration is that the present approach assumes that the available data (the representations of each example) contains everything we expect our categories to contain. If that is so, where do these item representations that are as semantically rich as our concepts come from? For example, if our concept of apple is merely a representation of physical features, how can that explain other information about apples like their role in appreciating teachers, avoiding doctors, discovering gravity, worms, cider, pesticides, bobbing, pies, etc. This issue becomes more extreme when considering categories that are even slightly more abstract (e.g., bag) where what makes examples similar is a construction rather than something directly derived from physical form. A promising proposal that has received only limited attention distinguishes between a *core* and an *identification* procedure for concepts (Miller & Johnson-Laird, 1976; Smith & Medin, 1981). The identification component is perceptually driven, while the core of the concept includes richly constructed semantic elements that arise from world knowledge and the interaction between humans and their environment.

Also in line with criticisms of similarity- or data-driven approaches is a **theory-driven approach** which considers categorization to be a process of explanation rather than similarity-based matching (Murphy & Medin, 1985). In this view, category representations are grounded in knowledge about what makes something a member that is not expressed in the same terms as item representation. In other words, a stimulus is not a chair because it has features that closely resemble the features of known chairs (or a summary of the features of known chairs); instead, the stimulus is a chair because the data (our sensory experience) is best explained in terms of the explanatory principles underlying chairs. What might such principles be? Researchers have looked to function and origin for such principles: Does it do what a chair should do? Was it built to be a chair? Is it used as a chair?

The classic example from Murphy and Medin (1985) asks how we categorize a fully clothed man in a pool. The suggestion is that we explain the available data in terms of the category of drunkenness by recognizing how explanatory principles like reduced coordination/judgment accord with what we see—it is not that we identify a close feature-by-feature resemblance between the man in the water and our prior experience of drunk people. The theory view of categorization provides an important critique of standard accounts: matching between stimuli and category representations requires solving the problem of identifying the "respects" for similarity—what are the features to compare upon and with what weights or importances?

In practice, researchers have had little success in translating this viewpoint into a mechanistic account of the processes and representations underlying categorization ability. Even so, much progress in the field can be seen as offshoots off the influence of the theory view. For example, an important idea rising in the field takes the perspective that categories are best represented as models of the statistical regularities that hold among category members; and the models are applied to categorize examples through a process of fitting the data rather than matching it (see Hot Topic). This resonates with a view that categories may be best understood in terms of schema theory as organized generic knowledge structures that can be activated and instantiated by filling slots with specific values (see Komatsu, 1992; Rumelhart, 1980). Another approach emphasizes the role of causal relationships in category learning and representation, for example the presence of wings on a bird and the bird's ability to fly (cf., Ahn & Kim, 2000; Rehder, 2003).

Murphy and colleagues have extended the impact of the theory view in a number of ways including a critique of the way category learning is typically studied in the laboratory that reinforces limited psychological accounts by excluding the critical role of prior knowledge about features, concepts, and general semantic memory (e.g., Murphy & Allopenna, 1994; Murphy, 2003; Wisniewski & Medin, 1994). Researchers have also been influenced by the theory view in expanding the problem of categorization beyond the ability to classify traditional taxonomic categories. There is a diversity of kinds of categories and a diversity of ways in which categories are learned and used (Markman & Ross, 2003; Medin, Lynch, & Solomon, 2000; Kurtz, 2015).

4.2 Modes of Category Learning

While the study of human category learning is ultimately about real-life concepts like *athletes* or *forks*, it is often difficult to answer questions about how *natural categories* like these are acquired because they have already been learned in unique and personal ways that cannot be easily controlled for. In order to get around this, cognitive psychologists create and teach artificial categories that can be more precisely controlled. These artificial categories are made up of members that participants have never seen before but that possess simpler versions of the kinds of features that exist in the real world. Examples are grouped into categories by researchers, often according to the same kinds of principles that we think real categories are grouped by. Participants are then taught which category each example belongs to, imitating the process by which we learn about categories in the real world. What people learn about the categories can be assessed by having them decide what category some new item is in or by asking them questions about trained examples (*How typical is this example of its category?*), features (*What category is a winged creature most likely to be in?*), or relationships between features (*How likely are winged creatures to have webbed feet?*). Specific aspects of the task (the stimuli, which examples are in which category, how many categories, etc.) can be manipulated to see in what way those changes affect how easily categories are learned, what kind of information is remembered, or how that knowledge is applied.

4.2.1 Learning Concepts Through Classification

Most commonly, concept learning is studied through a **supervised category learning** (see Figure 4.2), in which images are presented one at a time and learners decide which of usually two categories each belongs to. They are told whether they are right or wrong (this feedback is what makes the learning considered supervised) and over time they learn to correctly assign examples to the appropriate category, often with high accuracy. More than just memorizing what category each example is in, learners can pick up on relevant commonalities and differences between the categories, just like how we learn about what tends to be true of dogs and what distinguishes dogs from coyotes.

It is not hard to come up with real life instances that align with this kind of learning. For example, imagine you see an animal running across your lawn and think that it is a coyote before your friend informs you that it is in fact your neighbor's dog, Fluffy. Although we can think of cases fitting this kind of guess-and-correct classification, it is not likely the only or even primary way we learn. Concepts are most likely acquired through a combination of many modes of learning, in service of particular goals. What makes up your concept of dog likely comes from times in which you knew something was a dog before you saw it (e.g., your friend invites you over to meet her new dog), made inferences about a dog that ended up being true or not (e.g., you learn whether or not a dog will play catch), or learned about dogs incidentally while focusing on a specific task (e.g., picking out a pet from a pet store). Sometimes you may not even get feedback about whether your idea of category membership or predicted features are correct (e.g., you never find out whether the animal that ran across the lawn was a coyote or a dog).

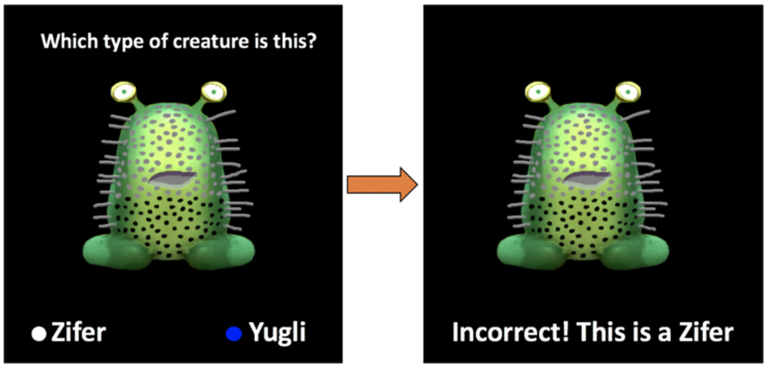

Figure 4.2: Example of one trial of a supervised classification task. The participant views an example and decides which of two categories it is in before receiving feedback.

4.2.2 Learning Concepts Through Inference, Use, and Observation

Research has provided evidence that differences in the way a concept is learned are important. Oftentimes, when a learning task is changed, different kinds of information are acquired. For example, when participants learn by predicting features of labeled examples, they often learn more about the most common features and the relationships between features (Markman & Ross, 2003; Yamauchi & Markman, 1998). The fact that certain features "go together", or are more typical or central, are aspects of the internal structure of categories. Knowledge of internal structure gives us a sense of what is generally true of a category, sometimes above and beyond what is necessary to figure out what something is. For example, the fact that silverware is typically made of metal may be useful to learn even if it does not help you determine if something is a spoon or a fork. In addition to *inference learning*, internal structure is also better learned through *indirect learning tasks* where organization into categories helps to accomplish some goal like predicting how much food animals would eat but categories are not explicitly learned (Minda & Ross, 2004). It is also better learn in *observational tasks* where category labels are provided before the example is shown, and guessing is not necessary (Levering & Kurtz, 2015). In essence, task demands during learning influence what is attended to and what becomes more central to the representation of a category. When a task focuses the learner on classification, the learner focuses on the information that is necessary for classification but when that focus is removed, more robust knowledge of internal structure can be acquired. Because categories in the real world are used for a multitude of different tasks, developing robust categories through multiple modes of learning is essential.

4.2.3 Organizing our Own Concepts

In many cases, we cannot rely on category membership being explicitly defined for us but rather we must organize our observations into categories using our own heuristics. For example, your concept of music genres (e.g., *classical music* or *hip hop*) has probably not come from listening to carefully labeled songs and learning the features associated with each genre. While some experiences may have been labeled for you (e.g., you hear a song while listening to a country radio station), you have largely constructed your own organization based on unlabeled examples. Research into purely *unsupervised* classification is often difficult because there are so many ways that a number of items can be organized. One common finding emerging from this research is that when asked to sort items into categories, people tend to focus on forming rules along single dimensions (e.g., Medin, Wattenmaker, & Hampson, 1987). For example, you may decide that any song being sung with a southern twang is country music and not need consider any other dimension.

Rather than completely unsupervised, our learning is often semi-supervised, meaning that we experience a combination of labeled and unlabeled examples. Studies on the role of unlabeled examples (relative to completely supervised learning) has been mixed, sometimes showing that they are helpful, sometimes hurtful, and sometimes having no effect. Recent research has suggested that labeled cases are important when categories are highly similar and therefore category membership is ambiguous. For example, it would be useful to have some labeled cases when distinguishing subtle differences in types of electronic music, but not when learning the broad difference between classical and punk music (Vong, Navarro, & Perfors, 2016).

Even when learning about a concept is supervised, it is sometimes possible for us to decide which examples we want to learn about and when. For example, on a trip to the zoo, a child may ask a parent to label certain unknown examples ("antelope?") but not others. This self-directed learning (also known as active or selective learning) is thought to be more effective than passive (receptive) learning, particularly when category distinctions are based on simple rules (Bruner, 1961, Markant & Gureckis, 2014). Differences in how people learn in these modes can be simulated in the lab by having one group of participants construct or select specific examples to learn about while another group is either given a random presentation order or a presentation order that matches a participant in the first group (this is called a yoked design). In these kind of studies, the participant who made the selection often learns the categories better despite being exposed to the exact same examples as their yoked counterparts (Schwartz, 1966). Possible reasons for this could be that self-directed learning is more engaging, results in deeper processing and better memory for examples, or allows for more focused attention oriented toward testing specific hypotheses about category membership (see Gureckis & Markant, 2012, for more information).

4.3 Kinds of Categories and Their Uses

An important early contribution in the empirical investigation of category structure was the finding that categories are organized at different hierarchical levels that serve different purposes—and specifically that an intermediate level, known as the **basic level of categorization**, appears to play a foremost role in guiding the way we access and use categories (Rosch & Mervis, 1975). Very specific categories (waterbuck antelope) capture tightly knit knowledge reflecting a large overlap in the features that each member has. This means that a great deal can be inferred with high confidence about a member of such a category. Very broad categories (mammal) are based on only a few core common properties that carry a great deal of weight in organizing knowledge, but do not provide much specific information about their members. The basic level (antelope) provides a compromise of reasonably high resemblance between members of a single category and low resemblance between members in different categories. Therefore, the basic level of categorization may be our most fluid and task-general way of making sense of everyday experience. Interestingly, the level of categorization that is privileged may not always be the basic level—instead it varies depending on factors including age, domain expertise, cultural norms, and the goals or tasks for which the category is being used (see Medin & Atran, 2004; Tanaka & Taylor, 1991).

As discussed above, the theory view suggests that concepts may not be sufficiently grounded by physical similarities (see Goldstone, 1994). This may or may not apply to ordinary entity concepts like dog and chair, but it has become clear that there are important kinds of categories that are certainly not subject to traditional similarity (high levels of match between features) as an organizing principle.

Barsalou (1983, 1985) demonstrated the existence and psychological role of *ad-hoc* categories that are generated in the moment (i.e., things to take out of a house in case of fire) as well as more stable categories that are *goal-derived* (i.e., things to eat on a diet). Critically, the members of these categories lack any traditional featural similarity to one another but do cohere systematically around functional *ideals* or goal-relevant properties (i.e., zero-calorie). More broadly, the term *relational* has been proposed (Gentner & Kurtz, 2005; Markman & Stillwell, 2001) to describe categories based on how objects relate to one another within scenes or situations. For example, an 'obstacle' is a category that can take nearly any concrete or abstract form, but that coheres around fulfillment of a relationship wherein one entity blocks the progress of another. Relational categories are grounded in structure-mapping theory (Gentner, 1983), which specifies how the alignment of structured representations (entities organized by filling roles in relations) drives psychological similarity. On this view, much of the meaning that people represent about the world is more complex than simple objects and requires specification of what elements relate to other elements in what. A great deal of empirical evidence shows that comparison processes (analogy, similarity, metaphor) play a major role in human cognition, and operate based on a search for identical sets of connected relationships between cases (see Gentner, 1983). Researchers are pursuing the study of relational categories with an important emphasis on real-world learning where challenges include mastering foundational concepts in formal instructional settings and promoting successful use of acquired knowledge when the context or surface-level form is not the same (Goldwater & Schalk, 2016; Kurtz & Honke, 2017; Loewenstein, 2010).

4.4 Future Directions in Concepts

While scientific progress toward an understanding of how people learn, represent, and use categories has been considerable, there remain significant frontiers and challenges. One is that researchers have found a number of explanatory principles that do a good job of accounting for at least some part of the overall problem, but it is not clear whether the categorization system is deeply multi-faceted (i.e., variable across domains, settings, learners, etc.) or whether the range of performance characteristics reflects different manifestations of a single universal, highly flexible mechanism. Another major challenge is unifying our account of real-world, everyday categorization with advances made using highly artificial tasks in the laboratory. Lastly, there is an important need for synthesis and integration of data and theory from perspectives outside of the core approach that have produced largely siloed progress. For example, developmental psychologists have made important progress in understanding the transitions from infant to child to adult forms of categorization (Carey, 2009; Keil, 1989; Sloutsky, 2010), but there is limited cross-talk despite the obvious value to be gained. Similarly, a subset of researchers has focused on neurobiologically-oriented accounts of categories and concepts with pockets of impact arising between the approaches (e.g., Ashby & Maddox, 2005; Barsalou et al., 2003; Tyler & Moss, 2001). In addition, a set of mathematically-formulated accounts of concept formation seem to exist as a largely independent enterprise (Feldman, 2000; Pape, Kurtz, & Sayama, 2015; Vigo, 2013). We end by noting an emerging counter-example: the burgeoning field of machine learning/data science in which classification tasks are one of the core problems addressed. In a promising development, researchers are increasingly finding value in drawing upon and contributing to research on learning and representation of categories in both humans and machines.

Summary

1. Concepts emerge from the discovery of fundamental similarities between category members. They are the building blocks of thought as they connect perception to memory and allow for reasoning about unknown properties.
2. Some theorists assert that concepts are abstractions of experienced category members, either in the form of definitional rules for membership (classical view) or sets of commonalities or averages that hold in most cases (prototype approach).
3. In contrast to abstraction, some theories assume that concepts are simply stored information about individual examples that have been associated over time with category labels (exemplar approach).
4. The theory-driven view focuses on the role of concepts in explanation and considers them to be embedded in rich theoretical systems of knowledge that inform our determination of what things are above and beyond how similar features are to previous examples.
5. Concept learning is most studied through a classification task in which examples are displayed and guesses followed by feedback result in learners developing knowledge of what differentiates between members of more than one category.
6. Category learning tasks outside of the traditional classification task (observation, inference, use) often result in more robust knowledge of the internal structure of a concept.
7. Learning concepts based largely on one’s own organization and in a self-directed way can result in better learning.

8. While we can think of categories at many levels, there is evidence of a basic level (e.g., *dog*) that is favored over other levels (e.g., *mammal* or *pit bull*), perhaps because of its compromise between generality and specificity.

9. In addition to taxonomic categories, categories can be created on the fly (ad hoc), created based on relevant tasks (goal-derived), or based on relationships between features (relational categories).

Review Questions

1. Consider your own everyday concepts of the world introspectively. Which psychological account of categorization seems most plausible?

2. How would you imagine neuroimaging techniques could be used to address open questions or debates in the study of categorization?

3. Can you think of a way to resolve the difficulty of studying concept formation in the laboratory without giving up ecological validity (naturalistic properties of the stimuli, setting, and task)?

4. How do you think concepts change from when examples are first encountered to their mature state? How do concepts change across the human lifespan?

5. How do you think that changes in how people function in a digital, connected world may alter the way concepts are learned, represented, and used?

6. What constructs from the psychological study of concepts do you think could be leveraged to develop artificial intelligence capable of learning and reasoning?

Hot Topic: Categorization as finding the best account of the data

Kimery Levering

Rather than using similarity to reference points, the theory view suggests that items are categorized based on how well the item's features are explained by a category. This notion of "well-explained" can be realized without departing the realm of data. For example, one could compute the likelihood of an item having the features that it does if it were a member of a particular category. This conditional probability is based on knowing how many category members have each feature (e.g., having spots) versus not. Following Bayes' Theorem, instead of using the features to directly predict the category, one uses the likelihood of dogs having spots (and the other observed features of the target) to predict how well the category fits the example. If the example has features that occur frequently among dogs and the category itself is sufficiently common then that is strong evidence of membership. Anderson (1991) proposed a *rational* account in which the goal of categorizing is to make the most accurate possible inferences given the data. In this way, categorization is explained as forming clusters (neighborhoods) of the items in a domain and then predicting the category based on how likely each item feature is relative to each cluster combined with the likelihood of the category within each cluster. Criticisms of this approach include evidence that people make predictions based on one assigned category rather than by combining likelihoods arising from each possible category, evidence that people do not treat category labels as just like any other feature to be predicted, and the issue that the Bayesian foundations underlying this account implausibly assume feature independence.

Kenneth Kurtz

Fortunately, there is another way to determine how "well-explained" an item's features are relative to a category. Kurtz (2007) proposed that categories can be understood in terms of: (1) a transformation function instantiated as a set of synapse-like connection weights between a layer of neuron-like nodes that encode the input feature values and a "hidden" layer that recodes the information in an internal learned feature space; and (2) reconstruction functions that predict what item features are most likely with respect to each category. The paired functions represent category knowledge in the form of expectations about what configurations of feature values are consistent with membership. Error-driven learning adjusts the function pairs to work harmoniously for items that belong in each category. When an item is consistent with these expectations, it passes through the functions relatively unchanged, but when input feature(s) are inconsistent, the functions yield reconstructive distortion—the expected features do not match the observed ones. The amount of such distortion indexes the likelihood of membership. When a cat is evaluated as a dog, the result is a shift toward category expectations (i.e., bigger size, barking call, greater sociality) and this degree of distortion indicates poor category fit. A connectionist model called DIVA (see Figure 4.3) based on these principles provides a better account of human categorization on some critical tests than reference point models (e.g., Conaway & Kurtz 2017).

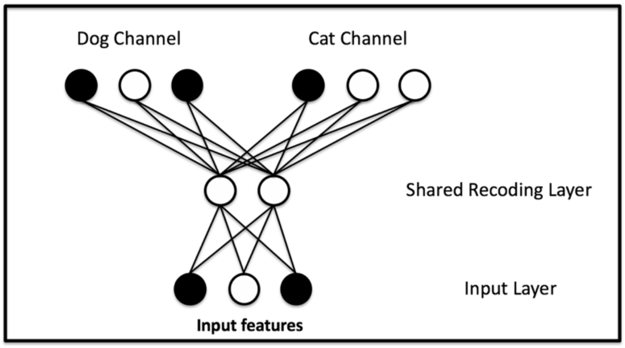

Figure 4.3: The structure of the connectionist model DIVA (Kurtz, 2007). In this example, a stimulus (three input features) is best reconstructed through the *dog* channel and so the model would classify it as a dog.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409–429. doi:10.1037/0033-295X.98.3.409

Conaway, N. & Kurtz, K. J. Similar to the category, but not the exemplars: A study of generalization. *Psychonomic Bulletin & Review*, *24*, 1312–1323. doi:10.3758/s13423-016-1208-1

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin & Review*, *14*, 560–576. doi:10.3758/BF03196806

References

Ahn, W. & Kim, N. S. (2000). The causal status effect in categorization: An overview. In D. Minda (Ed.), *The psychology of learning and motivation* (pp. 23–65), San Diego, CA: Academic Press. doi:10.1006/cogp.2000.0741

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition*, *13*, 263–308. doi:10.1006/cogp.2000.0741

Ashby, F. G. (1992). Multidimensional models of categorization. In F. G. Ashby (Ed.), *Muldimensional models of perception and cognition* (pp. 449–483). Hillsdale, NJ: Erlbaum. doi:10.4324/9781315807607

Ashby, F. G. & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149–178. doi:10.1146/annurev.psych.56.091103.070217

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*, 372–4000. doi:10.1006/jmps.1993.1023

Barsalou, L. W. (1983). Ad hoc categories. *Memory and Cognition*, *11*, 211–227. doi:10.3758/BF03196968

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 629–654. doi:10.1037/0278-7393.11.1-4.629

Barsalou, L.W. (2003). Situated simulation in the human conceptual system. *Language and Cognitive Processes*, *18*, 513–562. doi:10.1080/01690960344000026

Bruner, J. (1961). The act of discovery. *Harvard Educational Review*, *31*, 21–32.

Carey, S. (2009) *The origin of concepts*. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780195367638.001.0001

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*, 107–140. doi:10.1037/0096-3445.127.2.107

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*, 630–633. doi:10.1038/35036586

Gentner, D., & Kurtz, K. J. (2005). Relational categories. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. W. Wolff (Eds.), *Categorization inside and outside the lab* (pp. 151–175), Washington, DC: APA. doi:10.1037/11156-009

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155–170. doi:10.1207/s15516709cog0702_3

Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition*, *52*, 125–157. doi:10.1016/0010-0277(94)90065-5

Goldwater, M. B., & Schalk, L. (2016). Relational categories as a bridge between cognitive and educational research. *Psychological Bulletin*, *142*, 729–257. doi:10.1037/bul0000043

Gureckis, T. M., & Markant, D. M. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, *7*, 464–481. doi:10.1177/1745691612454304

Hampton, J. A. (1993). Prototype models of concept representation. In I. van Mechelen, J. A. Hampton, R. S. Michalski, & P. Theuns (Eds), *Categories and concepts: Theoretical views and inductive data analysis*, pp. 67–95. London: Academic Press.

Keil, F. C. (1989). *Concepts, kinds and cognitive development*. Cambridge, MA: MIT Press.

Komatsu, L. K. (1992). Recent views of conceptual structure. *Psychological Bulletin*, *112*, 500–526. doi:10.1037/0033-2909.112.3.500

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44. doi:10.1037/0033-295X.99.1.22

Kurtz, K. J., & Honke, G. (Under invited revision). Sorting out the problem of inert knowledge: Category construction to promote spontaneous transfer. doi:10.31234/osf.io/uq42r

Kurtz, K. J. (2015). Human category learning: Toward a broader explanatory account. *Psychology of Learning and Motivation*, *63*, 77–114. doi:10.1016/bs.plm.2015.03.001

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin & Review*, *14*, 560–576. doi:10.3758/BF03196806

Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Mem-*

ory & Cognition, *43*, 266–282. doi:10.3758/s13421-014-0458-2

Loewenstein, J. (2010). How one's hook is baited matters for catching an analogy. *Psychology of Learning and Motivation*, *53*, 149–182. doi:10.1016/S0079-7421(10)53004-4

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of human category learning. *Psychological Review*, *111*, 309–332. doi:10.1037/0033-295X.111.2.309

Lupyan, G. (2013). The difficulties of executing simple algorithms: Why brains make mistakes computers don't. *Cognition*, *129*, 615–636. doi:10.1016/j.cognition.2013.08.015

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*, 94–122. doi:10.1037/a0032108

Markman, A. B. & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*, 592–613. doi:10.1037/0033-2909.129.4.592

Markman, A.B. & Stilwell, C.H. (2001). Role-governed categories. *Journal of Experimental and Theoretical Artificial Intelligence*, *12*, 329–358. doi:10.1080/09528130110100252

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist*, *44*, 1469–1481. doi:10.1037/0003-066X.44.12.1469

Medin, D. L., Lynch, E. B., & Solomon, K. O. (2000). Are their kinds of concepts? *Annual Review of Psychology*, *51*, 121–147. doi:10.1146/annurev.psych.51.1.121

Medin, D. L. & Schaffer, M. M. (1978). Context theory of classification. *Psychological Review*, *85*, 207–238. doi:10.1037/0033-295X.85.3.207

Medin, D. L. & Atran, S. (2004). The native mind: Biological categorization and reasoning in development and across cultures. *Psychological Review*, *111*, 960–983. doi:10.1037/0033-295X.111.4.960

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive psychology*, *19*(2), 242–279. doi:10.1016/0010-0285(87)90012-0

Miller, G. A., & Johnson-Laird, P. N. (1976). *Language and perception*. Cambridge, MA, England: Belknap Press. doi:10.4159/harvard.9780674421288

Minda, J. P. & Ross, B. H. (2004). Learning categories by making predictions: An investigation of indirect category learning. *Memory & Cognition*, *32*, 1355–1368. doi:10.3758/BF03206326

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289–316. doi:10.1037/0033-295X.92.3.289

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT press. doi:10.7551/mitpress/1602.001.0001

Murphy, G. L. (2003). Ecological validity and the study of concepts. In B. H. Ross (Ed.), *The psychology of learning and motivation*, Vol. 43 (pp. 1–41), San Diego: Academic Press. doi:10.1016/s0079-7421(03)01010-7

Murphy, G. L. & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 904–919. doi:10.1037/0278-7393.20.4.904

Nosofsky, R. M. (1984). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 700–708. doi:10.1037/0278-7393.14.4.700

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57. doi:10.1037/0096-3445.115.1.39

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological review*, *101*, 53–79. doi:10.1037/0033-295X.101.1.53

Pape, A. D., Kurtz, K. J., & Sayama, H. (2015). Complexity measures and concept learning. *Journal of Mathematical Psychology*, *64–65*, 66–75. doi:10.1016/j.jmp.2015.01.001

Posner, M. I., & Keel, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353–363. doi:10.1037/h0025953

Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1141–1159. doi:10.1037/0278-7393.29.6.1141

Rosch, E. & Mervis, C. B. (1975). Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605. doi:10.1016/0010-0285(75)90024-9

Rumelhart, D. E. (1980). On evaluating story grammars. *Cognitive Science*, *4*, 313–316. doi:10.1207/s15516709cog0403_5

Schwartz, S. H. (1966). Trial-by-trial analysis of processes in simple and disjunctive concept-attainment tasks. *Journal of Experimental Psychology*, *72*, 456–465. doi:10.1037/h0023652

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*, 325–345. doi:10.1007/BF02288967

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*,1317–1323. doi:10.1126/science.3629243

Sloutsky, V. M. (2010). From perceptual categories to concepts: What develops? *Cognitive Science*, *34*, 1244–1286. doi:10.1111/j.1551-6709.2010.01129.x

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press. doi:10.4159/harvard.9780674866270

Smith, J. D., & Minda, J. P. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 775–799. doi:10.1037/0278-7393.27.3.775

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, *23*, 457–482. doi:10.1016/0010-0285(91)90016-H

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352. doi:10.1037/0033-295X.84.4.327

Tyler, L. K., & Moss, H. E. (2001). Towards a distributed account of conceptual knowledge. *Trends in Cognitive Science*, *5*, 244–252. doi:10.1016/S1364-6613(00)01651-X

Vigo, R. (2013). The GIST of concepts. *Cognition*, *129*, 138–162. doi:10.1016/j.cognition.2013.05.008

Vong, W. K., Navarro, D. J., & Perfors, A. (2016). The helpfulness of category labels in semi-supervised learning depends on category structure. *Psychonomic Bulletin & Review*, *23*, 230–238. doi:10.3758/s13423-015-0857-9

Wisneiwski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science*, *18*, 221–281. doi:10.1207/s15516709cog1802_2

Wittgenstein, L. (1953). *Philosophical investigations*. New York, NY: Macmillan.

Yamauchi, T. & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language*, *39*, 124–148. doi:10.1006/jmla.1998.2566

Glossary

abstraction The process of filtering out irrelevant details while preserving the most relevant, common, or significant parts. 56

basic level of categorization Intermediate level that provides the most cognitively useful compromise between being informative (members share many common traits) but also generic (glosses over minor differences). 63

category Collection of objects, people, events, or ideas in the world that are considered similar or treated similarly despite differences. 55

classical view Assumes a concept to be the necessary and sufficient conditions for membership in a category. 57

concept The mental representation of a category which can take on different forms depending on which theory is being considered. 55

exemplar approach Assumes a concept to be a collection of remembered instances that make up a category, with no abstraction. 58

graded structure When certain members of a category are thought to be better examples than others. 57

prototype approach Assumes a concept to be an abstracted list or full example consisting of common/average features that members are likely (but not required) to have. 58

selective attention A focus of resources on characteristics that are relevant for classification. 58

similarity The extent to which two or more concepts or examples are alike, either through having shared properties or close proximity in multidimensional psychological space. 56

supervised category learning Learning about a category when examples are labeled with what category they are in either initially or after a guess. 61

theory-driven approach Assumes a concept to be based on feature similarity but in service of and collaboration with knowledge-rich theories about the world. 60